

CHAPTER 1

INTRODUCTION

The World Wide Web creates a universal space of information that can be accessed by individuals, companies, government, universities, students, teachers and business people. In this chapter, the World Wide Web is introduced by describing its origin, growth and how it became one of the world's largest databases. Web Mining studies discover and analyze useful information from the World Wide Web. An overview of different topics in Web Mining studies is given. Finally, we describe and motivate our research statement, followed by the outline of this dissertation.

1.1 The World Wide Web

1.1.1 History

The World Wide Web (WWW or Web) originated in the 1980s at the ‘Conseil Européenne pour la Recherche Nucleaire’ (CERN), European Organization for Nuclear Research in Geneva, Switzerland (W3C, 2003). Here, Tim Berners-Lee worked as a researcher at the European High-Energy Particle Physics Lab. During that time there was an urgent need for collaboration between physicists and other researchers in the high-energy physics community. However, there was a great variety of computer and network systems, with hardly any common systems. Different types of information had to be accessed in different ways, systems were inconsistent and complicated leading to a big investment of effort by users. Information sharing resulted in frustration and inefficiency.

In 1989, Tim Berners-Lee wrote a proposal called *Hypertext and CERN* in order to create a solution by introducing a networked information project at CERN (W3C, 2003). Hypertext is a special type of database system, invented by Ted Nelson in the 1960’s, in which objects (text, pictures, music, programs) can be creatively linked to each other. When you select an object, you can see all the other objects that are linked to it. The proposal incorporated many new ideas and features like HyperText Markup Language (HTML), HyperText Transfer Protocol (HTTP) (Webopedia, 2002) and a web browser client software program. An important concept of the proposal included the consistence across all types of computer platforms of the client software’s program so that all users could access information from many types of computers. Finally, in May 1991 the first information-sharing system using HTML, HTTP and a client software program (called WorldWideWeb) was fully operational on the multi-platform computer network at the CERN laboratories in Switzerland (Hitmill, 2003). The physicists at CERN used the name ‘Web server’ for the main computer at CERN because it ‘served-up’ batches of cross-linked HTML documents. By the end of 1992 there where over 50 Web servers in the world. Many of these were located at universities or other research centres.

In February 1993, Marc Andreessen, an undergraduate student at the University of Illinois at Urbana-Champaign, was working on a project for the National Centre for Supercomputing Applications (NCSA) when he led a team that developed the graphic interface browser called *Mosaic*. People without computer expertise are now able to navigate the Web by just pointing and clicking on objects of their choice.

Finally, in April 1993, CERN's directors declared that World Wide Web technology would be freely usable by anyone, with no fees being payable to CERN. A milestone in the history of the World Wide Web.

1.1.2 Growth

The following years the World Wide Web grew quickly. In 1999, there were more than 720,000 public information servers. Subsequently, in 2001 there were over 24 million servers and the May 2003 survey of Netcraft (2003) collected more than 40 million servers. In Belgium, an analysis of ISPA Belgium (2003), organization of Belgian Internet providers, registered 370.000 Internet connections in July 1999 and 1,800,000 Internet connections in March 2003.

Today, the World Wide Web represents a universe of information through which people can communicate and collaborate by means of a system of Internet servers that support specially formatted documents. The documents are formatted in a script called HTML that supports links to other documents, as well as graphics, audio and video files. This means you can jump from one document to another simply by clicking on hot spots (Webopedia, 2002). Yet, the Web is not identical to the Internet. The Web is one of many Internet-based communication systems. On the Internet, you can run data services like electronic mail, file transfer, remote log-in, bulletin boards and the World Wide Web. The Internet is often compared with traffic: computers must use the cables in an agreed fashion to avoid chaos. Therefore, a common protocol is used called Transmission Control Protocol/Internet Protocol (TCP/IP). The Web is like a parcel delivery service on the Internet. At your request, World Wide Web servers will send you documents (CERN, 2002).

In order to define standards for the Web to evolve in a single direction rather than being splintered among competing factions, Tim Berners-Lee, also known as the architect of the WWW, founded the World Wide Web Consortium (W3C) in 1994. Standards exist for programming languages, operating systems, data formats, communications protocols and electrical interfaces.

The explosive growth of the World Wide Web, which is still expected to continue, gives rise to ethical and social concerns. Questions about privacy and intellectual property are still open for discussion in this virtual information space. Yet, some privacy policies for Web visitors are regulated by privacy statements (Mena, 2001). In general, it must be clear that gathering and storing customer information adds value to both parties: faster and customized services for the web visitor and commercial assets for the web owner.

1.2 Web Mining

1.2.1 Definition

Despite the fact that Web Mining is a relatively new study, many definitions about Web Mining are given (Chang et al, 2001; Cooley et al, 1997; Mena, 1999; Mobasher et al, 1996; Mulvenna et al, 2000; Zaïane, 2000). They all describe Web Mining from a particular point of view.

Cooley et al (1997) defines Web Mining as the discovery and analysis of useful information from the World Wide Web. Following Zaïane (1998) and Borges and Levene (2000), Web Mining is a common term for the application of data mining techniques on the Web. Data mining is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from data in large databases (Piatetsky-Shapiro et al, 1996; Zaïane, 1998). Databases are considered large when they contain several hundred thousand transactions (Mena, 2001). Besides, data mining is about knowledge discovery for a strategic, tangible, competitive business advantage (Mena, 2001). This implies that Web Mining includes all techniques that aim to discover knowledge from the World Wide Web. The goal of *knowledge discovery* is to uncover implicit knowledge not necessarily stated in any resource. This may not be confused with resource discovery, which goal is to find explicit information (Zaïane, 1998).

Finally, Mena (2001) gives a more practical definition. Web Mining is all about improving the customer experience while optimising business profitability. In general, three knowledge discovery domains pertain to Web Mining: Web Content Mining, Web Structure Mining and Web Usage Mining (Chang et al, 2001; Cooley et al, 1997; Mobasher et al, 1996; Zaïane, 2000). An overview is given in figure 1.1.

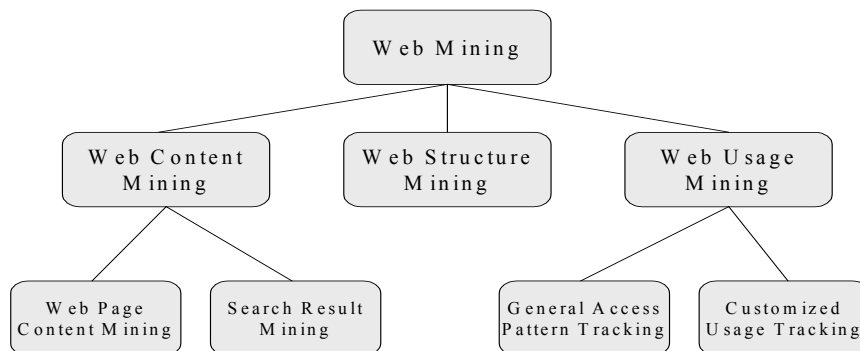


Figure 1.1: Taxonomy of Web Mining techniques.

1.2.2 Web Content Mining

Web Content Mining is the process of extracting knowledge from the content of documents and their descriptions. Two groups of Web Content Mining strategies exist. First, *Web Page Content Mining* mines directly the content of documents. For example, WebSQL (Arocena and Mendelzon, 1998) uses graph trees to extract knowledge and restructure web documents. Ahoy (Shakes et al, 1997) applies heuristics to distinguish personal home pages from other web pages and ShopBot (Doorenbos et al, 1997) looks for product prices within web pages. In Ester et al (2002), a new approach is introduced for spotting interesting information on the World Wide Web. Instead of classifying pages, more complex objects of complete web sites are spotted by means of several approaches for classification.

Second, improvement of the content search of other tools like search engines is accomplished by *Search Result Mining*. Sieg et al (2003) present ARCH, an interactive query formulation aid that is based on conceptual categories. The goal of the system is to meet the user's information needs by closing the gap between the user's stated query and the actual intent of the search. Kwok et al (2001) extent question-answering techniques, first studied in the information retrieval literature, to the web and experimentally evaluate their performance. Zamir and Etzioni (1998) present a tool for clustering documents retrieved by a set of search engines. The techniques are based on information provided in search results like phrases, URL's or snippets. Snippets are descriptions or first lines of the page content.

1.2.3 *Web Structure Mining*

Web Structure Mining is the process of inferring knowledge from the World Wide Web organization and links between pages in the Web. Links pointing to a document indicate the popularity of the document while links coming out of a document indicate the richness or the variety of topics covered in the document. Examples of techniques of Web Structure Mining are Hypersuit (Weiss et al, 1996), PageRank (Brin and Page, 1998) and CLEVER (Chakrabarti et al, 1998). Based on interconnections between web pages a weight is given to pages in order to find pertinent web pages. Another technique is the Multi Layered Database Approach MLDB (Han et al, 1995; Zaïane and Han, 1995), which uses a multi-level database representation of the Web in order to represent structure and content of the World Wide Web.

1.2.4 *Web Usage Mining*

Web Usage Mining, also called web log mining, is the process of extracting previously unknown and interesting usage patterns in web access logs (Cooley et al, 1999a; Zaïane, 1998). Web servers record data about user interactions on the web in log files. Analysing log files (or web access logs) of web sites can help us understand user behaviour. Applying data mining techniques on web access logs allows management to optimise the site for the benefit of visitors (Foss et al, 2001). Likewise, understanding and modelling visiting behaviour may lead to strategies for web personalization and in general strengthen competitive advantage.

Two main research areas exist in Web Usage Mining. First, *General Access Pattern Tracking* analyses log files to understand access patterns and trends. This information is used for optimising web sites through better structures, design, layout and page-links. Other applications are grouping of resource providers and web personalization techniques to provide better services to visitors. Examples of techniques to understand general access patterns and trends are WebLogMining (Zaïane et al, 1998), Speedtracer (Wu et al, 1998), Wum (Spiliopoulou and Faulstich, 1998) and WebSIFT (Cooley et al, 1999c). In Dai and Mobasher (2003), various approaches are explored for integrating semantic knowledge into the personalization process based on Web Usage Mining. In Perkowitz and Etzioni (2000), adaptive web sites mine the data buried in web server logs to produce more easily navigable web sites. Another example of General Access Pattern Tracking is given in Spiliopoulou et al

(2000), where the effectiveness of a web site is improved with web usage mining.

Second, *Customized Usage Tracking* analyses individual trends in order to customize web sites to users. Displayed information, depth of site structure as well as format of resources can all be dynamically customized for each user over time based on their access patterns. In Perkowitz and Etzioni (1997) automatically adaptive sites are presented through learning from user access patterns. SiteHelper (Ngu and Wu, 1997) provides individual web personalization and Web Watcher (Joachims et al, 1997) provides recommendations to web users.

1.3 Research statement and motivation

A well-known problem in studies of web usage behavior is extracting information about the order in which people visit web pages. If we are able to group temporary aspects of visiting patterns together based on how the order of pages occurs within patterns and if large groups are provided with similar order-based characteristics, we could use this information for the benefit of the visitor. For example, if 10% of the visiting patterns show that page x is visited before page y, and that page y is visited before page z, we may *predict* that page z will be visited after page x followed by page y. This way, a proxy server is able to provide faster deliveries of pages to web visitors. Also, if no direct hyperlinks exist from page x to y and from page y to z, we may suggest inserting *direct hyperlinks* for the ease and convenience of the visitor. Another advantage of using order-based information of groups of visiting patterns is suggesting cross-links between web pages. *Cross-links* are references to other documents based on common features. Especially for commercial web sites, cross-links may be an interesting asset. For example, people who visited product X are offered a cross-link to product Y. Through studying the order of visited pages we are able to examine which cross-links are used and which are not used. This means that, in the end, the results of cross-selling products X and Y through the web may be compared with the effect or use of cross-links. In other application domains like studies of the *learning curve* and *psychology* of the web visitor, the order of visited pages may provide information to construct guided tours for first time visitors and for visitors who frequently use the web site.

In order to provide order-based information that is useful for prediction, link optimization, learning curve and psychology studies, we used basic, existing or related work but we also introduced three new concepts in Web Usage Mining.

First, sequences are constructed holding pages as well as times. Second, sequences are clustered based on a distance measure using Sequence Alignment Methods instead of the commonly used Association distance measure. Third, the results are graphically presented showing the structure of the web site, including direct hyperlinks between web pages, and the most frequent occurring patterns depicting the order in which web pages are visited.

Generally, in this study we will concentrate on Web Usage Mining studies and more precisely on General Access Pattern Tracking. The technique that we will use throughout this thesis to understand general access patterns and trends from log files on web sites is clustering sequences representing navigations (and not visitors or users) based on Sequence Alignment Methods (SAM). The advantage of clustering in General Access Pattern Tracking is that groups of patterns are provided with small differences within the groups and large differences between the groups. The advantage of using *SAM* over classical clustering methods is that SAM group sequences together based on the *order* of occurrence of elements in a sequence. Moreover, two information types such as visited pages and visiting times are handled by 2-dimensional SAM (2-dim SAM). The advantage of using *2-dim SAM* over SAM is that 2-dim SAM group sequences together based on the *order* of occurrence of elements and on *relations* between visited pages and visiting times. Objectives of using SAM and 2-dim SAM on web usage data are given in chapter four, section 4.3.

No prior work has been found yet concerning the problem of mining navigation patterns using a measure that incorporates the order of elements within sequences. Cadez et al (2000) use a mixture of first order markov chain to model and cluster web behavior. They claim to measure order-based information. However, the main difference between SAM and the method used by Cadez et al (2000) is that SAM examines order-based information within the entire sequence whereas the approach of Cadez et al (2000) models only parts of the sequence. For example, the sequence AB is modeled by the approach of Cadez et al (2000) not taking into account whether the sequence is ABC or CAB or whatever. So they take into account only the previous web page instead of the entire sequence. This means that important or interesting information may be lost. For example, suppose one direct hyperlink exists from page A to B and no direct hyperlink exists from page B to page C. Obviously, the support of sequence AB will be much larger than the support of sequence BC. For this reason, the approach of Cadez et al (2000) concentrates on sequence AB. Yet, sequence BC may be interesting as well. Suppose the following situation occurs. Only 1% of the patterns provide sequence BC and 90% of the patterns holding B also hold C. This means that 90% of the visitors who went to B (or C) also went to C (or B), without the existence of a direct hyperlink between B and C. This is useful to know for link optimization

studies. Moreover, in order to support navigations through guided tours, entire sequences must be studied. For these reasons we analyze entire sequences instead of parts of sequences or sub-sequences.

1.4 Dissertation outline

In chapter two, existing or related work about web usage mining studies is given. Several data sources are described and definitions about frequently used concepts and terms throughout this thesis are given. In chapter three, the algorithm of SAM and 2-dim SAM are described and illustrated with examples, without yet considering clustering based on SAM or 2-dim SAM.

In order to show how SAM and 2-dim SAM provide advantages when looking for order-based information within sequences, chapter four discusses the surplus value of clustering based on SAM over classical clustering methods as well as the surplus value of clustering based on 2-dim SAM over SAM. After defining the objectives, SAM and 2-dim SAM are applied to three real data sets and sequences are clustered based on SAM and 2-dim SAM. Then, the clustering results based on SAM are compared with a classical clustering method. Also, the clustering results based on 2-dim SAM are compared with SAM.

Clustering sequences based on SAM or 2-dim SAM distance measures provide a general overview of large groups of visiting patterns on a web site. Also, general information about the order in which pages are visited and small differences within the groups along with large differences between the groups are studied. This also means that most of the general visiting patterns are indicated by the structure of the web site. Direct hyperlinks between web pages offer a 'road' to visitors leading to obvious visiting patterns, which are generally and mostly extracted by our algorithm of clustering based on SAM or 2-dim SAM. Yet, in order to search for navigation patterns that are interesting instead of general or obvious, in chapter five, SAM is extended with an interestingness measure. A pattern is interesting if it is unexpected or surprising. For example, if page x is usually followed by page y without a direct hyperlink from page x to page y the pattern x followed by y is interesting. Chapter five discusses the technique for measuring interestingness based on Baldwin's support logic since Baldwin introduced a general measure for interestingness that is easy to apply in different research domains. The technique is applied to a real data set and the results provide interesting navigations as well as non-existing navigations given a provided structure.

From the results, suggestions are given for optimizing the structure of the web site, analogue with how visitors behave.

In order to study the stability of the SAM algorithm, chapter six provides an analysis of the sensitivity of SAM towards changes in the parameters. To examine a broad range of parameter settings, the effects of small, medium or large changes in the parameters of SAM on the results are analyzed.

Although we may provide a good method for extracting order-based information within visiting behavior on a web site, we still face the problem of handling large databases by means of SAM-based clustering. Moreover, a good algorithm for analyzing sequences in Web Usage Mining studies must be able to handle large databases. Therefore, in chapter seven, the computational complexity of SAM is described and a heuristic for analyzing large databases by means of SAM is provided.

CHAPTER 2

WEB USAGE MINING: RELATED WORK

In this chapter, three different data sources, analyzed by Web Usage Mining studies, are described. Also, URL addresses of web sites providing sources of web data are publicly available on the World Wide Web. Besides data sources this chapter also provides descriptions of frequently used terms throughout this thesis. Different data types, data abstractions and page types are defined. Following, the Web Usage Mining process along with three analysis steps, which are called pre-processing, processing and post-processing are described. In each step some heuristics and techniques related to Web Usage Mining studies are given to clean and mine the data. Also, methods for analyzing the results are given. In addition, examples are provided of the approaches that are used throughout this project. Finally, in the pre-processing step of the Web Usage Mining process, typical problems such as how to deal with outliers, detect accesses that are not recorded by the data sources and user identification are discussed and methods for handling these problems are given.

2.1 Data sources

In order to discover and analyse web usage patterns, data needs to be collected from different sources. A general framework for data collection is given in Cooley (2000) and Srivastava et al (2000) where three levels of data sources are described: *server-level*, *client-level* and *proxy-level*. Each level differs in terms of format, scope, method of implementation, accuracy and reliability.

2.1.1 *Server-level data collection*

Server-level data collection often replies to data created by the web servers in *log files* (or server logs) where the browsing behaviour of visitors is recorded. These log files are stored in various formats, such as Common Log file Format (CLF) or Extended Common Log file Format (ECLF). The format of a common log file line has the following fields, separated by a space: IP address (or remote host name), user id (or remote login name of the user), date, request, status and bytes. An extended common log format file is a variant of the common log format file and, for each request, keeps track of two additional fields: referrer and user agent. Examples of lines in common and extended log file formats that we used throughout this thesis are shown in table 2.1. If data is missing, a minus sign is typically placed in the field. Other examples of log file data are given in Cooley et al (1999), Fu et al (1999), Srivastava et al (2000), Zaïane and Luo (2001), Zaïane et al (1998). The basic data fields in log files are explained in appendix 1.

ECLF							
CLF						Referrer	User agent
IP address	User ID	Date	Request	Status	Bytes	Referrer	User agent
195.238.3.198	-	2001-03-03 00:01:43	GET/bib/src/top/law-top.html	304	80	-	-
195.238.3.198	-	2001-03-03 00:01:45	GET/leeromgeving/trajecten_tew/0122/0122.htm	304	80	-	-
212.190.0.252	-	2001-03-03 00:01:46	GET/images/luthemehover/actuhov.gif	304	80	-	-
...
142.56.200.14	-	1999/01/31-23:59:07	GET/music/machines/manufacturers/Akai/MPC/samples/HTTP/1.1	403	1471	-	Mozilla/4.0(compatible; MSIE4.01; MSIECrawler; Windows 95)
100.77.86.90	-	1999/01/31-23:59:13	GET/music/machines/manufacturers/EMS/Overview/vcs3.gifHTTP/1.0	301	279	http://www.ems-synthi.demon.co.uk/emsprods.html#vcs3	Mozilla/4.05(Macintosh; I; PPC,Nav
100.77.86.90	-	1999/01/31-23:59:15	GET/music/machines/manufacturers/EMS/Overview/vcs3.gifHTTP/1.0	200	212111	http://www.ems-synthi.demon.co.uk/emsprods.html#vcs3	Mozilla/4.05(Macintosh; I; PPC,Nav
...

Table 2.1: Sample Common Log file Format (CLF) and Extended Common Log file Format (ECLF).

Collecting data from server-level data sources is, unfortunately, not as easy as it seems. Data stored in log files are not completely reliable due to several reasons:

1. Not all pages are recorded in log files because of client- and proxy-level caches. A client cache occurs when a visitor uses the 'back' or 'reload' button. A proxy-level cache occurs when the page, requested by the client, is delivered by the proxy server instead of the web server. A proxy server is described in section 2.1.3 Proxy-level data collection.
2. Logged time information may be inaccurate when a page is delivered by client or proxy caches. This means that the page view time of its previous page will be interpreted to be longer than it actually was. Also, the logged page view time which is often calculated as the time difference between two subsequent, logged, requests, may differ from the actual view time due to reasons like connection speed of the client, size of requested page file and network congestion.

3. When the user id is not available and clients request pages behind a proxy server, many requests are recorded with the same IP address (usually the proxy's host name). As a result, page views may seem to be erratic, with very short viewing times.

Methods to deal with these problems are cookies and packet sniffers or network monitors. A *cookie* is a message given to a web browser by a web server. The browser stores the message in a text file, called cookie.txt. This message is sent back to the server each time the browser requests a page from the server (Webopedia, 2002). Cookies may be used to identify users and to prepare customized web pages. A *packet sniffer* is a program that records all network packets that travel past a given network interface. It is used to analyse network traffic and helps a network manager to keep traffic flowing efficiently. It is also used to record hidden parameters that are not stored in log files. Unfortunately, besides legitimate use, sniffers are used as well for stealing information off a network. Examples of free packet sniffing tools are Ethereal, Ksniffer, Snort, IpGrab and IpLog (Packet Sniffing Tools, 2003). In Pitkow (1997), a complete discussion on the shortcomings of the current log standard and potential solutions are given.

Not only browsing behavior or usage data is collected from server sources but other data as well such as content data, structure data and web page meta information. For example, in Cooley et al (1999c), usage, content and structure data are integrated into an algorithm called Web Site Information Filter (WebSIFT) to identify interesting knowledge. Also, in Chan (1999), usage and content data are used to build user profiles. The different data types are explained in section 2.2.1 Data types.

2.1.2 Client-level data collection

Data processing now occurs at the client side instead of at the server side. Two important client level sources for data collection are available: remote agents and modified browsers. *Remote agents* are used to record single user – single site browsing behaviour. Examples of remote agents are Javascripts and Java applets. In Shahabi et al (1997), users navigation paths are detected using a Java based remote agent. Likewise, the WebSIFT system, which is a web usage mining system that discovers interesting behaviour on web sites, uses optional data such as remote agent logs to provide information for constructing information abstractions, such as page views and user sessions (Cooley et al, 1999c). Using *modified browsers* (such as Mosaic or Mozilla), single user – multi site browsing behaviour is collected. A practical example of employing the Mozilla browser and Microsoft Internet Explorer browser using JavaScript,

in order to capture behavioural aspects on the web, is introduced by the Web Event-Logging Tool (WET) (Etgen and Cantor, 1999).

The major advantage of implementing agents and modified browsers is that the problem of cached pages is resolved. In Srikant and Yang (2001), an algorithm is presented that can handle page caching by the browser. However, by using Java applets the problem of the actual page view time still exists and some overhead may occur. Then again, Javascripts reduce the overhead problem but they cannot capture all user clicks due to clicking the 'back' or 'reload' button by the site visitors. Another disadvantage of using both client level data sources is that user cooperation is required. Without approval of the user, the remote agent cannot be implemented or the modified browser cannot be used.

2.1.3 Proxy-level data collection

A third source for data collection that is used in web usage mining studies is the proxy server. A *proxy server* (for example an ISP provider) acts as intermediary between a client application, such as a web browser, and a real server. It intercepts all requests to the real server to see if it can fulfil the requests itself. If not, it forwards the requests to the real server. Proxy servers have two main purposes: performance improvement and filtering requests.

Performance improvement Because proxy servers save the results of all requests for a certain amount of time and because the user often is on the same network as the proxy server, the performance for groups of users is improved.

Filtering requests Proxy servers are also used to filter requests. A company might for example use a proxy server to prevent its employees from accessing a specific set of web sites.

Employing proxy level data sources, cached pages are collected from proxy traces in order to reveal the actual HTTP requests from multiple clients to multiple servers. Also the browsing behaviour of a group of users sharing the same proxy server may be identified. An example of how proxy servers are used for overcoming many of the problems with server-side and client-side logging, is given by the WebQuilt system (Hong and Landay, 2001).

2.1.4 World Wide Web

Other sources of data for Web Usage Mining studies are publicly available on the World Wide Web. Table 2.2 lists some addresses of web sites where data can be downloaded. In addition, a description of the data set is given.

Address	Data set description
http://www.kdnuggets.com/datasets/index.html	Data sets for all kinds of Data Mining applications within different research areas.
http://www.kdcentral.com/Data_Sets/Web_Log_Mining/	Data sets representing web click stream and purchase data from Gazelle.com, a leg wear and leg care web retailer.
http://maya.cs.depaul.edu/~classes/ect584/resource.html	Three data sets are presented. First, DePaul CTI Web Usage Data contains pre-processed and filtered sessionized data of visits to the main CTI site during a two-week period. Second, Movie Ratings Data presents real movie ratings from the www.movielens.org web site. The data set holds ratings on more than 1600 movies by 1000 users. Finally, the third data set, called UCI KDD Archive, is an online repository of large data sets, which encompasses a wide variety of data types, analysis tasks and application areas.
http://www.cs.washington.edu/ai/adaptive-data/	Here, web logs storing data of three years' surfing behaviour on the web site http://machines.hyperreal.org are given. Also, logs of user accesses to http://www.cs.washington.edu , the Department of Computer Science and Engineering at the University of Washington, from August 1998 and January to September 1999 are presented. These logs record data that was used in Perkowski and Etzioni (1997).

Table 2.2: Publicly available data sets for web usage mining.

2.2 Web usage terms

The data sources, which are described in the previous section, record different kinds of data that is structured in several ways. In order to create some consistency in the discussions that follow, we provide an overview of frequently used concepts in Web Usage Mining studies (Cooley, 2000; W3C, 2003).

2.2.1 Data types

Data is generally classified into several groups according to the type of information (Cooley, 2000; Cooley et al, 1999a; Cooley et al, 1999b; Srivastava et al, 2000):

Content data This is the real data in the web pages, the substantive or meaningful part of data the web page was designed to convey to the visitors. Some examples of content data are graphics and text.

Structure data The organization of the content is described by structure data. Two kinds of structure data exist: intra-page and inter-page structure data. The first includes the arrangement of various HTML or XML tags within a given page; the second contains information about the hyper-links connecting one page to another.

Usage data This type of data describes the pattern of usage of web pages. Examples of usage data are IP addresses, page references and date/time of access. A typical source that collects usage data is the ECLF server log.

2.2.2 Data abstractions

In order to discover patterns of web usage, different types of data, collected from different sources, must be organized and prepared in a way that pattern discovery techniques can be applied to the data. Therefore, data abstractions are defined. Definitions of frequently used data abstractions throughout this thesis are given below.

User A person using a client application to interact and retrieve resources from the server.

Client A role assumed by an application when retrieving resources from the server.

Request A message describing an atomic operation to be carried out in the context of a specified resource. For example, HTTP GET, POST, PUT and HEAD requests.

Web page A collection of resources identified by a single URL.

Web site A collection of interlinked web pages, including a host page, residing at the same network location. Interlinked means that any of a web site's constituent web pages can be accessed by following a sequence of references beginning at the site's host page.

Page view The rendered web page in a specific client application.

User session The click stream of page views for a single user across the entire web. The user requests pages from one or more web servers. For example, if a user visits the web sites <http://www.airplane.com> and <http://www.flightsim.com>, a user session might look like this:

<http://www.airplane.com/faq.php>, <http://www.airplane.com/features.php>,
<http://www.flightsim.com/login.htm>,
<http://www.flightsimnetwork.com/cgi/dcforum/dcboard.cgi>,
<http://www.airplane.com/contactus.php>

Server session / Visit The click stream of page views for a single user to a web site. The user requests pages from a single web server. For example, if a user visits the web site <http://www.diamonds.com>, a server session might look like this:

<http://www.diamonds.com>, <http://www.diamonds.com/necklaces.html>,
<http://www.diamonds.com/rings.html>

Session Uniform name to refer to user session and/or server session (visit)

Episode A subset of page views from a server session.

2.2.3 Page types

A third way for structuring data is classifying web pages into groups according to their intended use.

Head/Home page Also known as host page. This should be the first page that users visit when entering the web site. Examples of home pages used throughout this thesis are <http://www.luc.ac.be/tew> and <http://machines.hyperreal.org>.

Content page The purpose of this page is providing mainly content information of the web site. Examples of content pages are http://www.luc.ac.be/tew/opleidingen/basisopleidingen/opbouw_hi/3de_jaar_hi_kmo.htm and <http://machines.hyperreal.org/manufacturers/Yamaha/DX-100>.

Navigation page The purpose of this page is providing mainly links to guide users to content pages. Examples of navigation pages are <http://www.luc.ac.be/tew/information> and <http://machines.hyperreal.org/guide>.

Regarding content and navigation pages we remark that it may be difficult to define a-priori which pages are content or navigation pages because of the following reasons:

- Defining content and navigation pages is dependant of the user. For example, a user who visits the web site for the first time, will use particular web pages as ‘content’ while for users who frequently visit the web site, the same web pages are visited as ‘navigation’.
- Every web page is in fact a combination of ‘content’ and links, which means that, defining which pages are ‘content’ and which are ‘navigation’ is not an easy task.

To deal with these difficulties when a-priori categorizing web pages into ‘content’ and ‘navigation’ pages, we define *index* pages as navigation pages and non-index pages as content pages. Also, pages that are offered to the visitor as a ‘guide’ on the web site (guided tours) are often navigation pages. Yet, we still are aware of the fact that page x may be used a-posteriori as ‘navigation’ (or ‘content’) while being a-priori defined as ‘content’ (or ‘navigation’).

Dynamic page A dynamic page refers to web content that changes each time it is viewed. For example, the same URL could provide different

information depending on parameters such as geographic location of the reader, time of day, previous pages viewed by the reader or the reader's profile. There are many technologies for producing dynamic HTML, including Common Gateway Interface (CGI) scripts, Server-Side Includes (SSI), cookies and Java (Webopedia, 2002). The opposite of dynamic is static. In Nasraoui and Rojas (2003) an approach is proposed that considers Web Usage data as a reflection of a dynamic environment which therefore requires dynamic learning of the access patterns from non-stationary Web usage environments.

Static page A static page can only supply information that is written into the HTML and this information will not change unless the change is written into the source code (Webopedia, 2002). In this research project, static web pages are used in examining surfing behavior on web sites.

2.3 Web Usage Mining: a data mining process

Web Usage Mining is defined as the application of data mining techniques on web access logs allowing management to optimize the site for the benefit of visitors (Foss et al, 2001). Before proceeding to the details of Web Usage Mining we first define data mining.

A standard definition for data mining, also called knowledge discovery, is the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data (Agrawal, 1993; Zaïane, 1998). Another definition is that data mining is the automatic or (more usually) semi-automatic process of discovering patterns in large data sets, containing several hundred thousands, even millions, of records. The patterns discovered must be meaningful, i.e. leading to some advantage, usually an economic advantage. Besides, data mining is the acquisition of knowledge and the ability to use it (Witten and Frank, 2000). Mena (1999) defines data mining as a process that recognizes patterns through inductive data analysis. Also, data mining involves various techniques such as association, classification, clustering and segmentation, which are used for decision-making knowledge in areas like estimation, optimisation, sequencing, prediction and visualization. This means that data mining is not query or user-driven. Instead, it is driven by the need to uncover hidden undercurrents in the data. Finally, Agrawal (1999) and Thearling (2003) state that data mining converges three technologies: statistical and learning algorithms, increased computer power and improved data collection and management. An overview of data mining is given in Hand et al (2001). Eventually, the web creates one of the greatest opportunities for data mining,

due to a huge collection of data and a universal digital distribution medium, which makes data mining results actionable in fundamentally new ways (Agrawal, 1999).

Finally, a Web Usage Mining process consists of three steps occurring in the following order: pre-processing, processing (i.e. the actual data mining step) and post-processing. Output of each step is used as input for each subsequent step. The following sections describe each step within a Web Usage Mining process.

2.4 Pre-processing

Pre-processing, also called *data preparation*, is necessary to convert raw data, recorded in different data sources, into usable data in order to perform mining techniques on the usage data. Generally, the files analysed in this step are server log files, web site files or usage statistics from previous analyses. Outputs of this step are user session files, server session (visit) files, episodes, site topology or page types. Within Web Usage Mining, the pre-processing step is divided into two activities: data cleaning and transaction identification (Cooley, 2000).

2.4.1 Data cleaning

First, because our objective is analysing visiting behaviour on web sites, irrelevant items should be eliminated from the data source files. An *irrelevant item* is usually recognizable by the suffix of the URL name. For example, all log entries with filename suffix like gif, jpeg, jpg, GIF, JPEG, JPG and map should be removed. Furthermore, when analysing data, sometimes a value can be far from the others. Such a value is called an *outlier* and often these values are the result of an error in data entry. Outliers can be removed from the data, although never without any special attention. In our research project, we found seven user requests that were recorded in a log file without time information. These requests are very difficult to handle in the following steps of the analysis. Without any time information, they might erroneously be assigned to a server session. For this reason, we omitted the records from the data.

A second task to fulfil during the data cleaning process is *detecting accesses that are not recorded in log files*. Reasons why certain elements are not logged as well as methods to deal with this problem are given in section 2.1.1 Server-level data collection. In our research project, we did not use information from cookies nor did we use packet sniffers to detect proxy-level caches. However,

we did use a filtering method in order to overcome most of the problems of requests of different users recorded with the same IP address due to proxy servers. For example, the filtering method searched through the log files for requests recorded with equal IP addresses and time information. Also, if the information provided by the referrer corresponded with the 'road' indicated by the web site structure (taking into account the possibilities of using 'back' buttons), the filtering method kept the records for further analysis.

A third task is *user identification*. Because of the use of proxy servers, it is hard to identify each individual user who accesses and views pages on the web site. The main problems concerning user identification are the following:

Single IP address / multi server session ISP's have a pool of proxy servers through which users can access the web. This situation results in one IP address for different users, potentially over the same period of time.

Multi IP address / single server session Some ISP's randomly assign each request from a user to one of several IP addresses. In this case, a single server session can have different IP addresses.

Multi IP address / single user This occurs when a user accesses the web from different machines. Each session will have a different IP address dependent on what machine was used.

Multi server session / single user When a user opens up more than one browser window and accesses different portions of a web site simultaneously, the log file will collect different server sessions reporting erroneously accesses of different users.

Single client / multi user If more than one individual uses the same computer, such as families or public access machines, all accesses will be recorded as if it were only one user.

Examples of user tracking approaches are given in Shahabi and Banaei-Kashani (2001), Shahabi et al (2000), Shahabi et al (1997). A framework for the evaluation of the accuracy of sessionizing tools is presented in Berendt et al (2001). Finally, the main tracking mechanisms for handling problems of cached pages and user identification are presented in table 2.3 (Cooley, 2000). A description of each mechanism is given, along with privacy concern levels ranging from low to very high, advantages and disadvantages.

Mechanism	Description	Privacy concern	Advantages	Disadvantages
IP address and agent	Assumes that each unique IP address/agent pair is a unique user.	Low.	Always available; no extra technology is required.	No guarantee of referring to a unique user, generally because of random or rotating IP.
Embedded session ID	Uses dynamically generated pages to insert ID into every link.	Low/ Medium.	Always available and independent of IP address.	Repeated visits are not collected. Requires full dynamic site.
User registration	Users explicitly sign-in to the site.	Medium.	Can track single individuals, not just browsers.	Not all users may be willing to register.
Cookie	Saves an identifier on the client machine.	Medium/ High.	Can track repeat visits.	Can be disabled. Negative public image.
Software agent	A program is loaded into the browser that sends back usage data.	High.	Accurate usage data for a single web site is collected.	Likely to be refused. Negative public image.
Modified browser	The browser records usage data.	Very high.	Accurate usage data across the entire web is collected.	Users must explicitly ask for software.

Table 2.3: Tracking mechanisms for cached pages and user identification (Cooley, 2000).

Yet, in most research projects, cached pages are not recorded (Cadez et al, 2000). In the absence of cookies or dynamic web pages, the combination of IP address and user agent is used to identify users (Cooley et al, 1999b; Fu et al, 1999). Moreover, using only the information supplied by the log file for pre-processing tasks is supported and specified in the HTTP protocol by CERN and NCSA (Cooley et al, 1999a). In our research project we cleaned the data in log files following the approaches of Cadez et al (2000), Cooley et al (1999b) and Fu et al (1999).

2.4.2 *Transaction identification*

Before data mining techniques are applied on the web usage data, sequences of page accesses must be grouped into logical units representing server sessions (visits) or user sessions. Depending on the criteria of identifying transactions, the size of a transaction can range from single page reference to all page references in a user session or server session. In Cooley et al (1997), three different techniques are described to identify transactions. The selection of the technique is subject to the application area of the results of the Web Usage Mining project.

Navigation-content technique Each transaction consists of a single content page reference and all the navigation page references leading to this content page reference. This method is used to mine for traversal patterns.

Content-only technique Each transaction consists of all the content page references in a user session or server session. These transactions are used to discover associations between content pages. In Cadez et al (2000) and Banerjee and Ghosh (2001), content page references are grouped into a higher level called page categories or concepts. Each transaction contains categories of web pages or concepts instead of URL page requests.

Maximal forward reference technique Each transaction is defined as the set of pages in the session from the first page in the log file up to the pages before a backward reference is made. When the next forward reference is made, a new transaction is started. A backward reference is a page view already present in the set of pages of the current session. Similarly, a forward reference is a page view that is not present in the set of pages of the current session. This technique is used to discover path traversal patterns. An example is given in Srikant and Yang (2001), where the point from where visitors backtrack is identified as a possible expected location of a web page.

Practically, most research projects adopt a standard pre-processing technique for transaction identification that includes some time-out heuristic (Banerjee and Ghosh, 2001; Catledge and Pitkow, 1995; Cooley et al, 1999a; Zaïane and Luo, 2001), which is also used in our research project. Because it is very likely that users will visit the web site more than once, the goal of using time-out heuristics is to identify individual transactions.

2.5 Processing

The second step within a Web Usage Mining process represents the actual data mining process, also called *pattern discovery* or *pattern recognition*. The outputs of the pre-processing step are used as inputs for the processing step. Different techniques are applied to discover patterns in data. The choice of which technique is used depends on the needs of the analyst and the type of data. Here, we distinguish two groups of techniques. In the first, typical data mining methods are used to mine the usage data. In the second, modifications of the typical data mining methods are used.

2.5.1 Typical data mining methods

The typical data mining methods that are applied within Web Usage Mining are summarized in Cooley (2000), Cooley et al (1999a), Cooley et al (1997), Srivastava et al (2000). An overview of data mining techniques, without considering the application to web related data, is given in Agrawal and Srikant (1994a), Fayyad et al (1996), Witten and Frank (2000).

2.5.1.1 Association rules and frequent item sets

Association rules present unordered associations and correlations among data items where the presence of one set of items in a transaction implies the presence of other items. In Agrawal and Srikant (1994), Agrawal et al (1996), fast algorithms for mining association rules are given. The new association algorithms Apriori and AprioriTid are introduced and compared with AIS and SETM algorithms. Empirical evaluation shows that the new methods outperform the existing ones. The difference between apriori algorithms and other existing ones is that apriori generates and counts less item sets during candidate generation. Besides, before a new pass begins, it concludes *a priori* that some combinations are not possible because lack of minimum support. The concepts ‘support’ and ‘confidence’ are used to prune the search space. *Support* is a measure based on the number of occurrences, which means that it identifies the percent of transactions that contain the given pattern. *Confidence* (also known as accuracy) of a rule represents the number of transactions containing all of the items in a rule, divided by the number of transactions containing the rule antecedents. In Borgelt and Kruse (2002), a method for induction of association rules is presented and the performance of the classic apriori algorithm is optimised. In Dehaspe and Toivonen (2001), relational association rules are discovered. The process uses a relational database and the type of

patterns that are considered are SQL queries. Combinations of attribute-value pairs that have a pre-specified minimum support are *frequent item sets* (Witten and Frank, 2000). A tree projection algorithm for generation of frequent item sets is given in Agarwal et al (2001). Finally, association rules and frequent item sets are expressed by equations (2.1) and (2.2) (Goethals, 2002) below.

$$\begin{aligned}
 &\text{Association rule (AR):} && X \rightarrow Y \\
 &\text{Support (AR) in dataset } D = X \cup Y \text{ in } D \\
 &\text{Confidence (AR) in dataset } D \\
 &= [\text{support } (X \cup Y) \text{ in } D] / [\text{support } (X) \text{ in } D] \tag{2.1}
 \end{aligned}$$

$$\begin{aligned}
 &\text{Item set (IS):} && I = \{i_1, i_2, \dots, i_k\} \tag{2.2} \\
 &\text{Support (IS) in dataset } D = \text{support } (I) \text{ in } D = \text{support } (i_1 \wedge i_2 \wedge \dots \wedge i_k) \text{ in } D
 \end{aligned}$$

where X, Y and I are item sets;
 $X \cap Y = \{\}$;
 k = total number of items in item set I ;

The items in Web Usage Mining usually are represented by visited pages through their URL's. Some examples of association rules and item set, along with their support and confidence values, which resulted from the data used throughout this work, are given in table 2.4. Other practical examples are given in Cooley et al (1999a), Mobasher et al (2001), Zaiane and Luo (2001).

Association rules	Support (%)	Confidence (%)
If http://machines.hyperreal.org then http://machines.hyperreal.org/manufacturers/Moog	2.52	15.00
If http://machines.hyperreal.org then http://machines.hyperreal.org/manufacturers/Roland/TR-909	1.98	7.00
If http://machines.hyperreal.org/manufacturers/Moog then http://machines.hyperreal.org/manufacturers/Roland/TR-909	0.31	12.00
Item set	Support (%)	Confidence (%)
(http://machines.hyperreal.org , http://machines.hyperreal.org/manufacturers/Moog , http://machines.hyperreal.org/manufacturers/Roland/TR-909)	0.22	-

Table 2.4: Examples of association rules and item set in Web Usage Mining.

Association rules discovered from web usage data give an overview of the (set of) pages that are frequently visited together. These results may be applied to the optimisation of the web site through, for example, organising link structures between web pages. Also, association rules and frequent item sets may serve as a heuristic for pre-fetching documents in order to reduce loading time of pages from a remote site. Obviously, the inputs for this application are often sessions defined with the content-only technique.

2.5.1.2 Sequential Patterns

In Agrawal and Srikant (1994b) and Srikant and Agrawal (1995), sequential patterns are defined as follows. A *sequence* is an ordered list of item sets. In a set of sequences, a sequence s is maximal if s is not contained in any other sequence. Finally, every maximal sequence with a certain user-specified minimum support represents a *sequential pattern*.

In Web Usage Mining, sequential patterns find inter-session patterns in such a way that the presence of a set of pages is followed by another page in a time-ordered set of sessions or episodes (Cooley et al, 1997). Also, temporal relationships among data items in Web Usage Mining studies are presented in Cooley et al (1997) and Mannila et al (1995). The difference with association rules is that the former relates pages that are referenced together in a single session, so that they are defined as intra-session patterns. As with association rules, confidence and support values are used as thresholds in order to limit the number of rules discovered and reported. The resulting information is used to predict visiting patterns and to target advertising campaigns aimed at groups of users. Other applications of sequential patterns in Web Usage Mining are, for example, finding common characteristics of visitors who went to a particular page within a specific time period $[t_1, t_2]$. On the other hand, we may be interested in a time interval (within a day, week, month etc.) in which a particular page is frequently accessed.

In our research project, sequential patterns are represented by sequences of ordered page requests called open sequences (Capri, 2001). This is also shown in Büchner et al (1999). Examples of sequential patterns that we found in the data we used throughout our research project are given in table 2.5. Definitions of open sequences and equations for calculating support and confidence are provided in chapter four, section 4.8. The support value is equal to the number of server sessions holding the sequential pattern (open sequence) divided by the total number of server sessions in the data set. The confidence value is equal to the number of server sessions holding the sequential pattern (open sequence) divided by the number of server sessions holding all but the last element of the sequential pattern (open sequence). We remark that sequential patterns (open

sequences) are used in our research project in the final analysis step of the Web Usage Mining process (i.e. post-processing) whereas usually sequential patterns are used in the previous step (i.e. processing).

Sequential patterns	Support (%)	Confidence (%)
Visitors on the web site http://machines.hyperreal.org went to http://machines.hyperreal.org/the_Roland_TB-303 followed by http://machines.hyperreal.org/ecards	2	82
Visitors on the web site http://www.luc.ac.be/tew went to http://www.luc.ac.be/tew/opleidingen followed by http://www.luc.ac.be/tew/opleidingen/basisopleidingen/opbouw_hi	20	24

Table 2.5: Examples of sequential patterns in Web Usage Mining.

2.5.1.3 Clustering

In Fayyad et al (1996), clustering is defined as is a common descriptive task where one seeks to identify a finite set of categories or clusters to describe the data. The categories may be mutually exclusive and exhaustive, or consist of a richer representation such as hierarchical or overlapping categories. Besides, cluster analysis groups together items that have similar characteristics (Kaufman and Rousseeuw, 1990). The challenge is to find groups of items (without any predefinition) that naturally fall together, to assign instances to these groups and to be able to assign new instances to the groups or clusters. The values of attributes that measure different aspects of the instance characterize the instance (Witten and Frank, 2000).

Applied within Web Usage Mining, clustering identifies usage, transaction and page clusters. *Usage clusters* group URL references based on how often they occur together across sessions, rather than *clustering sessions* themselves. In Mobasher et al (2000) an example of usage clusters is given. Usage as well as session clusters represent groups of users having similar behavioural characteristics with regard to visited pages. Such knowledge is useful for inferring user demographics in order to perform market segmentation in E-commerce applications or to provide personalized web content to users that fall within a particular cluster. In Fu et al (1999), Nasraoui et al (1999), Shahabi et al (1997) and Yan et al (1996), user sessions are clustered to predict future user behaviour. In our research project, server sessions are clustered. Finally, *page clusters* discover groups of pages having similar content. Although page clustering actually falls within Web Content Mining we do mention it here because, as already stated in section 2.1.1 Server-level data collection, besides usage data, content data is collected as well to perform better results. For

example, in Heer and Chi (2001), user profiles are built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure and page content. Likewise, in Mobasher et al (2000), two techniques based on clustering of user and/or server sessions and clustering of pages, are presented in order to discover overlapping aggregate profiles that can be effectively used by recommender systems for real-time web personalization.

2.5.1.4 Classification

Fayyad et al (1996) describes classification as “learning a function that maps (classifies) a data item into one of several predefined classes”. Classification rules group items into a predefined profile according to their common attributes. Furthermore, new data items that are added to the database are classified into the profile. “In classification learning, a learning scheme takes a set of classified examples from which it is expected to learn a way of classifying unseen samples” (Witten and Frank, 2000).

In Web Usage Mining, classification algorithms develop a profile of users according to their demographic information or their access patterns. In Zaïane et al (1998) the WebLogMiner, a knowledge discovery tool for mining web server log files, classifies features in the web log data and generates classification rules from such models. These classification rules are used to describe each class, optimise the structure of the web site and customize answers to requests.

2.5.2 *Modifications of typical data mining methods*

Besides the typical data mining methods, modified algorithms are used to discover user profiles or to learn user navigation patterns in Web Usage Mining studies. For example, in Zaïane et al (1998) the WebLogMiner is introduced. The system interactively extracts implicit knowledge from access records by means of combining OLAP and data mining techniques with a multi-dimensional data cube. Each dimension is represented by one or more attributes. Examples of dimensions are URL, time, agent, user, server status etc. The multi-dimensional structure of the data cube provides a way to view the data from different perspectives. The strengths of WebLogMiner are scalability, interactivity, variety and flexibility. In Borges and Levene (2000a) user navigation data is modelled as a hypertext probabilistic grammar (HPG). HPG generates probability strings. The highest probability strings generated by the grammar correspond to the user preferred navigation trails. To deal with the drawbacks of returning a large number of rules when the cut-point is small and a small set of rules when the cut-point is high, a new heuristic is introduced. By

setting the value of the stopping criterion, the analyst can determine the number and the quality of rules. Another example of modified data mining techniques for Web Usage Mining studies is the Web Site Information Filter (WebSIFT) system (Cooley et al, 1999b). WebSIFT applies an interestingness measure to data mining methods in order to automatically discover interesting rules and patterns from web usage behaviour. Other examples are given in Kosala and Blockeel (2000). Finally, in our research project, server sessions are clustered using matrices of SAM distance measures instead of using the typical distance measures of clustering methods. More details about SAM and clustering based on SAM matrices are given in chapter three and four. Also, in chapter five, an interestingness measure for Web Usage Mining studies is integrated with SAM. SAM calculates distance measures between interesting combinations of pages in server sessions in order to discover interesting navigations.

For the analysis of web usage behaviour, often several typical data mining techniques, which are presented in different categories in section 2.5.1, are combined to obtain the desired information. Also, not only usage data but also content and topology data are used. Illustrations are given in Cooley et al (1999a) and Zaïane and Luo (2001). In our research project, content pages are distinguished from navigation pages in order to define categories of visiting page times. Also, topology data of the web site structure is used to provide information for extracting interesting navigations from non-interesting navigations. More details are given in chapter four and five.

Finally, statistical techniques provide a general insight into the data in order to 'get to know your data' before the actual analysis occurs. Examples of information provided by statistical analysis are frequency of page-clicks, average page-clicks per user or server session, longest/shortest user or server session, average viewing time, most frequently accessed pages etc. In Cooley et al (1999a), the WEBMINER system uses statistical techniques in combination with clustering, association mining and sequential pattern mining to provide interesting rules, patterns and statistics. Some examples of statistical information used throughout this thesis for describing data sets are: total number of server sessions, total number of distinct URL addresses (i.e. total number of distinct web pages), shortest/longest server session, server sessions' average length, total number of requested web pages, distribution of the length of server sessions, average visiting page times, etc.

2.6 Post-processing

In order to apply the results of the pattern discovery process of Web Usage Mining, they first have to be understood and interpreted. Therefore, the final process of Web Usage Mining, called *pattern analysis* or *pattern evaluation*, must be fulfilled. The discovered patterns are analysed using different techniques: visualisation tools, OLAP and data & knowledge querying.

2.6.1 Visualization tools

Visualization tools present graphically how the users visit the web site. The web is generally visualised as a directed graph with cycles, where nodes are represented as pages and (inter-page) hyperlinks are depicted as edges. Pitkow and Bharat (1994) have developed the Webviz system for visualising World Wide Web access patterns. Also, Spiliopoulou and Faulstich (1998) use graphical techniques for analysing discovered patterns. They developed a Web Utilization Miner (WUM) where aggregate trees and navigation patterns are drawn in a graph. Furthermore, in Kato et al (2000) a visualization technique using a polar coordinate system is introduced to assist web publishers in pattern analysis. Figure 2.1 illustrates how the polar coordinate system is represented by circles sharing the same centre (or origin). The inner circle has the smallest radius; the outer circle the largest. The analyst-selected target page (page A) is plotted at the origin (0,0) of the circle. User behaviour to and from the target page is shown by the pages plotted on the circles (pages B, C and D), with the number of visited pages increasing and the ratio of the number of users who visited the pages decreasing in proportion to the distance from the target page.

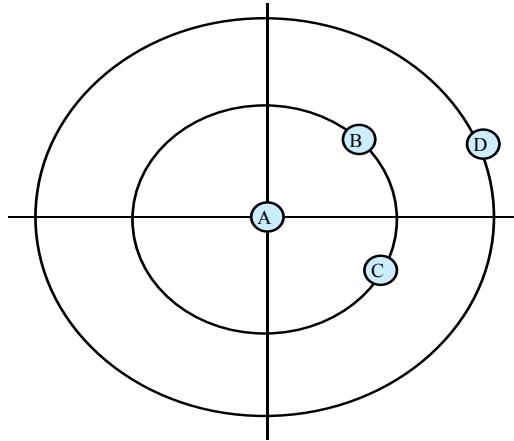


Figure 2.1: Polar coordinate system for user path visualization.

Another technique for visualizing information of web usage patterns is presented in Chi et al (2000). By means of the Dome Tree visualization, user paths are embedded in 3D and the analyst can peer into the Dome through the opening like a door. Also, path crossings are minimized in order to provide an overview of web usage paths.

In our research project navigation patterns are visualized by means of clusters of server sessions. Each cluster graphically presents the web site structure (i.e. pages and direct hyperlinks between pages) and navigations on the web site. For each cluster, open sequences with high support values are selected. The navigation patterns, represented by the selected open sequences, are drawn in the graph by means of arrows between pages. More details are provided in chapter four.

2.6.2 OLAP

OLAP stands for On-Line Analytical Processing. Pendse (2003) summarizes the OLAP definition in just five key words: Fast Analysis of Shared Multidimensional Information, or FASMI for short. OLAP is capable of analysing multi dimensional data, for example, time series and trend analysis views (Webopedia, 2002). Pattern analysis of web usage data by means of OLAP techniques is described in Zaïane et al (1998). By means of a data cube structure, the intersection of three dimensions of interest, for example traffic, weekend days and user agent, is generated interactively. The results are presented through summarization of the selected data cells. In Büchner and Mulvenna (1998), marketing intelligence is discovered through online analytical web usage mining.

2.6.3 Data & knowledge querying

A query is a request for information from a database (Webopedia, 2000). With a query language an application is allowed to express what conditions must be satisfied by the data it needs, rather than having to specify how to get the required data. In Cooley et al (1999a) an SQL-like querying language is used by the WEBMINER system. Spiliopoulou and Faulstich (1998) integrate a MINT query into the WUM system in order to obtain simple statistical information and new aggregate trees, to combine aggregate trees into a navigation pattern and to look for information about revisited nodes.

CHAPTER 3

SEQUENCE ALIGNMENT METHOD (SAM)

The Sequence Alignment Method (SAM) originated in biological studies to compare macromolecules. Within this research area it was important to develop a method that handles sequences of variable length and incorporates a measure for the order in which elements occur in sequences. Moreover, biological studies need to assign different costs for different types of work when equalizing one sequence with another. More precisely, SAM is developed to measure distances between sequences. Yet, some differences between SAM and mathematical distance measures may occur. It is important to recognize for which parameter settings SAM distance differs from mathematical distance. In preliminary studies, SAM is applied to sequences in Web Usage Mining studies in Hay et al (2001a).

Calculating SAM distance measures between sequences is a combinatory problem and dynamic programming is used for SAM calculations. Besides SAM, also multi-dimensional SAM (MDSAM) is presented. Descriptions of SAM and MDSAM illustrated with examples in Web Usage Mining studies are provided in Hay et al (2001b, 2002a, 2002b, 2003b, 2003c, 2003d). While SAM calculates distances between sequences consisting of one attribute, MDSAM calculates distances between sequences consisting of more than one attribute, without losing the characteristics of SAM. Moreover, MDSAM takes into account the inter-attribute relationships. An attribute or dimension is defined as a particular type of information in a sequence. For example, visited pages and visiting times are two different attributes within Web Usage Mining studies. Also, particular relations between visited pages and visiting times may occur, which is recognized by MDSAM. Finally, heuristic algorithms based on dynamic programming and genetic algorithms are given to calculate MDSAM between multi-dimensional sequences.

3.1 Biological background

Originally, the Sequence Alignment Method (SAM), also known as *edit distance*, *alignment distance* or *Levenshtein distance* (Sankoff and Kruskal, 1983), was developed in biology to obtain knowledge with regard to homology (i.e. correspondence) of macromolecules. These macromolecules are considered as long sequences of subunits linked together sequentially in a chain. Examples of such macromolecules are DNA or RNA sequences. SAM measures the biological distances between such sequences. Essentially, the amount of work required to equalize two sequences of information is treated as a measure of distance. The method deals with common characteristics of biological sequences. This means that it handles *variable-length sequences* and incorporates *sequential information*, i.e. the order in which elements appear in a sequence, into its distance measure. It also gives opportunities to treat some mutations as more unlikely than others, using different *weights* for different types of work during the equalization process. In Sankoff and Kruskal (1983) an overview is given of pattern recognition in macromolecular sequences. In chapter six, the effects on the results of using different weights are analyzed for web usage mining studies.

3.2 Calculating SAM distance: the basic algorithm

In general, the distance or similarity between two sequences, based on SAM, is reflected by the number of operations necessary to convert one sequence into the other. As a result, SAM distance is represented by a score. The higher/lower the score, the more/less effort it takes to equalize sequences and the less/more similar sequences are. In addition, SAM scores for the following operations during the equalization process: *Deletion* and *insertion* operations are applied to elements of the source (first) sequence in order to change the source into the target (second) sequence. *Substitution* operations indicate deletion + insertion. Note that throughout this thesis, operations are applied to the source sequence in order to change (equalize) the source into (with) the target. This way, additional complexities about the method are avoided and uniform procedures are followed in every chapter. Finally, SAM represents the minimum cost (optimal distance) for equalizing two sequences.

In particular, SAM distance measure between two sequences $S_1 = s_{11}, s_{12}, \dots, s_{1m}$ and $S_2 = s_{21}, s_{22}, \dots, s_{2n}$ is calculated using the following formula (Sankoff and Kruskal, 1983):

$$d_{\text{SAM}}(S_1, S_2) = \min [(w_d D + w_i I) + w_s S] \quad (3.1)$$

where

- d_{SAM} is the distance between two sequences S_1 and S_2 , based on SAM;
- w_d is the weight value for the deletion operations, a positive constant not equal to 0, determined by the researcher ($w_d > 0$);
- w_i is the weight value for the insertion operations, a positive constant not equal to 0, determined by the researcher ($w_i > 0$);
- w_s is the weight value for the substitution operations and equals $w_d + w_i$;
- D is the number of deletion operations;
- I is the number of insertion operations;
- S is the number of substitution operations;

and

- m is the length of the first sequence (source);
- n is the length of the second sequence (target);
- s_{ij} is an element, representing a particular character, of a sequence;
- i identifies the sequence number, $i = 1, \dots, N$;
- N is the total number of sequences in the analysis;
- j identifies the position in a sequence, $j = 1, \dots, m$ or $j = 1, \dots, n$;

Equation (3.1) indicates that the score, represented by SAM distance measure between two sequences, consists of the minimum costs for deleting, inserting and substituting elements.

SAM is illustrated by means of four examples in different domains. Table 3.1 presents sequences used within biology, human speech and time use studies. We also provide a preliminary example how we will use SAM in web usage mining studies. For each sequence pair, SAM distance is given along with the operations that are necessary to convert the source into the target. The weight values used to calculate SAM are 1 for insertion/deletion and 2 for substitution.

Domain	Sequence pair		SAM distance $w_i = 1$ $w_d = 1$ $w_s = 2$	Operation		
				I	D	S
Biology	Source	AACAAA	1	0	1	0
	Target	AAAAA				
Human Speech	Source	INDUSTRY	8	3	3	1
	Target	INTEREST				
Time use studies	Source	breakfast work shopping dinner sport	4	1	3	0
	Target	work housekeeping dinner				
Web usage mining	Source	page x page y	1	1	0	0
	Target	page x page y page z				

Table 3.1: Sequence comparison based on SAM.

Practically, we provide an algorithm to structure the equalizing process between two sequences in a fast and easy way. For illustrations we use the examples given in table 3.1. We remark that it has not yet been proven that the following three steps always lead to an optimum solution. Yet, they mostly do.

Step 1 Identify maximum identities or the longest common sub strings respecting the sequential order of elements.

 Example: Sequence pairs

Longest common sub strings

A A C A A A
 | | / / /
 A A A A A

AA-AAA

I N D U S T R Y I N D U S T R Y
 | | / / / | | / / /
 I N T E R E S T I N T E R E S T

IN-ST or IN-TR

breakfast work shopping dinner sport

work-dinner

work housekeeping dinner

page x page y

page x - page y

page x page y page z

Step 2 Identify elements, which are not included in the sub string and appear in the source and target sequence. Count one substitution (= deletion + insertion) operation for each such identified element.

Example: Sequence pairs

```

A A C A A A
| | / / /
A A A A A

```

```

I N D U S T R Y I N D U S T R Y
| | / / / / /
I N T E R E S T I N T E R E S T

```

```

breakfast work shopping dinner sport
/ / / / /
work housekeeping dinner

```

```

page x page y
| |
page x page y page z

```

Elements not included in sub string and appearing in source and target sequence

None (0 substitution)

R (1 substitution)
or S (1 substitution)

None (0 substitution)

None (0 substitution)

At the end of this step, the order of substituted elements has been changed. In the example above, R is changed in the source sequence in order to obtain the same order of elements as in the target sequence. Likewise, if IN-TR is chosen as longest common subsequence, S changes places coming after TR in the source sequence. In the following step the results are given for changes made in the source sequence.

Step 3 Identify elements, which are not included in the sub string and appear in either one of the compared sequences. Count one deletion operation for each element found in the source sequence. Count one insertion operation for each element found in the target sequence. As a matter of fact, elements found in the source sequence are 'deleted' from the source sequence; elements found in the target sequence are 'inserted' into the source sequence, respecting the positions of the elements.

 Example: Sequence pairs

A A ~~C~~ A A A
 | | / / /
 A A A A A

 T E E E E T
 I N ~~D~~ ~~U~~ **R** S T ~~Y~~ I N ~~D~~ ~~U~~ T R **S** ~~Y~~
 | | \ \ | | / / /
 I N T E **R** E S T I N T E R E **S** T

 housekeeping
 breakfast work ~~shopping~~ dinner sport
 / / /
 work housekeeping dinner

page x page y page z
 | | |
 page x page y page z

Elements not included in sub string and appearing in either one of the compared sequences

in source: C (1 deletion)
 in target: none (0 insertion)

in source: D, U, Y (3 deletions)
 in target: T, E, E (3 insertions)

in source: breakfast, shopping, sport (3 deletions)
 in target: housekeeping (1 insertion)

in source: -
 in target: page z (1 insertion)

3.3 Mathematical distance versus SAM distance

In mathematical science, the word ‘distance’ is used to indicate a function d , which satisfies the following *metric axioms* between the points a , b and c (Sankoff and Kruskal, 1983):

- nonnegative property: $d(a, b) \geq 0$;
- zero property: $d(a, b) = 0$ if $a = b$;
- triangle inequality: $d(a, b) + d(b, c) \geq d(a, c)$;
- symmetry: $d(a, b) = d(b, a)$;

Comparing the metric axioms illustrated above with the way we use SAM distance, the first three metric axioms always hold, while the fourth metric axiom does not always hold. This is illustrated in table 3.2, using S_1 , S_2 and S_3 representing three sequences within web usage mining studies.

SAM distance between sequence pairs	$S_1 = \text{page x page y}$ $S_2 = \text{page x page y page z}$ $S_3 = \text{page y page x}$	
	$w_d = w_i = 1 ; w_s = 2$	$w_d = 1 ; w_i = 2 ; w_s = 3$
$d(S_1, S_2)$	1	2
$d(S_2, S_1)$	1	1
$d(S_1, S_3)$	2	3
$d(S_3, S_1)$	2	3
$d(S_2, S_3)$	3	4
$d(S_3, S_2)$	3	5

Table 3.2: Comparing SAM distance with metric axioms of mathematical distance.

Dependant of the weights for deletions, insertions and substitutions, SAM distance measures are not always symmetric. If equal weight values for deletion and insertion are used, SAM satisfies all of the mathematical metric axioms of a distance measure. For example, in column two of table 3.2, weight values are $w_d = w_i = 1$. Yet, if unequal weight values for deletion and insertion are used, SAM satisfies all but one of the axioms. SAM becomes an asymmetric distance measure. For example, in column three of table 3.2, weight values are $w_d = 1, w_i = 2$. Therefore we remark that, throughout this thesis, if we use the term ‘distance’ we refer to SAM distance and not to mathematical distance measures, unless otherwise mentioned. In chapter five and seven, equal weight values of insertion and deletion with substitution = insertion + deletion are used. In chapter six, equal as well as unequal weight

values of insertion and deletion with substitution =, > or < insertion + deletion are used in order to investigate how sensitive the results are towards changes in parameter values of SAM. Although it is generally desirable to use a function d satisfying all of the metric axioms mentioned above, exceptions are given in Sankoff et al (1983). The use of SAM as an asymmetric function in speech recognition and without zero property in telecommunications is stated. Within Web Usage Mining studies, the use of SAM as an asymmetric function is discussed in chapter six.

3.4 Dynamic programming

3.4.1 Combinatory problem

Calculation of SAM distance measures between sequences is a combinatory problem, due to many different possible trajectories. A *trajectory*, also known as array (Sankoff and Kruskal, 1983), is a path of equalizations between two sequences (Joh et al, 2001) that may be optimal (i.e. representing the minimum total operational costs) or not. To understand the nature of this problem, we first summarize the basic SAM algorithm in figure 3.1. *Common elements* are elements appearing in both of the compared sequences whereas *unique elements* appear in either one of them. Following, we present a dynamic programming model and illustrate operational efforts using a comparison table. The *comparison table*, also called computational array, is a two-dimensional table representing the elements of source and target sequences and trajectories between them, shown by a set of moves.

```
begin
read source sequence           //first sequence of sequence pair//
read target sequence          //second sequence of sequence pair//
calculate maximum identity     //identity = common elements occurring in the same
                               order//

define other common elements
define unique elements
calculate SAM cost between source and target sequence
write source sequence, target sequence, SAM cost
end;
```

Figure 3.1: Summarization of basic SAM algorithm.

In practical applications, dynamic programming algorithms are used to resolve combinatory problems (Joh et al., 2001; Mannila and Ronkainen, 1997; Wilson, 1998). *Dynamic programming* is a mathematical programming technique for making interrelated decisions. It provides a systematic procedure for determining the optimal combination of decisions (Hillier and Lieberman, 1990). In Silver and Peterson (1985), dynamic programming is described as a mathematical procedure for solving sequential decision problems, where the outcome of the decision at one point has effect on the outcome at later decision points. The combinatory problem of SAM is mathematically presented by means of a dynamic programming model, given by equation (3.2) below.

$$d(S_1, S_2) = \text{cell}(m, n) \quad (3.2)$$

where

$m = \text{length of } S_1;$

$n = \text{length of } S_2;$

$\text{cell}(0, 0) = 0;$

$\text{cell}(i, 0) = \text{cell}(i-1, 0) + \text{cost};$

$\text{cell}(0, j) = \text{cell}(0, j-1) + \text{cost};$

$\text{cell}(i, j) = \min [\text{cell}(i-1, j), \text{cell}(i, j-1), \text{cell}(i-1, j-1)] + \text{cost};$

and

$i = 1, \dots, m;$

$j = 1, \dots, n;$

$\text{cost} = 0$ if $s_{1i} = s_{2j}$ and $i = j;$

$\text{cost} = w_d$ if $i > j;$

$\text{cost} = w_i$ if $j > i;$

In a comparison table, elements of S_1 (i.e. $s_{1i} = s_{11}, s_{12}, \dots, s_{1m}$) are written vertically and elements of S_2 (i.e. $s_{2j} = s_{21}, s_{22}, \dots, s_{2n}$) are written horizontally. The cells between S_1 and S_2 in the comparison table are systematically filled with numbers, starting from cell (0, 0) and ending with cell (m, n). Cells (i, 0) and (0, j) represent the margin cells of the first column and the first row in the comparison table. In chapter four and five, the basic SAM algorithm presented in figure 3.1 along with equation (3.2) are applied to real data sets consisting of server sessions providing information about visited pages on web sites.

The following sequence pair illustrates the combinatory problem of SAM by means of comparison table 3.3. General examples of sequences in web usage mining studies are given where elements in sequences are represented by visited web pages, ordered sequentially. Consider two sequences S_1 (source) = page x page y page z page x page w and S_2 (target) = page x page y page z

page x . S_1 is presented vertically, S_2 horizontally in the comparison table. Also, positions of elements are written next to or above the sequences. In a comparison table, each sequence starts with a blank representing a null element at position 0. Each cell (i, j) , with $i = 0, \dots, m$ and $j = 0, \dots, n$; $m = \text{length of source}$ and $n = \text{length of target}$, represents the optimal (= minimum) equalization cost or SAM distance between the elements at positions up to i of the source and the elements at positions up to j of the target. For example, cell $(3, 2)$ represents the equalization cost between – page x page y page z and – page x page y . Indeed, one deletion operation of element ‘page z ’ in the source equalizes these sequences, leading to an equalization cost of 1. Proceeding downwards to cell $(4, 2)$, an equalization cost of 2 is given for equalizing – page x page y page z page x and – page x page y . Obviously, two deletion operations of elements ‘page z ’ and ‘page x ’ in the source change the source into the target. Also cell $(4, 3)$ gives an equalization cost of 3 as distance measure between – page x page y page z page x and – page x page y page w . In this case, optimal distance is obtained by deleting elements ‘page z ’ and ‘page x ’ in the source and inserting element ‘page w ’ into the source. Ultimately, the final SAM distance measure between S_1 and S_2 is given in cell $(5, 4)$.

Besides cell values representing minimum equalization costs, also the types of operations can be read from a comparison table through trajectories. The equalization process starts at cell $(0, 0)$ and ends at cell $(5, 4)$. Each operation is represented by moves in the table. A *horizontal move* indicates an insertion, a *vertical move* represents a deletion and a *diagonal move* stands for an identity (if elements are equal) or substitution (if elements are not equal). The *optimal path*, showing the types of operations that lead to the total minimum equalization cost, is found in the comparison table through backtracking. Now the opposite direction of the equalization process is followed, starting at cell $(5, 4)$ and ending at cell $(0, 0)$. Each time the minimum cell value is chosen of the cell above, left and above-left. The direction that is chosen identifies the operation type. For example, starting at cell $(5, 4)$, the cell values above, to the left and above-left are respectively 2, 2 and 3. We can step above or step left. If we step left, a horizontal move occurs (1 insertion). Now we are at cell $(5, 3)$ from where the values of the cells above, left and above-left are now respectively 3, 3, 2. This means that we have to move diagonally to cell $(4, 2)$. This move does not change the equalization costs, which means that an identity occurs. Furthermore, the optimal path proceeds to cell $(3, 2)$, identifying 1 deletion, cell $(2, 2)$, again 1 deletion and finally to cells $(1, 1)$ and $(0, 0)$. Ultimately, the operation types leading to the total minimum equalization costs in cell $(5, 4)$ are one insertion and two deletions. More precisely, looking at s_1 and s_2 , element ‘page w ’ is inserted into the source at position 3, element ‘page z ’ is deleted from the source at position 3 and finally element ‘page w ’ is

deleted from the source at position 5. Note that a deletion and insertion of the same element affecting the same sequence is the same as one *reordering operation* (Joh et al, 2001a). In fact, we may also say that one reordering and one deletion are the operations leading to the total minimum equalization cost. In this research project, we will use the term reordering (or substitution) if the order of common elements is changed. The parameter value or operational weight for reordering will be denoted as “ η ”. Yet, reordering is used in Joh et al (2001a) in a position-sensitive SAM algorithm. We apply SAM’s basic operations. This means that, if we mention reordering (or substitution) we do not include the number of positions over which changes in order of elements occur into our measurement. Finally, during an equalization process, several optimal paths may occur. Remark that optimal paths are trajectories. However, a trajectory is not always an optimal path.

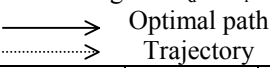
Operation weights: $w_d = w_i = 1$; $w_s = w_d + w_i$						
						
Position	Target	0	1	2	3	4
Source	Element	-	Page x	page y	page w	page x
0	-	0	1	2	3	4
1	page x	1	0	1	2	3
2	page y	2	1	0	1	2
3	page z	3	2	1	2	3
4	page x	4	3	2	3	2
5	page w	5	4	3	2	3

Table 3.3: Comparison table.

3.4.2 Advantages of dynamic programming

The dynamic programming method is preferable to other methods because of several reasons (Kruskal and Sankoff, 1983).

1. Separation of evaluation from algorithm. Dynamic programming permits the user to understand the evaluation system (through for example a comparison table), without studying the algorithm. Moreover, the user may change the evaluation system (through for example using other operational weights than the most elementary ones, also known as the default values for parameters in SAM-dynamic programming presented in table 3.3), without changing the algorithm.

2. Global optimality. Dynamic programming guarantees to find the best optimal path (or several best optimal paths) out of all possible trajectories. This means that the optimal path(s) with the best (minimum) equalization score is (are) found. Many other methods do not link their algorithm to an equalization score. Neither do they guarantee to find the optimal path.
3. Stable parameters. In some methods, it is frequently necessary to adjust the parameters for new comparisons and considerable expertise may be required to make proper choices. Likewise, the result is heavily dependent on these parameter settings. However, through dynamic programming sensible results are obtained by using the default values, which are all equal to 1 (re. chapter six).

Yet, despite these advantages, dynamic programming is also associated with a disadvantage (Hillier and Lieberman, 1990). Computing time may become extremely large due to the number of calculations or possibilities to examine.

3.5 Multi-dimensional SAM (MDSAM)

In order to measure similarities between sequences based on more than one attribute the multi-dimensional SAM or MDSAM is used. We define an *attribute* or *dimension* as a particular type of information in a sequence. Examples of different attributes in Web Usage Mining studies are pages (URL-addresses), time spent on pages, page type, page category, web structure level etc. By means of MDSAM, multi-dimensional or multivariate web usage patterns are analyzed on such typical attributes.

MDSAM is developed by Joh et al (2001) and is concerned with differences in element composition, sequential order of elements and handles inter-attribute relationships. An *element* of a sequence represents the lowest level of information and belongs to (or describes) a particular attribute. Examples of elements used in this research project are page1 (www.luc.ac.be/tew), page2 (www.luc.ac.be/tew/diensten/diensten_voor_students), 3 seconds, 1 minute etc. Most multi-dimensional distance measures sum one-dimensional distance measures across several attributes to produce an overall measure of similarity (Boyle and Flowerdew 1997; Les and Maher 1998; Murray 2000). However, this ignores correlations between attributes. For example, the choice between page and time spent on page is probably interrelated due to the fact that visitors spent short times on navigation pages and longer times on content pages.

Therefore, any valid similarity measure should incorporate such interdependencies in the quantification of the degree of similarity. This characteristic distinguishes MDSAM from other multi-dimensional distance measures.

Basically, MDSAM is a multi-dimensional extension of the one-dimensional SAM. This is explained in section 3.5.1. To overcome the problems of combinatorial explosion (re. section 3.5.2), Joh et al (2001) generated acceptable solutions by means of alternative heuristics based on dynamic programming and genetic algorithms, while maintaining the characteristics of MDSAM. More details of the approach are given in section 3.5.3.

3.5.1 Multi-dimensional sequence comparison

Multi-dimensional sequences are represented by K attribute sequences. Each attribute sequence consists of a set of elements. A pair of K -dimensional sequences consisting of m and n positions is represented by $K \times m$ and $K \times n$ matrices of qualitative elements. The problem of comparing these two sequences is to measure the effort required to equalize the two matrices. In particular, the equalization costs for two one-dimensional sequences is defined as the minimum sum of weights for deletion, insertion and substitution operations that are required to equalize the sequences (Gusfield, 1997; Sankoff and Kruskal, 1983; Waterman 1995). The problem that arises in multi-dimensional analysis of sequences is that the equalization costs are not equal to a simple sum of one-dimensional costs because of the interrelations between attributes. In other words, a variety of interdependency relationships between attributes complicate the problem of calculating the minimum-effort equalization.

3.5.2 Combinatory problem

The same operations applied to elements at the same positions across attributes require a multi-dimensionally integrated operational weight that is smaller than the simple sum of one-dimensional weights of the operations involved, implying interdependency between attributes. Therefore, *MDSAM bundles operations of the same kind that are applied to the same positions into the same sequences, across attributes, into a single operation*. However, combinatorial explosion occurs due to the fact that a multitude of ways can be envisioned to align one-dimensional sequences. To illustrate the combinatory problem of MDSAM, consider the following two-dimensional sequences.

$S_1 =$	page x	page y	page z	page w	(attribute 1)	
	a	b	c	d	(attribute 2)	
$S_2 =$	page x	page t	page u	page y	page v	(attribute 1)
	a	c	d	b	a	(attribute 2)

suppose: $w_d = w_i = 1$ and $\eta = \text{reordering} = w_d + w_i$

For each sequence pair, a *trajectory* contains the kind of operations (d = deletion; i = insertion) as well as the positions and the sequences that are affected by the operations. During the equalization process, trajectories are composed for each attribute. Some trajectories represent the optimal path or minimum costs (i.e. one-dimensional SAM), others are non-optimal. However, due to the interdependencies between attributes, all possible trajectories must be evaluated to calculate the total minimum cost, which is MDSAM distance between S_1 and S_2 .

trajectory = {d3S ₁ , d4S ₁ , i2S ₁ , i3S ₁ , i5S ₁ }	-optimal-	(attribute 1)
trajectory = {d3S ₁ , d4S ₁ , d2S ₂ , d3S ₂ , d5S ₂ }	-optimal-	
trajectory = {i3S ₂ , i4S ₂ , d2S ₂ , d3S ₂ , d5S ₂ }	-optimal-	
trajectory = {i3S ₂ , i4S ₂ , i2S ₁ , i3S ₁ , i5S ₁ }	-optimal-	
trajectory = {i2S ₂ , i4S ₁ , d3S ₁ , d4S ₁ , d2S ₂ , d3S ₂ , d5S ₂ }	-non-optimal-	
trajectory = {d2S ₁ , d4S ₂ , d3S ₁ , d4S ₁ , d2S ₂ , d3S ₂ , d5S ₂ }	-non-optimal-	
...		
trajectory = {d2S ₁ , d4S ₂ , d5S ₂ }	-optimal-	(attribute 2)
trajectory = {d2S ₁ , i4S ₁ , d5S ₂ }	-optimal-	
trajectory = {d2S ₁ , i4S ₁ , i5S ₁ }	-optimal-	
trajectory = {d2S ₁ , d4S ₂ , i5S ₁ }	-optimal-	
trajectory = {i4S ₁ , i2S ₂ , d5S ₂ }	-optimal-	
trajectory = {i4S ₁ , i2S ₂ , i5S ₁ }	-optimal-	
trajectory = {d3S ₁ , d4S ₁ , d2S ₂ , d3S ₂ , d5S ₂ }	-non-optimal-	
trajectory = {d3S ₁ , d4S ₁ , i2S ₁ , i3S ₁ , d5S ₂ }	-non-optimal-	
trajectory = {d3S ₁ , d4S ₁ , i2S ₁ , i3S ₁ , i5S ₁ }	-non-optimal-	
...		

The first trajectory of attribute 1 combined with the ninth trajectory of attribute 2 provides the minimum total equalization cost between S_1 and S_2 , because 5 operations could be bundled up into single operations:

trajectory = {d3S ₁ , d4S ₁ , i2S ₁ , i3S ₁ , i5S ₁ }	-optimal-	(attribute 1)
trajectory = {d3S ₁ , d4S ₁ , i2S ₁ , i3S ₁ , i5S ₁ }	-non-optimal-	(attribute 2)

Ultimately, MDSAM distance between S_1 and S_2 equals 5. This example shows that also non-optimal one-dimensional costs may result into optimal multi-dimensional costs.

3.5.3 Heuristics

Enumerating all possible solutions to find MDSAM and guarantee optimality is, due to combinatorial explosion and in terms of computing time, not a realistic approach. Therefore Joh et al (2001) introduced heuristics based on dynamic programming and genetic algorithms to compute near optimal solutions within acceptable computing times.

3.5.3.1 A heuristic based on genetic algorithms

Genetic algorithms are modeled in analogue to evolutionary processes of biological species. What makes genetic algorithms particularly interesting is that they do not search the entire space of a possibly infinite number of solution candidates, but reduce the solution search space by considering populations. Each time a new population is created genetically from the old one, resulting in better fitness values. Finally, a near optimal solution is found by means of a stop condition (e.g. fitness value of several populations during consecutive generations does not improve).

A heuristic algorithm for MDSAM, based on genetic algorithms, is developed as follows (Joh et al, 2001). First of all, some terms are defined. A set of *moves* in the comparison table constitutes a *trajectory*, a set of trajectories is called a *trajectory set*. A trajectory set represents K one-dimensional trajectories (K = number of attributes). Ultimately, a set of trajectory sets constitutes a *population* of the current *generation*.

Trajectory sets are selected based on their fitness values. The fitness value of a trajectory set is the sum of the costs for the insertion and deletion operations included in the multi-dimensional operation sets. The lower the multi-dimensional alignment cost, the better the fitness value of a trajectory set. The employed genetic algorithms start randomly with the population of the 0th generation and probabilistically select the trajectory sets for generating new populations in proportion to their fitness values. The *selection probability* $S_t(u)$ of a trajectory set u for the population of the t^{th} generation is expressed in equation (3.3).

$$S_t(u) = \frac{\sum_{u'=1}^U C^{u'}}{C^u} \quad (3.3)$$

$$\sum_{u'=1}^U \left(\frac{\sum_{u''=1}^U C^{u''}}{C^{u''}} \right)$$

where

C = fitness value;

U = total number of trajectory sets within the population of the t^{th} generation;

$u = u^{\text{th}}$ trajectory set within the population of the t^{th} generation;

and

$1 \leq u \leq U$;

$u' = 1, \dots, u, \dots, U$;

$1 \leq U \leq \prod_{k=1}^K A_k$;

K = total number of attributes;

A_k = total number of trajectories for attribute k ;

Equation (3.3) says that the probability of the u^{th} trajectory set selection is defined as the goodness-of-fit of that particular u^{th} trajectory set (numerator) and the sum of goodness-of-fit of all trajectory sets (denominator) within the population of the t^{th} generation. The selection probability of the u^{th} trajectory set is also called the *survival rate*.

The population of the next generation is created by means of the following *genetic operators*: reproduction, crossover and mutation (Mena, 1999). Reproduction is the process by which a program evaluates and copies strings according to the desired output. When crossover occurs, two strings exchange information that yields new combinations. Mutation is a source of variation used to maintain diversity in a population. Figure 3.2 summarizes the MDSAM heuristic based on genetic algorithms.

```

begin
   $t=0$  //  $t$  indicates the  $t^{\text{th}}$  generation //
   $no\_improve = 0$ 
  initialize  $E(t)$  //  $E(t)$  is the population //
  calculate  $C^\circ(t)$  //  $C^\circ(t)$  is the fitness of  $E(t)$  //
  Best_Fitness =  $C^\circ(t)$ 
  while not ( $no\_improve \geq convergence\_rate$ ) do
    begin
      select a genetic operator
      create  $E_1(t)$  by selecting and copying a subset of  $E(t)$ 
       $t = t+1$ 
       $no\_improve = no\_improve + 1$ 
      create  $E_1(t)$  by applying the selected genetic operator to  $E_1(t-1)$ 
      create  $E_2(t)$  by selecting and copying a subset of  $E(t-1)$ 
      create  $E(t)$  by summing  $E_1(t)$  and  $E_2(t)$ 
      calculate  $C^\circ(t)$ 
      if  $C^\circ(t) < Best\_Fitness$  then Best_Fitness =  $C^\circ(t)$  and  $no\_improve = 0$ 
    end
  end

```

Figure 3.2: Summarization of the MDSAM heuristic based on genetic algorithms.

3.5.3.2 A heuristic based on dynamic programming

In theory, the one-dimensional optimal trajectories do not always provide the optimal multi-dimensional solution, as discussed in section 3.5.2. Yet, in practice we may expect that the integration of one-dimensional optimal trajectories will provide a solution that is near to the multi-dimensional optimum due to the fact that optimal trajectories involve the largest number of cost-free identities (Joh et al, 2001). However, often many one-dimensional optimum trajectories occur and the number of combinations across attributes to consider will cause combinatorial explosion. Therefore, Joh et al (2001) provide a heuristic based on dynamic programming, which considers for each attribute only one optimum trajectory along (or at) the diagonal region of the comparison table. Besides, Sankoff and Kruskal (1983) and States and Boguski (1991) have proven that most one-dimensional optimum trajectories run along the diagonal region of the comparison table. Equation (3.4) shows how optimum trajectories along (or at) the diagonal regions are found (Joh et al, 2001).

$$F(Q_{vk}) = \sum_I^{Rk} e_{rv} \quad (3.4)$$

where

e_{rv}	is a dichotomous value denoting whether the coordinate of the r^{th} identity operation of the v^{th} optimum trajectory falls in the diagonal region; = 1 if $e(i, j, k)_{rv} \in D\{e(i, j, k)\}$ = 0 otherwise;
$e(i, j, k)_r$	is the r^{th} identity operation applied to the i^{th} source element and the j^{th} target element of the k^{th} attribute;
$e(i, j, k)_{rv}$	is $e(i, j, k)_r$ of the v^{th} optimum trajectory;
R_k	is the total number of identity operations of the optimum trajectory of the k^{th} attribute;
Q_{vk}	is the identity operation set of the v^{th} optimum trajectory of the k^{th} attribute;
$F(Q_{vk})$	is a diagonal function measuring how much Q_{vk} is involved with the diagonal region of the k^{th} attribute, denoted by $D\{e(i, j, k)\}$;

and

$F(Q_k) = F(Q_{vk}^\circ) = \max [F(Q_{1k}), \dots, F(Q_{vk}), \dots, F(Q_{Rk})]$;	
Q_k	is the cost-free identity operation set of an optimum alignment of the k^{th} attribute;
V_k	is the number of optimum trajectories that can be traced in the comparison table of the k^{th} attribute;
$Q_k = \{q q = e(i, j, k)_1, \dots, e(i, j, k)_r, \dots, e(i, j, k)_{Rk}\}$;	
O_k	is the cost-taking deletion and insertion operation set of the k^{th} attribute;
$O_k = \text{conv}(Q_k) = \{p p = d(i, k) \vee i(j, k)\}$;	
$\text{conv}(Q_k)$	is a procedural function that converts Q_k into O_k ;
$d(i, k), i(j, k)$	are the deletion and insertion operations applied to the i^{th} and j^{th} elements of the k^{th} attribute;

More specifically, equation (3.5) defines the value of e_{rv} (Joh et al, 2001):

$$e_{rv} = \begin{cases} = 1 & \text{if } p \leq q \leq (|m-n| + p) \\ = 0 & \text{otherwise} \end{cases} \quad (3.5)$$

where

p, q	are the positions of the shorter pattern and the longer pattern appeared in $e(i, j, k)_{rv}$ respectively;
--------	---

Finally, Q_k can easily be converted into O_k by comparing the coordinates of two adjacent identity operations. All the source and target elements in-between these are identified as the elements that are respectively deleted and inserted. When there are multiple Q_{vk} 's with the maximum value of the diagonal function, one of such Q_{vk} 's is arbitrarily selected as Q_{vk}° .

The major difference between the dynamic programming and genetic algorithm heuristics is that the dynamic programming approach considers, for each attribute, only *one optimum trajectory* within or nearest to the diagonal region. The genetic algorithm heuristic seeks, for each attribute, *more trajectories*, also including *non-optimum* trajectories. Figure 3.4 summarized the MDSAM heuristic based on dynamic programming.

```

begin                                // search for each attribute an optimal trajectory along or at the
  for k=1 to K do                       diagonal region //
    begin
       $F(Q_k) = F(Q_{vk}^{\circ}) = \max [F(Q_{1k}), \dots, F(Q_{vk}), \dots, F(Q_{vk})]$ 
      Optimal_trajectory_k =  $F(Q_{vk}^{\circ})$ 
    end;
  end;
  no_improve = 0
  Best_Fitness = 100,000,000           // give initial value to Best_Fitness //
  while not (no_improve >= convergence_rate) do
    begin
      multi_dim_cost = integration of optimal trajectories across K dimensions;
      no_improve = no_improve + 1
      if multidim_cost < Best_Fitness then Best_Fitness = multidim_cost and no_improve = 0
    end
  end

```

Figure 3.3: Summarization of the MDSAM heuristic based on dynamic programming.

3.5.3.3 A heuristic based on combining genetic algorithms and dynamic programming

A third heuristic combines genetic algorithms and dynamic programming as follows (Joh et al, 2002). The heuristics based on dynamic programming outperform those based on genetic algorithms in terms of computing time. However, the heuristics based on genetic algorithms improve the solution accuracy significantly. Therefore, genetic algorithms use the solution of dynamic programming as a starting value (re. figure 3.3) instead of randomly starting with a population of the 0th generation. For a further search to solution the random application of genetic operators (re. figure 3.2) is used. Finally, the new heuristic sharply improves the solution accuracy while only moderately increasing computing time. In chapter four, the heuristic based on genetic

algorithms and dynamic programming, described in figures 3.3 and 3.2, is applied to real data sets consisting of server sessions providing information about visited web pages and time spent on pages.

3.5.4 *Two-dimensional SAM*

In the following chapter, SAM and MDSAM are applied to different data sets. More precisely, two different attributes are used: visited web pages and visiting time of web pages. Other attributes like page type or web structure level are not used because the inter-attribute relationships with visited web pages are too strong, which means that, if one attribute is known, the other attributes are known as well. Analysis of such attributes by means of MDSAM produces the same outcome as when only visited pages are analysed. For example, a web page is always related to one page type and one web structure level whereas several different visiting times may be related to the same web page. Therefore, in the following chapters, we will use the term two-dimensional (2-dim) SAM to refer to MDSAM applied to two-dimensional sequences (i.e. server sessions consisting of visited web pages and visiting times). Finally, the next chapter starts with an overview of the surplus value of SAM and two-dim SAM and examples of situations are given when to apply SAM and 2-dim SAM to server sessions.

CHAPTER 4

APPLICATIONS OF SAM AND 2-DIM SAM TO WEB USAGE DATA

This chapter examines the contributions of SAM and 2-dim SAM for Web Usage Mining studies. SAM is applied to web usage data in order to discover visiting profiles providing information of *visited pages* and the *order* in which pages are visited on a web site. 2-dim SAM is applied to web usage data in order to discover visiting profiles providing information of *visited pages* and the time that people stay on a page, also called *categories of visiting page time*. Moreover, 2-dim SAM also takes into account the *order* in which pages are visited on a web site and *relations* between visited pages and categories of visiting page time.

Preliminary studies of SAM and 2-dim SAM applications on real web data are given in Hay et al (2003b), (2003c), (2003d), (2002a), (2002b). In this chapter, SAM and 2-dim SAM distance measures are applied to real log files of visiting behaviour on three different web sites. First, our approach of Web Usage Mining is explained in section 4.5. We also provide a method for defining categories of visiting page time and illustrate its application on web usage data stored in log files. In section 4.6, the data that is used throughout experimental tests are described and illustrated by means of statistics and other graphical presentations.

Following section 4.7, the actual ‘mining’ step of our approach of Web Usage Mining is described. Here, server sessions are clustered based on SAM or 2-dim SAM distance measures. In order to define the right number of clusters, several criteria are used for defining a trade off between number of clusters and model fit. Finally, section 4.7 provides cluster solutions for SAM applied to server sessions consisting of visited pages. In section 4.8, these cluster solutions are examined on visited pages, the order of occurrence of visited pages and on the length of server sessions. In order to provide a general, graphical overview how people visit a web site, groups of surfing behaviour,

represented by visiting profiles, are presented, for each of the three data sets, in section 4.9. The graphical presentations provide information about navigations and the structure of the web site as well as direct hyperlinks between web pages. URL addresses of web pages are also shown in the graphs. Finally, in section 4.10, the results of applying SAM to server sessions and clustering based on SAM are deployed. For each web site, examples are given how the structure of the web site may be adjusted conform to visiting profiles, providing information about the order of visited pages. Also, in order to provide better and faster services to web visitors, examples of page prediction are given. Finally, besides structure improvement and page prediction, other topics for applying the results are given.

In order to show that SAM is a better method for measuring the order of visited pages in server sessions, the same data sets are used in section 4.11, where server sessions are clustered based on Association distance. Association distance measures are commonly used Euclidean based distance measures between sequences, which do not take into account the order of elements (Everitt, 1980). Preliminary tests with Association distance applied to server sessions are illustrated in Hay et al (2003b, 2003c). Likewise, clusters are examined on visited pages, the order of occurrence of visited pages and on the length of server sessions. Finally, comparisons are made with the clustering results based on SAM distance measures.

In order to show that 2-dim SAM is a better method for measuring relations between visited pages and categories of visiting page time, without losing its capacities of measuring sequential information, SAM is compared with 2-dim SAM in section 4.12. The data sets of server sessions consisting of visited pages that were previously used in experimental tests of SAM and Association, are now enlarged with categories of visiting page time. First, server sessions consisting of visited pages and categories of visiting page time are clustered based on 2-dim SAM distance measures. The clusters are examined on the order of and relations between visited pages and categories of visiting page time. Second, server sessions consisting of visited pages and categories of visiting page time are clustered based on SAM distance measures. The clusters are examined on the order of and relations between visited pages and categories of visiting page time. The results of the first and second approach are compared. Finally, some illustrations are given how information provided by profiles, resulting from clusters based on 2-dim SAM distance measures, may be deployed. For example, profiles based on 2-dim SAM distance measures between server sessions consisting of visited pages and categories of visiting page time, may not only predict that page y is visited after page x but also the time that people will stay on page y and page x. This may suggest the urgency of delivering pages by web servers. Also, extracted information given by

profiles based on 2-dim SAM distance measures might be used to verify whether navigation and content pages are actually used by the visitors conform to the intentions of the web developer.

Finally, conclusions are given about the experimental tests using SAM, Association and 2-dim SAM distance measures between server sessions from three different data sets. Indications are provided when to use SAM, Association and 2-dim SAM. Ultimately, avenues for future research are given.

4.1 Surplus value of SAM

The surplus value of SAM, compared to other distance measures, is that SAM incorporates the *order of elements* in addition to measuring distances between sequences. This means that, if SAM is used as distance measure for clustering, sequences are grouped based on the order of occurrence of elements as well as equalities of elements. Elements are defined in section 3.4 of the previous chapter.

We illustrate this feature of SAM by means of an example, given in table 4.1. In the first row, four sequences S_1, S_2, S_3, S_4 are given, holding three to five elements. Instead of using the general examples of server sessions in the previous chapters (for example, page x page y page z), from now on each web page is identified with a unique integer value (for example, 1, 2, 3). In the following rows of table 4.1, SAM distance measures and Association distance measures are calculated between each sequence pair, using equations (3.1) and (4.1) respectively. Association distance measures are commonly used Euclidean based distance measures between sequences, which do not take into account the order of elements (Everitt, 1980). Other methods, which are often used for measuring distances between sequences, are given in appendix four. Note that none of them incorporates the order of elements. For non-metric data, Association distance is measured by transforming each sequence into a vector and counting the number of dissimilarities at each position of the sequence. Missing values in either one of the compared sequences are treated as dissimilarity. In particular, the distance between two sequences S_1 and S_2 , based on Association distance, is presented with the following formula (Hay et al, 2003b, 2003c):

$$d_{ASS}(S_1, S_2) = \sum_{i=1}^n f_i \quad (4.1)$$

$$\text{with } \begin{cases} f_i = 1 & \text{if } S_1(i) \neq S_2(i) \\ f_i = 0 & \text{otherwise} \end{cases}$$

where

d_{ASS} is the distance between two sequences S_1 and S_2 , based on Association distance;

$\sum_{i=1}^n f_i$ is the sum of dissimilarities between two sequences S_1 and S_2 , from positions i to n ;

n is the number of positions of S_1 or S_2 if the sequences are of equal length, otherwise n is equal to the number of positions of the longest sequence;

$S_1 = 1\ 2\ 3\ 4\ 5$ $S_2 = 2\ 3\ 4$ $S_3 = 2\ 1\ 4\ 5$ $S_4 = 1\ 2\ 1\ 4\ 5$		
Sequence pair	$d_{SAM} (w_d = w_i = 1; \eta = 2)$	d_{ASS}
(S_1, S_2)	2	5
(S_1, S_3)	3	5
(S_1, S_4)	2	1
(S_2, S_3)	3	2
(S_2, S_4)	4	5
(S_3, S_4)	1	5
Ward clustering based on distance matrix		
Cluster	SAM	Association
1	S_1 and S_2	S_1 and S_4
2	S_3 and S_4	S_2 and S_3

Table 4.1: Example of clustering sequences based on SAM and Association distance.

The objective is to cluster sequences based on equalities and the order of occurrence of elements. In the example, SAM recognizes the longest common sub strings respecting the order of elements, between S_1 and S_2 (i.e. pattern 2, 3, 4) and between S_3 and S_4 (i.e. pattern 2, 1, 4, 5). Instead of comparing elements within sequences based on positions only, SAM is able to search for patterns of elements across positions. Therefore, clustering based on the SAM distance matrix groups S_1 and S_2 in cluster 1 and S_3 and S_4 in cluster 2. However, clustering based on the Association distance matrix does not recognize these patterns and groups S_1 and S_4 in one cluster and S_2 and S_3 in the other. Note that, instead of using Ward for hierarchical clustering, other methods may be used as well like for example median, centroid, complete or single linkage. Other clustering methods are described in section 4.7, table 4.8 of this chapter. Figure 4.1 plots the dendrograms that resulted from clustering sequences S_1, S_2, S_3, S_4 , based on SAM and Association distance.

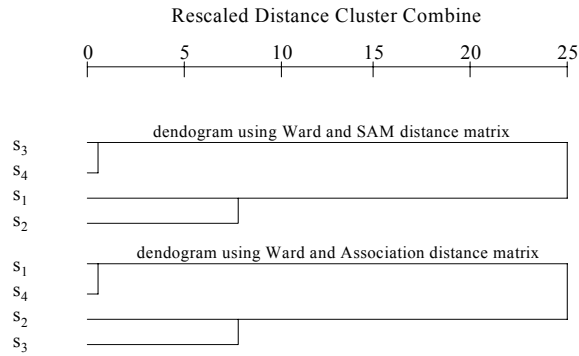


Figure 4.1: Example of dendrograms resulted from clustering sequences based on SAM and Association distance.

4.2 Surplus value of 2-dim SAM

2-dim SAM measures similarities between sequences based on two attributes. In section 3.4 of the previous chapter, an *attribute* or *dimension* is defined as a particular type of information in a sequence, for example pages or times. In addition to SAM, 2-dim SAM measures not only the order and equalities of elements between two sequences, but captures also the *inter-attribute relationships* within sequences. An example of an inter-attribute relationship is “3” (of attribute 1) and “2” (of attribute 2) within the following two-dimensional sequences:

$$\begin{aligned}
 S_x = & \quad 1 \ 2 \ 3 \text{ (attribute 1)} \\
 & \quad 0 \ 0 \ 2 \text{ (attribute 2)} \\
 S_y = & \quad 1 \ 2 \ 3 \text{ (attribute 1)} \\
 & \quad 1 \ 2 \ 2 \text{ (attribute 2)}
 \end{aligned}$$

This means that, if 2-dim SAM is used as distance measure for clustering, 2-dimensional sequences are grouped based on three characteristics: the order of occurrence of elements, equalities of elements and inter-attribute relationships. However, if SAM is used as distance measure for clustering, two-dimensional sequences are grouped based on two characteristics: the order of occurrence of elements and equalities of elements. SAM is not able to recognize inter-attribute relationships between two-dimensional patterns.

We illustrate the difference between SAM and 2-dim SAM by means of an example, given in table 4.2. In the first row, four two-dimensional sequences are given, holding three to five elements. In the following rows, SAM and 2-dim SAM distance measures are calculated between each sequence pair. If SAM is used to calculate the distance between 2-dimensional sequences, equations (3.1) and (3.2) are used to compute the distance for each attribute. Then, the individual distance measures for each attribute are summed together to represent the total SAM distance between 2-dimensional sequences. For example, if operation weights $d = i = 1$; $\eta = 2$ are used, $d_{\text{SAM}}(S_1, S_4)$ for attribute 1 equals 2, based on the trajectory $\{d3S_1, i3S_1\}$ and $d_{\text{SAM}}(S_1, S_4)$ for attribute 2 equals 4, based on trajectories $\{d3S_1, i2S_1, d5S_1, i5S_1\}$, $\{d2S_1, i2S_1, d5S_1, i5S_1\}$ or $\{d1S_1, i2S_1, d5S_1, i5S_1\}$. The total SAM distance between S_1 and S_4 equals 6. If 2-dim SAM is used to calculate the distance between 2-dimensional sequences, equations (3.4), (3.5) and (3.3) (dynamic programming and genetic algorithms) are used to compute the distance across attributes, while searching for the most equal trajectories between attributes. For example, based on trajectories $\{d3S_1, i3S_1\}$ for attribute 1 and $\{d3S_1, i2S_1, d5S_1, i5S_1\}$ for attribute 2, $d_{2\text{-dim SAM}}(S_1, S_4)$ equals 5.

$S_1 = 1\ 2\ 3\ 4\ 5$ (attribute 1) $= 0\ 0\ 0\ 1\ 4$ (attribute 2) $S_2 = 2\ 3\ 4$ (attribute 1) $= 0\ 0\ 4$ (attribute 2) $S_3 = 2\ 1\ 4\ 5$ (attribute 1) $= 0\ 0\ 1\ 4$ (attribute 2) $S_4 = 1\ 2\ 1\ 4\ 5$ (attribute 1) $= 0\ 1\ 0\ 1\ 1$ (attribute 2)		
Operation weights used for attribute 1 and 2: $d = i = 1$; $\eta = 2$		
Sequence pair	d_{SAM}	$d_{2\text{-dim SAM}}$
(S_1, S_2)	4	4
(S_1, S_3)	4	3
(S_1, S_4)	6	5
(S_2, S_3)	4	4
(S_2, S_4)	8	7
(S_3, S_4)	4	4
Ward clustering based on distance matrix		
Cluster	SAM	2-dim SAM
1	S_1 and S_2	S_1, S_3 and S_2
2	S_3 and S_4	S_4

Table 4.2: Example of clustering sequences based on SAM and 2-DIM SAM.

The objective is to cluster sequences based on equalities and the order of occurrence of elements as well as inter-attribute relationships. In the example,

SAM and 2-dim SAM both recognize pattern 2, 3, 4 of the first attribute in S_1 and S_2 . However, only 2-dim SAM recognizes inter-attribute relationships and therefore clusters S_1 , S_2 and S_3 together whereas SAM groups S_1 with S_2 in cluster 1 and S_3 with S_4 in cluster 2. In other words, 2-dim SAM considers the relations between attribute 1 and 2 of 1-0, 2-0, 3-0 or 5-4 in S_1 , S_2 or S_3 as relatively strong inter-attribute relationships within the data. In S_4 , different relations between attribute 1 and 2 are shown. For example, 2-1 and 5-1. For this reason, 2-dim SAM considers S_4 to be more distant from S_1 , S_2 and S_3 . Figure 4.2 plots the dendrograms that resulted from clustering sequences S_1 , S_2 , S_3 and S_4 , based on SAM and 2-dim SAM distance.

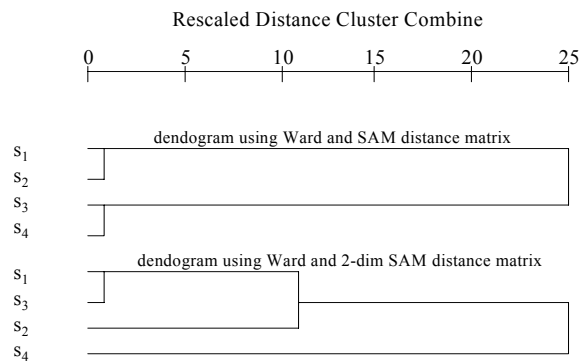


Figure 4.2: Example of dendrograms resulted from clustering sequences based on SAM and 2-DIM SAM.

4.3 Objectives

Within research area of General Access Pattern Tracking, which is described in chapter one, section 1.3, the objective of applying SAM or 2-dim SAM to Web Usage Data is to discover profiles of visiting behaviour on a web site. *SAM* is used to measure distances between server sessions of one attribute (dimension), which is visited pages. Profiles resulting from SAM distance measures provide information about *visited pages* and the *order* in which pages are visited. For example, page y is visited after page x. *2-dim SAM* is used to measure distances between server sessions of two attributes (dimensions), which are visited pages and categories of visiting page time. Profiles resulting from 2-dim SAM distance measures provide information about *visited pages*, *categories of visiting page time*, the *order* in which pages and times occur and *inter-attribute relationships* between pages and times. For example, page y is visited after page x while page y is related with time t2 and page x is related with time t1.

The extracted profiles of visiting behaviour are represented by clusters, which resulted from clustering server sessions based on distance matrices holding SAM or 2-dim SAM distance measures. If SAM is used as distance measure between server sessions consisting of visited pages, server sessions are clustered together based on equalities of pages and the order in which pages occur within server sessions. If 2-dim SAM is used as distance measure between server sessions consisting of visited pages and categories of visiting page time, server sessions are clustered together based on equalities of pages and times, the order in which pages and times occur within server sessions and relations between particular pages with particular times.

The extracted profile information may be used by web personalization systems to provide better and faster services to web visitors. For example, profiles based on SAM distance measures between server sessions consisting of visited pages may predict that page y is visited after page x and/or that page z is visited after pages y and x. Moreover, profiles based on 2-dim SAM distance measures between server sessions consisting of visited pages and categories of visiting page time, may not only predict that page y is visited after page x but also the time that people will stay on page y and page x. This may suggest the urgency of delivering pages by web servers. Also, for the convenience of the visitor, the structure of the web site may be adjusted conform to the profiles. For example, if page y is visited after page x without the presence of a link from page x to y, we may suggest inserting a direct hyperlink from page x to y. Also, extracted information given by profiles based on 2-dim SAM distance measures might be used to verify whether navigation and content pages are actually used by the visitors conform to the intentions of the web developer.

For example, suppose that page x is a navigation page leading to page z, we expect visitors to follow the ‘road’ from page x to z while staying longer on page z than on page x. If this is contradicted by the extracted profiles, we may provide information for adjusting the web site for the convenience of visitors. Other applications for using the information provided by SAM (2-dim SAM)-based clustering is offering different guided tours to different groups of web-visitors, distinguishing visiting behaviour of ‘first-time’ visitors from regular visitors, inserting cross-links between particular web pages etc. More details are given in section 4.10 and 4.12.2 Deploying the results.

4.4 Three steps in a Web Usage Mining process

Before proceeding to the applications of SAM and 2-dim SAM to real log files of Web Usage Data, we describe in this section our approach of Web Usage Mining. A general overview of the three steps in a Web Usage Mining process is given in chapter two. The details of our approach by means of the SAM and 2-dim SAM applications are provided in figure 4.3. Output of each step is used as input in the following step. First, the raw data, registered in log files, are pre-processed into server sessions so as to become useful for mining. Second, dependent on the objectives, SAM or 2-dim SAM distance measures are calculated between the server sessions. Hierarchical clustering algorithms are invoked on the distance measures in order to obtain, based on several information criteria for defining the number of clusters, clusters of server sessions. Third, for the SAM application, clusters are examined on equalities of pages, the order in which pages occur within the server sessions and on the length of server sessions. For the 2-dim SAM application, clusters are examined on the order of occurrence and inter-attribute relationships between pages and times. Open sequences are used for examining the order of pages (times) in clusters of server sessions. Open sequences are defined in section 4.8 of this chapter.

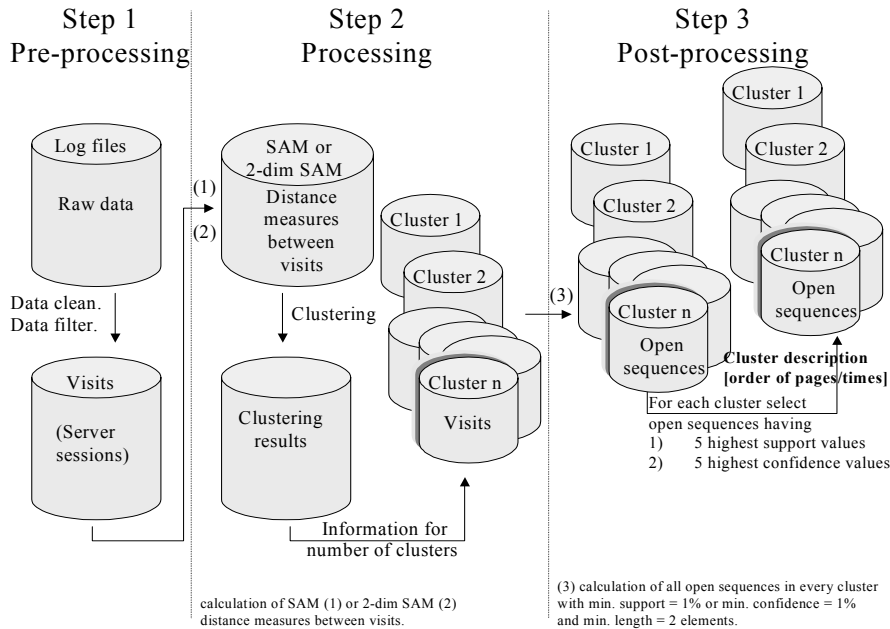


Figure 4.3: Web Usage Mining by means of SAM or 2-dim SAM.

Each step of our approach as well as the data sets used in the SAM and 2-dim SAM applications are described in detail in the following sections. In section 4.5 and 4.6, the raw log file data is pre-processed into server sessions and a description is given of the data sets that are used throughout this project. In section 4.7, SAM distance measures are calculated between server sessions. Then, server sessions are clustered based on SAM distance measures. In section 4.8, clusters based on SAM distance measures are examined on `page_ids`, the order of `page_ids` and on the length of server sessions. After describing and deploying the results in section 4.9 and 4.10, SAM is compared with Association distance in section 4.11. The clusters based on SAM distance between server sessions are compared with those based on Association distance in order to show that SAM groups server sessions together based on the order of pages. In section 4.12, SAM is compared with 2-dim SAM in order to show the capability of 2-dim SAM to cluster server sessions together based on relations between pages and times. Also, the results of the 2-dim SAM application are deployed. This chapter ends with conclusions about applying (2-dim) SAM to real web data stored in log files and avenues for future research in section 4.13.

4.5 Step 1: Pre-processing

In order to analyse visiting behaviour on a web site, sessions of web click stream data must be defined from the raw log data. A *server session* or *visit* is defined in chapter two as the click stream of page views for a single visit of a user to a web site (Cooley, 2000). We will use server session or visit interchangeably. In the absence of cookies or dynamic web pages, in this study server sessions are composed heuristically using the information supplied by the extended log files (Cooley et al, 1999a; CERN, 2002) as follows. First, the data are cleaned in such a way that only URL page requests of the form ‘GET...htm(l)’ are maintained. Then, a unique code is given to each distinct combination of ip address and user agent (Cooley et al, 1999b; Fu et al, 1999). Likewise, a unique code is given to each distinct URL.

Furthermore, server sessions are identified using some time-out heuristic (Banerjee and Ghosh, 2001; Catledge and Pitkow, 1995; Cooley et al, 1999a; Zaïane and Luo, 2001) based on a threshold of 30 minutes visiting page time. *Visiting page time* is the time difference (expressed in seconds) between consecutive page requests (Shahabi et al, 1997) and is calculated for each page that is registered as a request in the log file. Unfortunately, missing values occur when visiting page time is calculated for the last page of a server session. To deal with this problem, visiting page time for the last page of a server session is substituted by the average visiting page time of that particular page taken across all sessions in which the page is not the last page request (Witten and Frank, 2000). Using a threshold of 30 minutes visiting page time means that, for the same ip address, a new session is created when the time between subsequent page requests exceeds 30 minutes. In general, a server session is created when a new ip address and/or user agent is met in the log file. Finally, a filtering method is invoked on the sequences in order to identify visitors using the same ip address and user agent. Note that the focus of this project is to provide profiles of visiting behaviour showing visited pages and the order of visited pages on a web site, rather than examining the various engineering issues with regard to sessionizing or user identification. Besides, within our research area of General Access Pattern Tracking (re. chapter one and two), opposed to Customized Usage Tracking, general instead of individual trends are analysed in order to customize web sites to users. It is important to note that other heuristics or algorithms for identifying users and server sessions may be employed in future research.

After defining the server sessions, holding consecutive page requests, categories of visiting page time are added. The reason for adding time information to page information in server sessions is to examine whether

visitors actually use the site as web designers expect the site is being used. Moreover, time information provides the ability to predict which pages will be requested within certain time limits, which optimises the speed of delivering web pages to users. The approach of using time windows is illustrated in Cooley et al (1999a). Likewise, in Cooley et al (1997) time windows are used to find common characteristics of users that visited a particular page within the time period $[t_1, t_2]$. Yet, in our approach time windows are applied to investigate whether visitors actually (a-posteriori) use the site as web designers expect the site is being used (a-priori) as well as for delivery speed of page predictions. A definition of a-priori defined web pages is given in chapter two, section 2.2.3. Examination is done whether the a-priori defined web pages as ‘content’ and ‘navigation’ match with the a-posteriori visited web pages. In order to define the a-posteriori ‘content’ and ‘navigation’ web pages, an estimated cutoff visiting page time is defined between ‘content’ and ‘navigation’ pages by means of the following equation (Cooley, 2000):

$$t_{\text{cutoff}} = -\ln(1 - \gamma) / \lambda \quad (4.2)$$

where

- γ is the number of navigation pages divided by the number of total pages in the analysis;
- λ is the reciprocal of the observed mean visiting page time in the analysis;

Equation (4.2) is derived from integrating the formula for an exponential distribution from γ to zero (Cooley, 2000). The maximum likelihood estimate for the exponential distribution is the observed mean. Practically, t_{cutoff} is defined by taking the \ln -function of the scale $(1 - \gamma)$ and dividing (i.e. standardizing) it by the observed mean visiting page time. The reason for applying ‘-ln’ instead of ‘ln’ is because the scale $(1 - \gamma)$ will always be less than zero. Taking ‘ln’ of a value less than zero ends up with a negative value. Taking ‘-ln’ of a value less than zero ends up with a positive value.

On the one hand, if for page x actual visiting time is at or below t_{cutoff} , page x is used by the visitor as ‘navigation’ page (a-posteriori), irrelevant of whether the web developer constructed page x as ‘content’ or ‘navigation’ (a-priori) in the web site. On the other, if for page x actual visiting time is above t_{cutoff} , page x is used by the visitor as ‘content’ page (a-posteriori), irrelevant of whether the web developer constructed page x as content or navigation (a-priori) in the web site.

In order to examine, by means of 2-dim SAM, the delivery speed of page predictions, more categories besides t_0 (a-posteriori use of navigation page)

and t_1 (a-posteriori use of content page) are defined. If we want to know whether web pages must be delivered relatively fast (i.e. urgent), at an average speed or relatively slow (i.e. not urgent), the distribution of actual visiting page times is used to define several categories of visiting page times for pages that are actually used as a content page by the visitor. For example, t_1 , t_2 , t_3 are used to notify the page is actually used as a content page by the visitor with visiting page time between t_{cutoff} and 60 seconds, 61-300 seconds, above 300 seconds respectively. The reason for using categories of time instead of continuous time information is due to the SAM (and 2-dim SAM) algorithm. When two sequences are equalized by means of SAM (and 2-dim SAM), sequences must hold categorical elements. Numerical elements are difficult to handle by means of SAM (and 2-dim SAM). Future research will discuss algorithms for analysing continuous data for visiting page time.

Eventually, for the analysis of SAM and 2-dim SAM, server sessions are built in the form of `session_id, {(<unique code for URL request>); (<category of visiting page time>)}` representing consecutive pages requested by the same user with corresponding time information. Examples of how data is pre-processed into server sessions, using the heuristics described in this section, are given in table 4.3. The records are ordered based on the time of the request. Suppose $t_{\text{cutoff}} = 5$, defining t_0 . Likewise, suppose t_1 , t_2 and t_3 are categories of equally distributed visiting page times above 5 seconds. This means that, the number of occurrences of visiting page times in t_1 , t_2 and t_3 are equal. The web developer has constructed pages 68, 65, 9 and 1 for navigational use. In the last row of table 4.3, the first, third and fourth server sessions demonstrate visiting behaviour as expected by the web developer i.e. navigational pages are visited during less than 5 seconds; content pages are visited for more than 5 seconds.

Code ip address & user agent	Date	Time	Code URL
1	2001-02-15	00:01:43	68
1	2001-02-15	00:01:45	65
1	2001-02-15	00:01:47	55
2	2001-02-15	00:01:47	68
3	2001-02-15	00:01:48	1
4	2001-02-15	00:01:52	13
2	2001-02-15	00:02:05	9
1	2001-02-15	00:02:11	70
2	2001-02-15	00:02:30	68
2	2001-02-15	00:02:31	71
...
$t_{\text{cutoff}} = 5$ $t_0 \leq t_{\text{cutoff}}$ $t_{\text{cutoff}} < t_1 \leq 60$ $60 < t_2 \leq 300$ $300 < t_3$	Average visiting page time for page 70 = 95 for page 9 = 3 for page 1 = 8 for page 13 = 306 for page 71 = 75 ...		
Server sessions 1, {(68, 65, 55, 70); (t0, t0, t1, t2)} 2, {(68, 9, 68, 71); (t1, t1, t0, t2)} 3, {(1); (t1)} 4, {(13); (t3)} ...			

Table 4.3: Examples of constructing server sessions.

Finally, we give two remarks. First, as mentioned in chapter two, the time of the request is the time the request is received (and logged) by the web server. This means that the visiting page times, calculated from the logged time data, may differ from the real visiting (i.e view) time, since overload of network traffic may delay deliveries of pages to the user. However, most of the studies in Web Usage Mining rely on information supplied by (extended) log files and, although time differences may occur, logged time data are considered to provide reliable information (CERN, 2002, Cooley et al, 1999a). Future research discusses ways for handling differences between real and logged time data within Web Usage Mining studies. Second, server sessions consisting of one page only are not excluded from the analysis because they might provide profiles of visiting behaviour to web pages that are directly accessed using the URL address instead of using the navigational pages. Yet, care must be taken when interpreting the visiting page time since average values are used to replace missing values. Future research mentions excluding server sessions of one page from the analysis and verifies whether the results are significantly different.

4.6 Describing the data

SAM and 2-dim SAM are applied to three data sets. For each data set, visiting behaviour towards a different web site is examined. The first analysis concerns log files of our university web site, Faculty of Applied Economic Sciences (<http://www.luc.ac.be/tew>). This site consists of information and course material of a bachelors and masters degree in Applied Economic Sciences and Economic Engineering at the Limburg University Center (LUC) in Belgium. The second uses logged data of the Music Machines web site (<http://machines.hyperreal.org>), home of musical electronics on the web. Music Machines offers images, software, schematics, synthesizers, effects, drum machines, recording equipment etc. Visiting behaviour on this web site is also analysed by Perkowski and Etzioni (2000). Adaptive web sites mine the data buried in server logs to produce more easily navigable web sites. Through index page synthesis, a site could offer an alternative organization of its contents based on user access patterns. Finally, in the third experiment log files of the web site of a Belgian telecom provider are analysed. Due to privacy agreements we are not able to provide name/URL address of their web site. Generally, the site provides information about products, prices, subscriptions, business solutions, customer services, jobs, FAQ, press release etc.

After pre-processing the data using the heuristics given in the previous section, the raw data in the log files are converted into server sessions. Table 4.4 presents, for each data set, the number of server sessions. Also, the period of data registration and the number of distinct URL addresses that were logged in the files, are given. Each URL address is represented by a unique code, also called `page_id` or (web) page. For the first data set, a total number of 2764 server sessions are defined over 71 different `page_ids`. The second data set provides 3131 server sessions including 1159 different `page_ids`. Finally, in the third data set, 773 server sessions are defined from a web site of 492 web pages.

Data set	URL address of web site	Period of data registration	Total number of server sessions	Total number of distinct <code>page_ids</code>
1	http://www.luc.ac.be/tew	15/02/2001 - 22/07/2001	2764	71
2	http://machines.hyperreal.org	01/02/1999- 03/02/1999	3131	1159
3	'Belgian telecom provider'	20/02/1999- 28/02/1999	773	492

Table 4.4: Three pre-processed data sets used for SAM and 2-dim SAM experiments.

Table 4.5 provides statistics about the server sessions of each data set. All of the data sets hold at least one server session consisting of one element or page. The longest and average lengths of the server sessions differ among the data sets. In the last row, the total number requests are the total number of requested web pages in the data sets.

Statistics	Server sessions		
	Data set 1	Data set 2	Data set 3
Shortest	1	1	1
Longest	55	20	38
Average length	3	2.5	6
Total number of requests	8308	7887	4605

Table 4.5: Describing server sessions used for SAM and 2-dim SAM experiments.

Figure 4.4 provides, for each data set, the distribution of the server sessions' length. On the horizontal axis, the length of the server sessions, ranging from 1 to 55, is given. On the vertical axis the relative frequency (number of server sessions of the corresponding length divided by the total number of server sessions in the data set, multiplied by 100) is given. For example, considering the first data set of web usage behaviour on <http://www.luc.ac.be/tew>, 46.92% (i.e. $[1297 / 2764] * 100$) of the server sessions are one page long. The length of the server sessions in data set one and two follow approximately the same distribution. In the third data set, 31.95% and 18.63% of the server sessions are respectively five and six pages long. Although the first data set contains server sessions up to 55 pages long, merely 0.24% of the server sessions are longer than 30 pages. In the second data set 89.71% of the server sessions are one to five pages long. Finally, in the third data set, 89.53% of the server sessions are one to ten pages long.

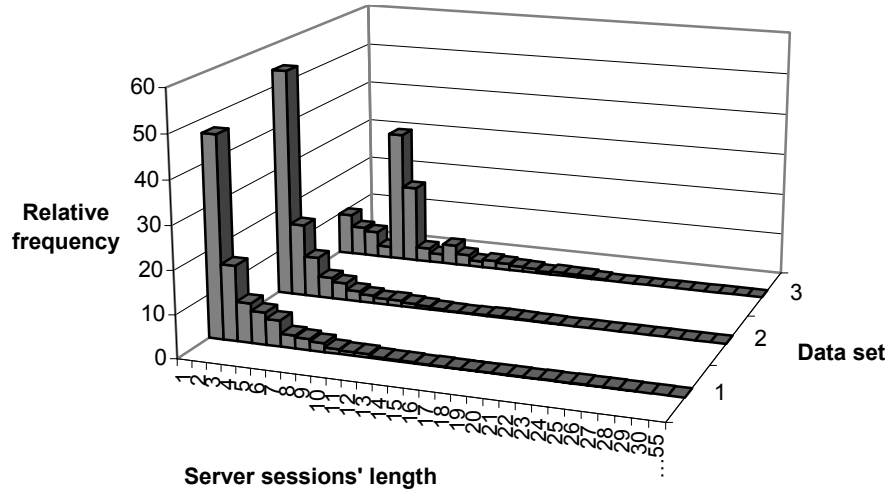


Figure 4.4: Distribution of the length of server sessions.

Figures 4.5 to 4.7 provide, for each data set, the distribution of distinct page_ids. In figure 4.5, on the horizontal axis, 71 distinct page_ids are presented. On the vertical axis, relative frequencies (number of requests of the corresponding page_id divided by the total number of requests (i.e. 8308) in the file, multiplied by 100) are given. For example, 19.03% ($[1,581 / 8308] * 100$) of the requested pages in the first data set are page 68. For the three highest relative frequencies, url-addresses are written in the graph.

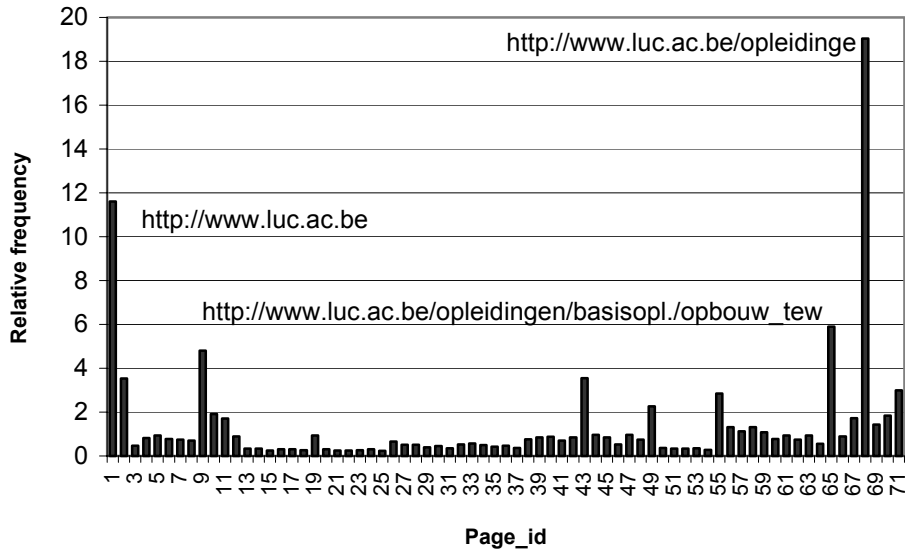


Figure 4.5: Distribution of 71 distinct page_ids in data set 1.

In figure 4.6, 1,159 distinct page_ids are represented by means of 50 groups. In the first stage of the pre-processing step, web pages were ordered alphabetically on URL address before a unique code (i.e. page_id) is assigned to each distinct URL address, starting with page 1 for <http://machines.hyperreal.org/addressbook> and ending with page 1,159 for <http://machines.hyperreal.org/software>. Here, each group reflects 23 web pages, except for the last group. For example, group 1 reflects page 1 to 23, group 2 reflects page 24 to 46, group 3 reflects page 47 to 69 etc. Finally, group 50 reflects page 1,128 to 1,159. On the vertical axis, the frequency values (number of requests of the page_ids within the corresponding group divided by the total number of requests (i.e. 7,887) in the file, multiplied by 100) are given. The graph shows that 25.28% of the visited pages in the second data set are pages within group 29, reflecting web pages 645 to 667 (including 645 and 667). The following two highest relative frequency values are 7.24% for group 45, reflecting web pages 1013 to 1035 (including 1013 and 1035) and 5.44% for group 50, reflecting web pages 1128 to 1159 (including 1128 and 1159).

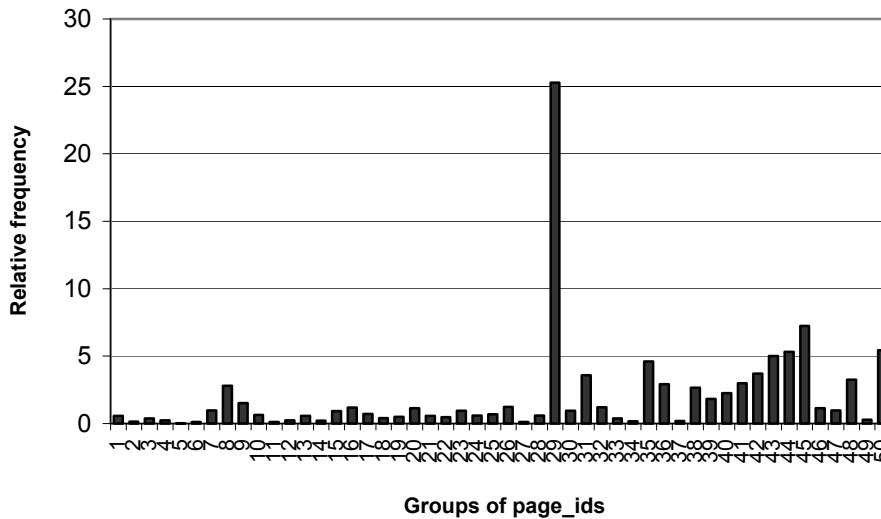


Figure 4.6: Distribution of 1159 distinct page_ids, represented in 50 groups, in data set 2.

In figure 4.7, 492 distinct page_ids are represented by means of 20 groups. Likewise, in the first stage of the pre-processing step, web pages were ordered alphabetically on URL address before a unique code (i.e. page_id) is assigned to each distinct URL address, starting with page 1 for 'www...../a...' and ending with page 492 for 'www.../p...'. Here, each group reflects 25 web pages, except for the last group. For example, group 1 reflects page 1 to 25, group 2 reflects page 26 to 50, group 3 reflects page 51 to 75 etc. Finally, group 20 reflects page 476 to 492. On the vertical axis, the frequency values (number of requests of the page_ids within the corresponding group divided by the total number of requests (i.e. 4,605) in the file, multiplied by 100) are given. The three highest relative frequency values are shown for group 15 (15.08%), 12 (14.83%) and 20 (12.65%).

We remark that groups of page_ids are used throughout this thesis for presentational reasons only, because scaling 1,159 and 492 different page_ids will be too large. Yet, every analysis is executed on individual page_ids and not on groups of page_ids. Future research discusses ways of analysing web pages at a higher hierarchical level.

We also remark that, instead of using alphabetically ordered URL addresses, we could group page_ids together based on classes, like for example manufacturers, software or samples (data set 2) and products, prices or services

(data set 3). Figures of groups of page_ids based on classes for data set 2 and 3 are given in appendix four.

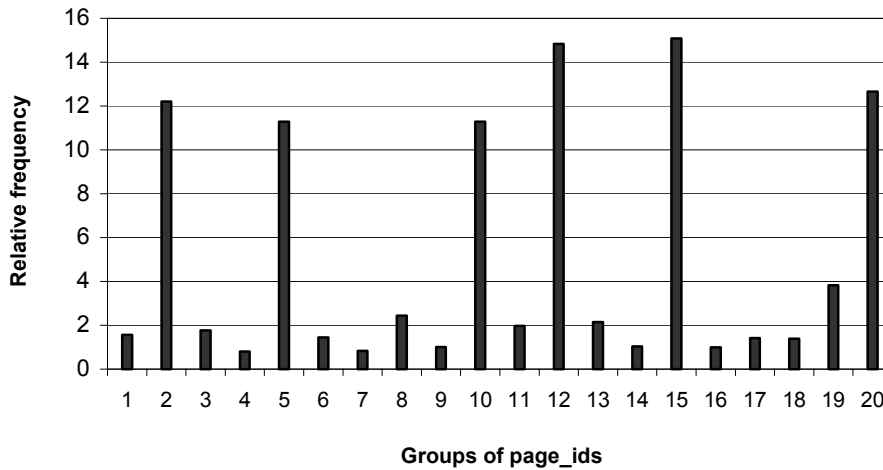


Figure 4.7: Distribution of 492 distinct page_ids, represented in 20 groups, in data set 3.

For each data set, categories of visiting page time are defined by t_{cutoff} using equation (4.2) as follows. Table 4.6 provides the values of γ and λ for each data set. In the first data set, 7 out of 71 pages are structured by the web developer as navigational pages. Examples of navigation pages are <http://www.luc.ac.be/tew/information> and <http://wwwq.luc.ac.be/tew/education>. In the second data set, the main navigational pages are <http://machines.hyperreal.org> (home page), <http://machines.hyperreal.org/gear> (manufacturers page), <http://machines.hyperreal.org/links> (provides references to other sources of information on the web), <http://machines.hyperreal.org/guide> (explains how to use Music Machines and how it's structured) and <http://machines.hyperreal.org/email> (guides you to the email account in order to contact the Music Machines crew). Other examples of navigation pages are /categories, /images, /software etc. We also believe that, for each manufacturer, one navigation page is defined, for example <http://machines.hyperreal.org/gear/ARP>, <http://machines.hyperreal.org/gear/Akai>, <http://machines.hyperreal.org/gear/Yamaha> etc. Ultimately, from the basic structure of the Music Machines web site, 99 pages are defined as navigation

pages. In the third data set, 50 out of 492 web pages are navigation pages. Some examples of navigation pages are /products&services, /sales, /tariff. In order to define the second parameter λ , the observed mean visiting page time is calculated across all of the page requests in each data set. In the first data set, pages are visited during an average time of 2.07 minutes or 124.20 seconds. In the second and third data set, the mean visiting page times are 107.89 and 54.72 seconds respectively. Finally, t_{cutoff} equals 12.89 in the first data set, indicating that, if visitors stay less than 12.89 seconds on a page, they actually use this page as a navigation page. In the second and third data set, t_{cutoff} equals 9.63 and 5.86 respectively.

Parameters	Data set		
	1	2	3
γ	0.09859	0.08541	0.10162
λ	0.00805	0.00927	0.01827
t_{cutoff}	12.89	9.63	5.86

Table 4.6: Calculating t_{cutoff} for each data set.

Also, for each data set, the distribution of visiting page times above t_{cutoff} is used to define categories of visiting page time when visitors actually use pages as content pages. Because the actual duration of visits to content pages may be spread out over a time range of $[t_{\text{cutoff}}, 1799]$ seconds, several categories are defined based on equal distributions, also known as equal frequency binning. Equal frequency binning is used in our research project for the following reasons:

- Rarely occurring categories are avoided. Categories that may be interesting and rarely occur in the analysis are often not presented in the results of the analysis.
- Pre-defined input is avoided. If the analyst pre-defines categories, the results will provide information that was already known before the analysis took place.
- All of the three web sites in our experimental tests offer more content pages and less navigational pages, which advises using several time categories for content (i.e. index) pages. However, we must remark, if web sites are analysed with more navigational pages than content pages, it might be wise to create several time categories for navigational instead of content pages.

Figures 4.8 to 4.10 present the distribution of visiting page times for data set one, two and three. Within the group of visiting page times above t_{cutoff} we define three categories: short, medium and long stay on a web page. For each

data set, the boundaries of time categories are given in table 4.7. Based on the distribution of actual visiting page times above t_{cutoff} , the boundaries are defined as follows. The total number of requests is the same within each time category. For example, in the first data set, the total number of requests with visiting page time above 12.89 seconds is equal to 6,033. Following, actual visiting page times are ordered ascending starting with 12.891 up to 1,799 seconds. Each time one actual visiting page time is assigned to t_1 until the total number of requests within time category t_1 is equal to 2,011 (i.e. $6,033 / 3$). The boundary of t_1 is now equal to the last visiting page time that was added to t_1 before the total number of requests that were added to t_1 exceeded 2,011. Then, visiting page times are assigned to t_2 and t_3 accordingly. Future research discusses another boundary calculation, based on equal total visiting page time within each category.

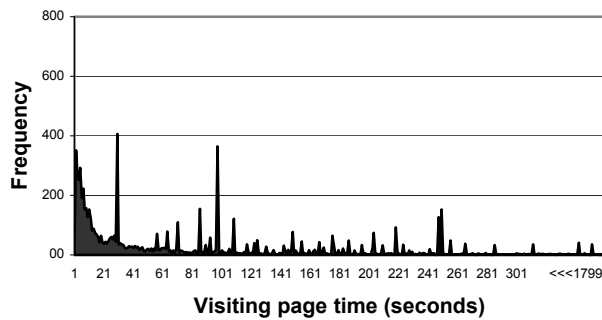


Figure 4.8: Distribution of visiting page times for data set 1.

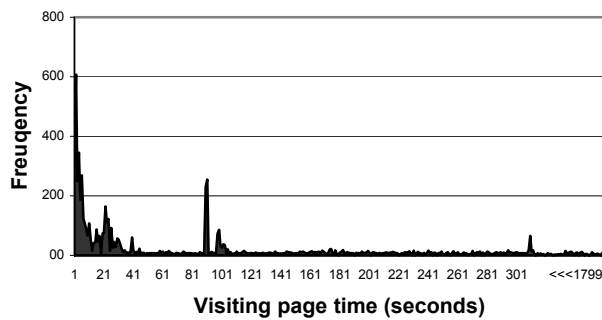


Figure 4.9: Distribution of visiting page times for data set 2.

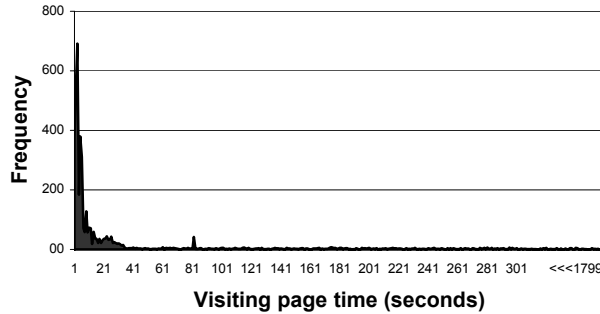


Figure 4.10: Distribution of visiting page times for data set 3.

Dataset		
1	2	3
$0 < t_0 \leq 12.89$	$0 < t_0 \leq 9.63$	$0 < t_0 \leq 5.86$
$12.89 < t_1 \leq 56$	$9.63 < t_1 \leq 68$	$5.86 < t_1 \leq 13$
$56 < t_2 \leq 166$	$68 < t_2 \leq 204$	$13 < t_2 \leq 107$
$166 < t_3$	$204 < t_3$	$107 < t_3$

Table 4.7: Visiting page time categories based on equal number of requests.

Figures 4.11 to 4.13 present average visiting page times for data set 1, 2 and 3. We remark that, in figures 4.12 and 4.13, the horizontal axis represents groups of distinct web pages, as previously mentioned and conform with figures 4.6 and 4.7. *Average visiting page time* in data set 1 is calculated for each `page_id` by summing the corresponding visiting page times, expressed in seconds, and dividing it by the number of requests of the corresponding `page_id`. In data set 2 and 3, average visiting page times are first calculated for each `page_id`. Then, the average is taken for each group. In the first data set, pages with identification number 68, 2 and 43 present the lowest average visiting page times of respectively 30, 57 and 64 seconds. In the second data set, group 14 (reflecting `page_id` 300 till 322) and 15 (reflecting `page_id` 323 till 345) present the lowest average visiting page times of respectively 14.04 and 17.61 seconds. Also group 49 (reflecting `page_id` 1105 to 1127), group 5 (reflecting `page_id` 93 to 115) and group 37 (reflecting `page_id` 829 to 851) have low average visiting page times of respectively 26.13, 28 and 28.52 seconds. In particular, `page_ids` 163, 349, 713, 984, 1082, 933, 815, 151, 947 and 1129 have low average visiting page times of respectively 2, 2, 4, 5, 5, 6, 7, 8, 8 and 8 seconds. Finally, in the third data set, the lowest average visiting

page time (24.20 seconds) is given for group 16 (reflecting page_id 376 till 400). Particularly page_ids 81, 492, 249, 250, 436 and 27 are visited during a short time of respectively 1, 1, 2, 2, 2 and 3 seconds.

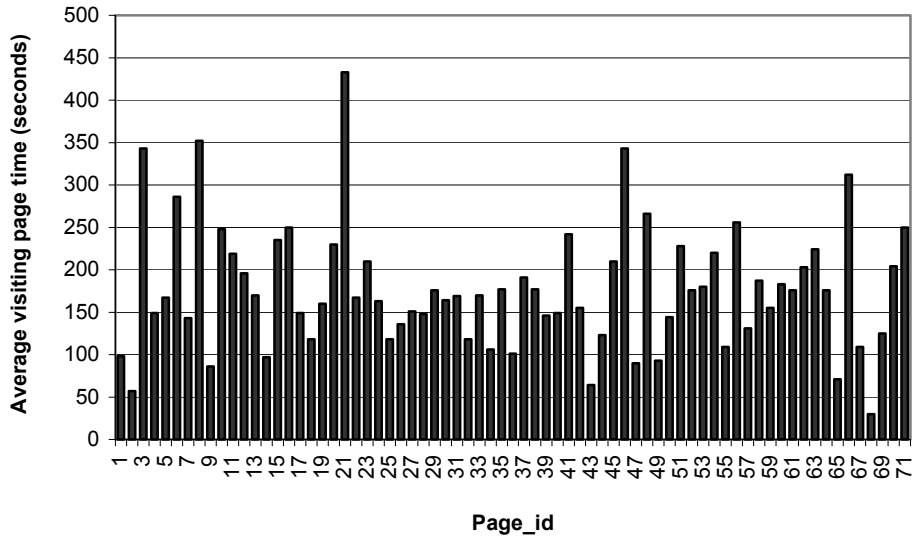


Figure 4.11: Average visiting page times in data set 1.

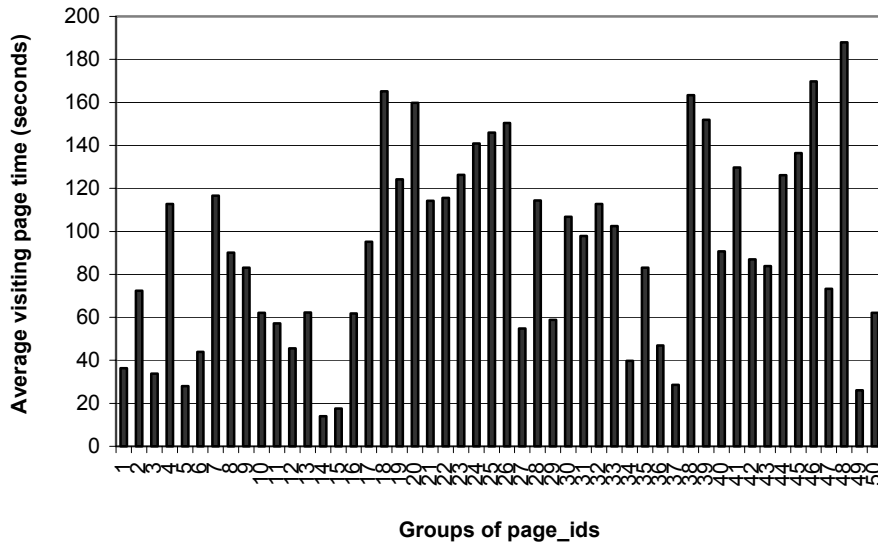


Figure 4.12: Average visiting page times in data set 2.

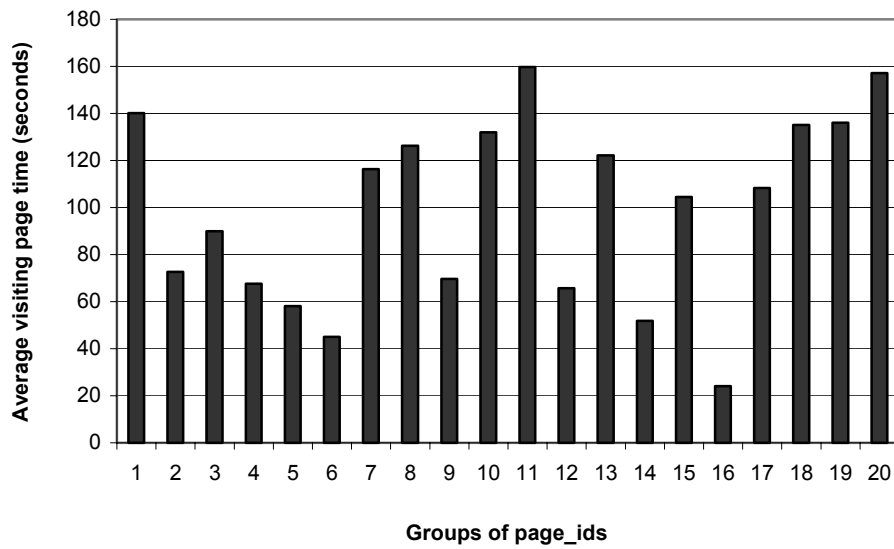


Figure 4.13: Average visiting page times in data set 3.

4.7 Step 2: Processing

In the second step of our approach in the Web Usage Mining process (re. figure 4.3), SAM distance measures are calculated between the server sessions in each data set, using equation (3.1) and the algorithm summarized in figure 3.1. We remark that, in this section, SAM is applied to three data sets. In section 4.11 and 4.12, Association distance and 2-dim SAM are applied to three data sets.

The operation weights are defined in its most basic and natural way i.e. $d = i = 1$ and $\eta = 2$ (Sankoff and Kruskal, 1983), indicating that the effort of deleting an element is the same as inserting an element and reordering is the sum of one insertion plus one deletion. In chapter six, the influence of changes in operation weights on the results are examined by means of sensitivity analyses.

For each data set given in table 4.4, one distance matrix holding pair wise SAM distance measures between server sessions is used as distance measure for clustering. For example, if 2764 server sessions, $S_1, S_2, \dots, S_{2764}$, are analysed, SAM calculates distance measures between every pair of sessions i.e. between S_1 and S_2, S_1 and S_3, \dots, S_1 and S_{2764} , between S_2 and S_3, S_2 and S_4, \dots, S_2 and S_{2764}, \dots , and finally between S_{2763} and S_{2764} . These distance measures are inserted into a matrix where columns and rows represent the sequences $S_1, S_2, \dots, S_{2764}$. The diagonal elements of the matrix are zero because they represent the distance between equal server sessions.

Several hierarchical clustering methods like Ward, Single-, Complete-, Average-, or Centroid linkage (Hair et al, 1998; Kaufman and Rousseeuw, 1990) may be invoked on the distance matrices. The clustering methods that are used in the experiments are *agglomerative* which means that the algorithms start with n clusters and proceed by successive fusions until a single cluster is obtained holding all of the server sessions in the data set (Kaufman and Rousseeuw, 1990). *Divisive* clustering techniques proceed in the opposite order and are computationally more complex than agglomerative techniques. Future research discusses heuristics for divisive clustering techniques.

The distance between two points i and j is calculated by the clustering methods as follows (Kaufmann and Rousseeuw, 1990):

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 \dots + (x_{in} - x_{jn})^2} \quad (4.3)$$

where

- i = 1, 2, ..., n = server session i ;
- j = 1, 2, ..., n = server session j ;
- x_{ab} = SAM distance measure in distance matrix with $a = i$ or j and $b = 1, 2, \dots, n$;

Two clusters are joined together if the dissimilarity between them is minimal. In table 4.8, equations indicate how the dissimilarity between two clusters, C_a and C_b , is computed. For example, using single linkage, the dissimilarity between two clusters C_a and C_b is equal to the minimum distance between a pair of points, one in C_a and one in C_b . Single linkage is often called nearest neighbour method. Yet, complete linkage defines the dissimilarity between two clusters as the largest distance between one point in one cluster and one point in the other. In average linkage, the dissimilarity between two clusters is the average distance between pairs of server sessions. Furthermore, in the centroid method, the dissimilarity between two clusters is defined as the Euclidean distance between their centroids or means.

Clustering method	Dissimilarity [C_a, C_b]
Ward	$= \ \text{avg}(x(C_a)) - \text{avg}(x(C_b)) \ ^2 / [(1/n_{C_a}) + (1/n_{C_b})]$
Single linkage (nearest neighbour)	$= \min d(i, j)$
Complete linkage (furthest neighbour)	$= \max d(i, j)$
Average linkage	$= [(1/n_{C_a}) + (1/n_{C_b})] \sum_{i \in C_a} \sum_{j \in C_b} d(i, j)$
Centroid linkage	$= \ \text{avg}(x(C_a)) - \text{avg}(x(C_b)) \ ^2$
where $\text{avg}(x(C_a))$ is the average distance in cluster a; $\text{avg}(x(C_b))$ is the average distance in cluster b; n_{C_a} is the number of server sessions in cluster a; n_{C_b} is the number of server sessions in cluster b; $i \in C_a$; $j \in C_b$;	

Table 4.8: Computing dissimilarities between two clusters (Kaufmann and Rousseeuw, 1990).

Ward's method joins clusters with a small number of observations and is strongly biased towards producing clusters with roughly the same number of observations (Milligan, 1980). Because we are interested in web usage behaviour of large groups of visitors, Ward's method is chosen for further analyses.

It is important to define the right number of clusters because specifying too few ignores group differences, while specifying too many causes the model to be unstable or computational demands will be extremely high. There are many approaches for determining the number of clusters and none of them has been proven to be the best (Bock, 1985; Everitt, 1979; Hartigan, 1985). However,

several criteria have proven to be useful for defining a trade off between number of clusters and model fit. *R-squared* is used as a goodness-of-fit measure during clustering processing and equals to the proportion of variation explained by the model. R-squared ranges in values from zero to one. Obviously, the level of R-squared increases with the number of clusters. Small values of R-squared indicate that the model does not fit the data well, whereas measures of 0.6 and higher are considered acceptable (Hair et al, 1998). Ultimately, we will define a stop-criterion when the incremental values of R-squared flatten out if additional clusters are formed. Opposed to R-squared, *semi-partial R-squared* represents the decrease in the proportion of variance accounted for by joining two clusters. Furthermore, Cooper and Milligan (1988) and Milligan and Cooper (1985) have compared thirty methods for estimating the number of clusters using hierarchical clustering methods. The criteria that performed best in these simulation studies were *pseudo F statistic (PSF)*, developed by Calinski and Harabasz (1974) and *T-squared statistic (TST)*, originated by Duda and Hart (1973). Relatively large values given by the pseudo F statistic indicate a stopping point. A general rule for interpreting the values of the T-squared statistic is to move towards joining of clusters and find values markedly larger than previous values. Finally, another method for judging the number of clusters in a data set is the *root mean squared standard deviation (RMSSTD)*, which provides a measure of homogeneity for the cluster solution. The smaller this value, the more homogeneous are the clusters. Equations for calculating the criteria are given in table 4.9.

Information criterion	Equation
R-squared	$= 1 - [(\sum_{a=1}^G W_{Ca}) / (\sum_{i=1}^n \ x_i - \text{avg}(x)\ ^2)]$
Semi-partial R-squared	$= (W_{Cm} - W_{Ca} - W_{Cb}) / (\sum_{i=1}^n \ x_i - \text{avg}(x)\ ^2)$
Pseudo F statistic (PSF)	$= (1) / (2)$ $(1) = [(\sum_{i=1}^n \ x_i - \text{avg}(x)\ ^2) - \sum_{a=1}^G W_{Ca}] / G - 1$ $(2) = (\sum_{a=1}^G W_{Ca}) / n - G$
T-squared statistic (TST)	$= (W_{Cm} - W_{Ca} - W_{Cb}) / [(W_{Ca} + W_{Cb}) / (n_{Ca} + n_{Cb} - 2)]$
Root mean squared standard deviation (RMSSTD)	$= \sqrt{[W_{Ca} / v (n_{Ca} - 1)]}$
<p>where</p> <p>n is the number of server sessions;</p> <p>n_{Ca} is the number of server sessions in cluster a;</p> <p>n_{Cb} is the number of server sessions in cluster b;</p> <p>v is the number of variables;</p> <p>avg(x) is the average distance in the data set;</p> <p>avg(x(C_a)) is the average distance in cluster a;</p> <p>$W_{Cm} = \sum_{i \in C_m} \ x_i - \text{avg}(x(C_m))\ ^2$;</p> <p>$W_{Ca} = \sum_{i \in C_a} \ x_i - \text{avg}(x(C_a))\ ^2$;</p> <p>$W_{Cb} = \sum_{i \in C_b} \ x_i - \text{avg}(x(C_b))\ ^2$;</p> <p>$C_m = C_a \cup C_b$;</p> <p>G is G-th level of the hierarchy and the number of clusters for the summation;</p>	

Table 4.9: Information criteria for defining the number of clusters.

For the SAM applications on the data sets given in table 4.4, the values of the information criteria for defining the number of clusters are graphically presented in figures 4.14 to 4.16. Note that figures 4.14 till 4.16 are constructed using Ward's method (re. table 4.8) for hierarchical clustering. Future research discusses examining the influence of using other methods than Ward on the final clustering results. Looking for consensus among the criteria, the following cluster solutions are defined.

4.7.1 Defining the number of clusters for data set 1 (<http://www.luc.ac.be/tew>)

If server sessions, consisting of visited pages, of the first data set are clustered using SAM distance measures, all of the five criteria suggest six clusters (re. figure 4.14). The pseudo F statistic has the largest value for six clusters and the T-squared statistic, starting from seven clusters, rises when two clusters are joined together providing a solution of six clusters. Likewise, the value of R-squared indicates that 88.04% of the variance is explained by the model if six clusters are defined. Besides, the graph shows that the incremental values of R-squared flatten out if more than six clusters are formed. This is also shown by the values of semi-partial R-squared. The additional variance explained by the model from six to seven clusters reaches barely 0.0138, which is quite low. Likewise, the root mean squared standard deviation has a relatively low value for six clusters, indicating that the homogeneity of the data in six clusters is relatively high.

Yet, not only six clusters appear to be a good clustering solution, two clusters might be interesting as well. For practical reasons, if we are interested in adjusting the structure of the web site to the largest group of visitors, two clusters might be a good solution. Unfortunately, the variance in the data explained by two clusters is only 50.92%, which is below the minimum level (Hair et al., 1998). An overview of the number of server sessions for each cluster of data set 1 is given in appendix 4.

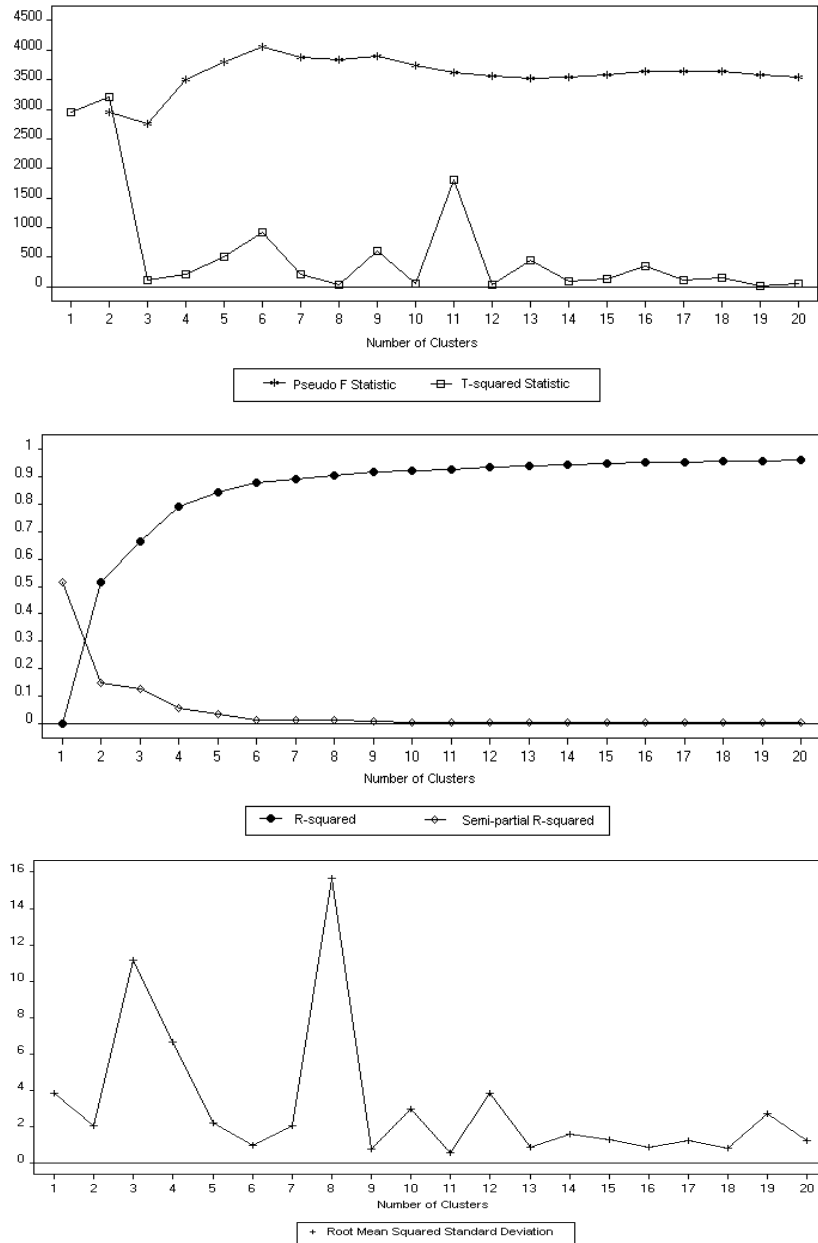


Figure 4.14: Applying SAM to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

4.7.2 Defining the number of clusters for data set 2 (<http://machines.hyperreal.org>)

If server sessions of data set two are clustered using SAM distance measures, a consensus is reached among all the criteria at five clusters (re. figure 4.15). The pseudo F statistic is relatively high and T-squared shows that five clusters are better than four or six. Also, R-squared indicates that five clusters are a good solution explaining 88% of the variance in the data. Additional variances incorporated in the model by a higher number of cluster solutions than five are very small, which indicate the level-out effect starting at more than five clusters. Likewise, root mean squared standard deviation is relatively low at this point, suggesting that five clusters are a good solution.

Yet, not only five clusters appear to be a good clustering solution, two clusters might be interesting as well. For practical reasons, if we are interested in adjusting the structure of the web site to the largest group of visitors, two clusters might be a good solution. Unfortunately 59.22% of the variance in the data is explained by two clusters, which is still below the minimum level (Hair et al., 1998). An overview of the number of server sessions for each cluster of data set 2 is given in appendix 4.

4.7.3 Defining the number of clusters for data set 3 (Belgian telecom provider)

If server sessions in data set three are clustered using SAM distance measures, T-squared statistic strongly suggests four clusters (re. figure 4.16). R-squared reaches a level of 0.64, which is satisfactory. Also, the root mean squared standard deviation is relatively low at this point. Although the F-statistic does not reach the highest value at four clusters, compared with the values of 772 other cluster solutions 439.99 is still relatively high. An overview of the number of server sessions for each cluster of data set 3 is given in appendix 4.

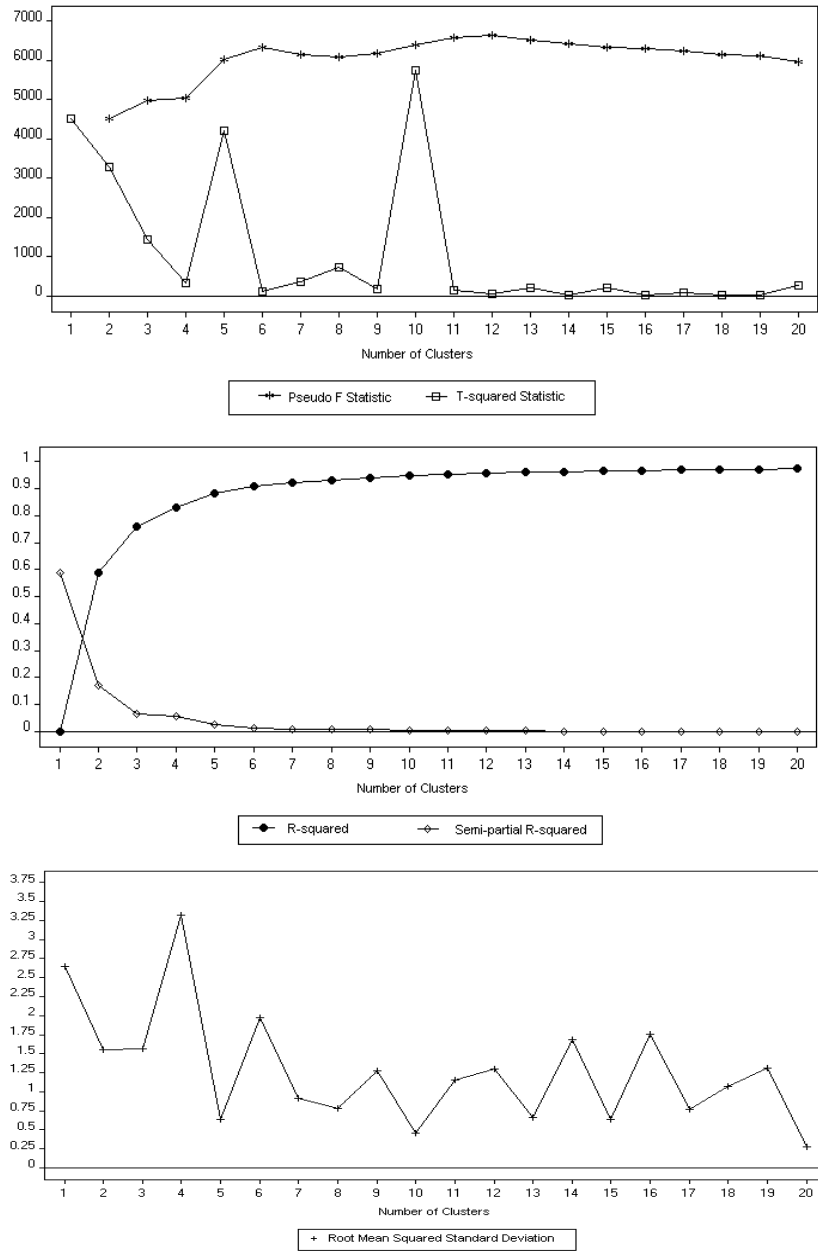


Figure 4.15: Applying SAM to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

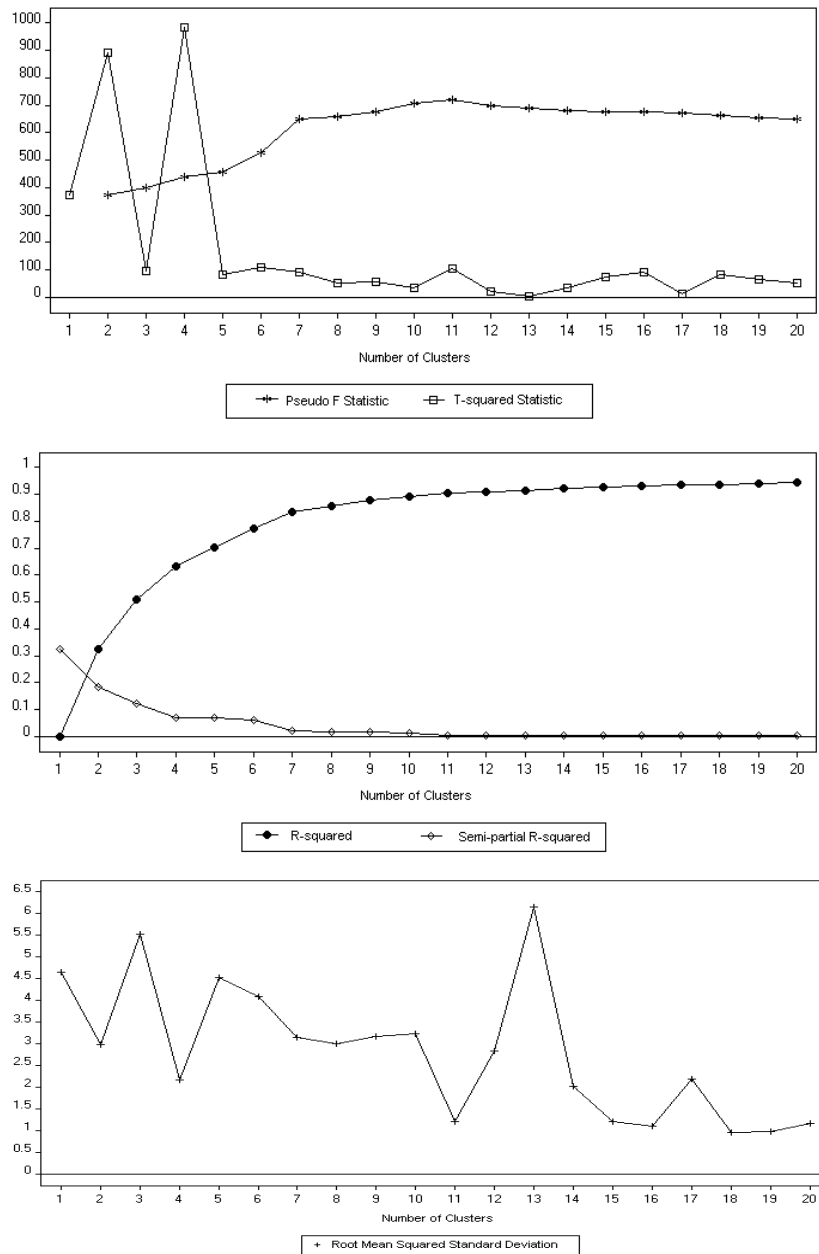


Figure 4.16: Applying SAM to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

4.7.4 *Remarks about information criteria*

Before proceeding to the third step of our approach of Web Usage Mining (re. figure 4.3), we provide some explications about the graphical presentations in figures 4.14 to 4.16. In figure 4.14, PSF reaches a peak at six clusters because dispersion in the data set (before clustering) is much higher than the sum of dispersions within six clusters, indicating a good clustering solution. TST from five to six clusters rises because $W_{C_m} > W_{C_a} + W_{C_b}$, which means that splitting up cases improves dispersion and homogeneity. Finally, in figure 4.16 PSF provides a lower value compared to TST at four clusters. The main reason is that on the one hand, PSF mainly investigates the difference between dispersion in the data set and dispersion for G clusters, with G being the level of the hierarchy and the number of clusters for the summation. On the other, TST investigates dispersion for two cluster solutions i.e. one before merging C_a with C_b and one after merging C_a with C_b , with $C_m = C_a \cup C_b$ and taking into account the number of server sessions in C_a and C_b .

4.8 Step 3: Post-processing

In the third step of our approach of Web Usage Mining (re. figure 4.3), every cluster is examined on *page_id*'s, the *order* in which *page_id*'s are requested and on the length of server sessions. Practically, every cluster represents visiting profiles providing a general view of not only page requests but also order-based information.

In this section, three techniques are used for cluster examination. First, clusters are graphically explored on *page_id*'s without considering the order of pages. Second, open sequences (Capri, 2001) are used to measure how well every cluster extracts order-based information from the data set. *Open sequences* are sequences with the same elements occurring in the same order and irrelevant of the positions of the elements. In Büchner et al (1999), open sequences are used to discover structural information within navigation patterns. For example, open sequence (68, 65, 55) indicates that page 65 is requested after page 68 and before page 55. Moreover, open sequence (68, 65, 55) is an element of the following server sessions:

- $\{(2, 68, 65, 70, 55); (t_0, t_1, t_1, t_1, t_2)\}$
- $\{(68, 55, 65, 33, 55, 22); (t_1, t_1, t_1, t_3, t_0, t_3)\}$
- $\{(68, 65, 55); (t_0, t_1, t_2)\}$

2-dimensional open sequences contain *page_id*'s and categories of visiting page time. For example, open sequence (68, 65, 55); (t1, t1, t2) indicates that page 68 and 65 are visited within time category t1 and that page 55 is visited within time category t2, without losing the order of visited pages. Only the first server session in the example above holds the 2-dimensional open sequence.

Open sequences are valued by means of two criteria: support and confidence. *Support* specifies the number of server sessions within a cluster presenting the open sequence divided by the total number of server sessions within that cluster. *Confidence* expresses the probability that, if a server session contains all but the last element (in respective order) of the open sequence, the server session will also hold the last element of the open sequence. Equations (4.4) to (4.7) show how support and confidence values for open sequences are calculated.

$$\text{Support } (p_1, p_2, \dots, p_e) = \sum_{i=1}^{n_{Ca}} S_i / n_{Ca} \quad (4.4)$$

$$\text{and } \begin{aligned} S_i &= 1 \text{ if } (p_1, p_2, \dots, p_e) \in_{\text{order}} S_i \\ S_i &= 0 \text{ otherwise} \end{aligned}$$

$$\text{Confidence } (p_1, p_2, \dots, p_e) = \sum_{i=1}^{n_{Ca}} S_i / \sum_{j=1}^{n_{Ca}} S_j \quad (4.5)$$

$$\text{and } \begin{aligned} S_i &= 1 \text{ if } (p_1, p_2, \dots, p_e) \in_{\text{order}} S_i \\ S_i &= 0 \text{ otherwise} \\ S_j &= 1 \text{ if } (p_1, p_2, \dots, p_{e-1}) \in_{\text{order}} S_j \\ S_j &= 0 \text{ otherwise} \end{aligned}$$

$$\text{Support } (p_1, p_2, \dots, p_e); (t_1, t_2, \dots, t_e) = \sum_{i=1}^{n_{Ca}} S_i / n_{Ca} \quad (4.6)$$

$$\text{and } \begin{aligned} S_i &= 1 \text{ if } (p_1, p_2, \dots, p_e); (t_1, t_2, \dots, t_e) \in_{\text{order}} S_i \\ S_i &= 0 \text{ otherwise} \end{aligned}$$

$$\text{Confidence } (p_1, p_2, \dots, p_e); (t_1, t_2, \dots, t_e) = \sum_{i=1}^{n_{Ca}} S_i / \sum_{j=1}^{n_{Ca}} S_j \quad (4.7)$$

$$\text{and } \begin{aligned} S_i &= 1 \text{ if } (p_1, p_2, \dots, p_e); (t_1, t_2, \dots, t_e) \in_{\text{order}} S_i \\ S_i &= 0 \text{ otherwise} \\ S_j &= 1 \text{ if } (p_1, p_2, \dots, p_{e-1}); (t_1, t_2, \dots, t_{e-1}) \in_{\text{order}} S_j \\ S_j &= 0 \text{ otherwise} \end{aligned}$$

where

S_i represents server session i ;

S_j represents server session j ;

n_{Ca} is the number of server sessions in cluster a ;

p_e is the last element of the page_id 's in the open sequence;

t_e is the last element of the time categories in the open sequence;

p_{e-1} is the one but last element of the page_id 's in the open sequence;

t_{e-1} is the one but last element of the time categories in the open sequence;

\in_{order} is an element of, respecting the order of occurrence of elements in the open sequence;

For example, support as well as confidence for open sequence (68, 65, 55) is 1 or 100% for cluster C_a holding the three server sessions presented above. For the 2-dimensional open sequence (68, 65, 55); (t_1, t_1, t_2) support is 0.33 or 33% while confidence is 0.5 or 50%. This means that, if page 65 is requested

after page 68 with category of visiting page time of t_1 for both pages, the chance that page 55 is requested AND visited during time t_2 is 50%.

We remark that, in our experiments, open sequences are used in the post-processing step, i.e. after clustering based on SAM distance measures. If open sequences were searched first, followed by SAM and clustering, we would be clustering open sequences instead of server sessions and might lose information embedded in server sessions. Moreover, applying alignment methods on open sequences instead of server sessions will treat open sequences with relatively high support and confidence the same way as open sequences with relatively low support and confidence. In the end, clusters may end up holding open sequences with relatively low support and confidence values. This means that a cutoff value needs to be defined to extract open sequences with high support and/or confidence values. Defining a cutoff value before the mining process takes place means that only part of the data will be analysed and valuable information might be lost. For this reason, we apply sequence alignment methods before open sequences and not the other way around. Nevertheless, further research should explore how open sequences might be used before the mining process.

Finally, a third way of looking at clusters is to examine the *length of server sessions* within the clusters. In order to know whether SAM is (in)sensitive to the length of the server sessions, clusters need to be analysed on the length of server sessions.

4.8.1 Examining clusters on *page_ids*

4.8.1.1 Data set 1 (<http://www.luc.ac.be/tew>)

Figures 4.17 to 4.22 are three-dimensional graphs presenting, for each cluster, the *page_id* on the horizontal axis and relative frequency or exclusivity on the vertical axis. *Relative frequency for page_id* x ($x = 1, 2, \dots, 71$) in cluster C_a ($a = 1, 2, \dots, 6$) is equal to the number of requests of x in C_a divided by the total number of requests in C_a , multiplied by 100. Relative frequencies provide a view of the distribution of pages within clusters. *Exclusivity for page_id* x in cluster C_a is defined as the number of requests of x in C_a divided by the total number of requests of x in the data set. Exclusivities of nearly one indicate that the clusters are well separated and that pages are exclusively represented in clusters. Exclusivities provide a view of the distribution of pages across clusters. Tables of relative frequencies and exclusivities for data set 1 are given in appendix 4.

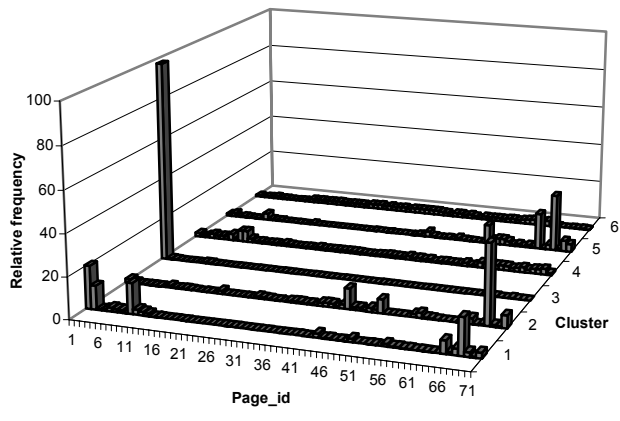


Figure 4.17: SAM applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Distribution of web pages in six clusters.

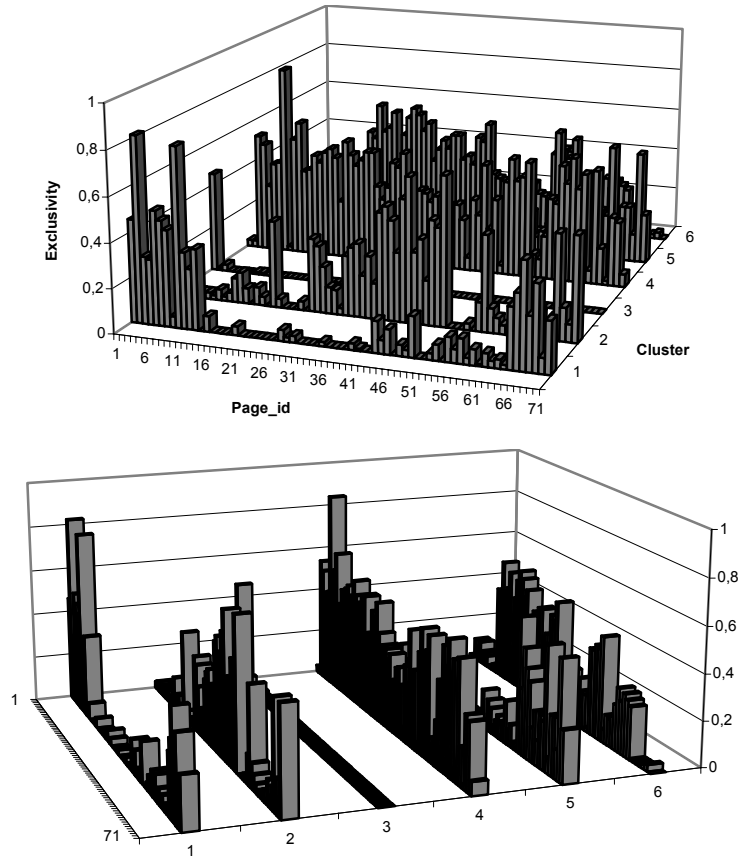


Figure 4.18: SAM applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Exclusivity of web pages in six clusters.

Figure 4.17 shows, if SAM is applied to server sessions consisting of visited pages in the first data set, different distributions of pages are provided within the clusters. 20.48% of the pages in cluster *one* are requests for page 1. Also, pages 68 (17.61%), 9 (14.83%) and 2 (11.35%) are well represented in cluster one. In cluster *two*, the most frequent request is page 68 (38.73%). Yet, 97.78% of the pages in cluster *three* are page 1. Cluster *four* and *six* provide a more or less equal spread among all of the 71 pages in the data set. Finally, cluster *five* shows peaks at page 68 (28.10%) and 65 (17.92%).

Figure 4.18 shows, if SAM is applied to server sessions consisting of visited pages in the first data set, exclusivities for most of the pages are relatively high in one of the six clusters. For example, 83% of pages 2 in the data set are grouped in the first cluster. Also, 80% of pages 9 in the data set are grouped in

the first cluster. In cluster two, page 49 and 43 show exclusivities of 0.67 and 0.65 respectively. Note that, although 97.78% of the pages in cluster three are page 1, exclusivity of page 1 in cluster three is 0.46, indicating that 46% of pages 1 in the data set are grouped in cluster three. Yet, exclusivities of other pages in cluster three are nearly zero. In cluster four, page 8 provides an exclusivity of 0.88. In cluster five, page 55, 65 and 70 show exclusivities of 0.59, 0.54 and 0.52 respectively. Finally, cluster six provides exclusivities between 0.50 and 0.55 for page 15, 18, 21, 22, 23 and 37.

4.8.1.2 Dataset 2 (<http://machines.hyperreal.org>)

Figures 4.19 to 4.20 present, for each cluster, groups of page_ids on the horizontal axis and relative frequency or exclusivity on the vertical axis. *Relative frequency for group y* ($y = 1, 2, \dots, 50$) in cluster C_a ($a = 1, 2, \dots, 5$) is equal to the number of requests of pages in group y within C_a divided by the total number of requests in C_a , multiplied by 100. Relative frequencies provide a view of the distribution of groups of pages within clusters. *Exclusivity for group y* in cluster C_a is defined as the number of requests of pages in group y within C_a divided by the total number of requests of pages in group y within the data set. Exclusivities of nearly one indicate that the clusters are well separated and that groups of pages are exclusively represented in clusters. Exclusivities provide a view of the distribution of groups of pages across clusters. Tables of relative frequencies and exclusivities for groups of pages in data set 2 are given in appendix 4.

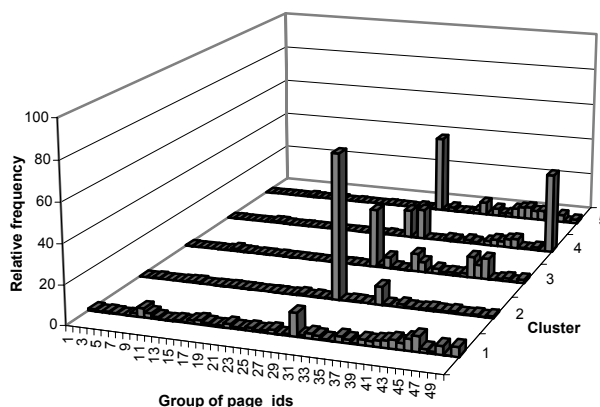


Figure 4.19: SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of groups of page_ids in five clusters.

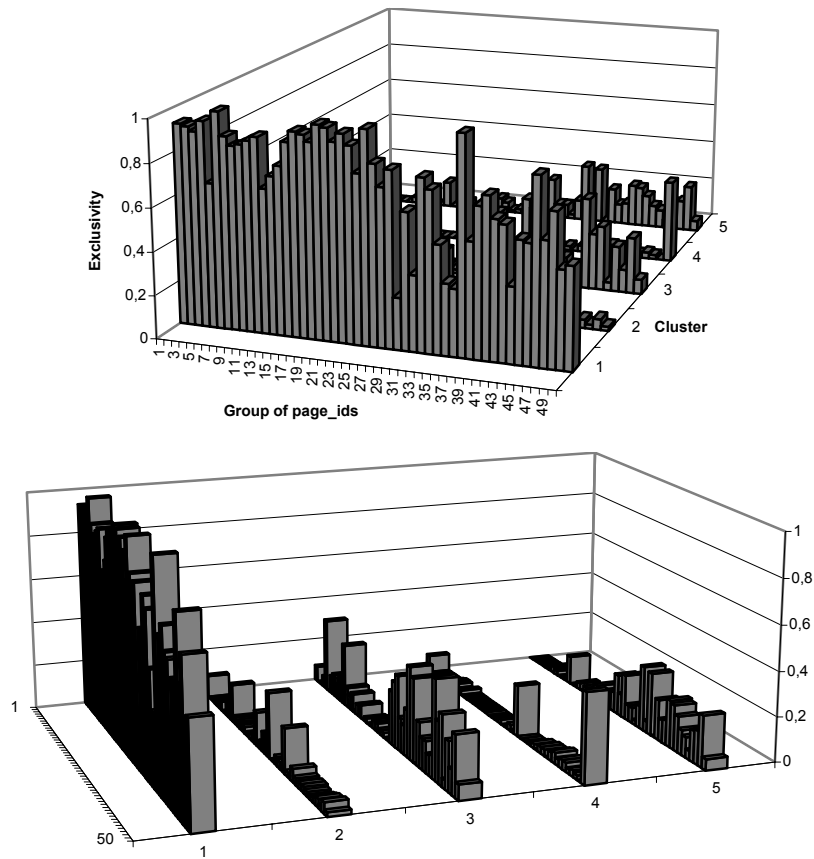


Figure 4.20: SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Exclusivity of groups of page_ids in five clusters.

Figure 4.19 shows, if SAM is applied to server sessions consisting of visited pages in the second data set, the distributions of pages within the clusters have one feature in common. Group 29, representing web pages 645 to 667 (including 645 and 667), is well represented in each cluster. The highest relative frequency for group 29 is shown by cluster *two*. Here, 73.53% of the pages are pages of or between 645-667. Compared with figure 4.6, 25.28% of the visited pages in the second data set are pages of group 29. In cluster *four*, high relative frequency is shown for group 50. Here, 41.43% of the requests are pages of or between 1128-1159.

Figure 4.20 shows, if SAM is applied to server sessions consisting of visited pages in the second data set, exclusivities for groups of page_ids in cluster one are relatively high compared to the four remaining clusters. In particular, exclusivities for groups of page_ids in cluster one are minimum 0.50 except for the following groups: 29 (0.24), 31 (0.35), 35 (0.33), 36 (0.31), 43 (0.35), 49 (0.45) and 50 (0.47). Also, exclusivities equal to 1 are shown in cluster one for group 6 and 37. In the remaining four clusters, exclusivities for groups of page_ids are relatively low. However, this does not mean that clusters are not well separated, because high exclusivities for individual pages are shown in every cluster. For example, exclusivities of 1 are shown for the following pages in the following clusters: page 105 (cluster three), page 284 and 287 (cluster two) and page 288 (cluster five). The reason why exclusivities for groups of page_ids may differ with exclusivities for individual pages is given by the following example. Suppose page x ($x = 1, 2, \dots, 5$) \in group 1 and the number of requests for page x in the data set and in two clusters derived from the data set is given in table 4.10. Exclusivity for group 1 is 0.9 in the first cluster and 0.1 in the second. Yet, page 5 is exclusively represented in cluster two. Although care must be taken when interpreting exclusivities for groups of page_ids, the approach is used for data sets with a relatively large number of different page_ids (for example, dataset 2 and 3 consist of 1,159 and 492 different page_ids respectively). Since we are unable to provide relative frequencies and exclusivities for each individual page_id, groups of page_ids are used.

Group	Page x	Number of requests		
		Data set	Cluster	
			1	2
1	1	15	15	0
	2	3	3	0
	3	0	0	0
	4	0	0	0
	5	2	0	2
Total number of requests		20	18	2
Exclusivity		Group 1	0.9	0.1
		Page 1	1	0
		Page 2	1	0
		Page 3	0	0
		Page 4	0	0
		Page 5	0	1

Table 4.10: Differences between exclusivities for a group of pages and for individual pages.

4.8.1.3 Dataset 3 (Belgian telecom provider)

Figures 4.21 to 4.22 present, for each cluster, groups of `page_ids` on the horizontal axis and relative frequency or exclusivity on the vertical axis. *Relative frequency for group z* ($z = 1, 2, \dots, 20$) in cluster C_a ($a = 1, 2, 3, 4$) is equal to the number of requests of pages in group z within C_a divided by the total number of requests in C_a , multiplied by 100. *Exclusivity for group z* in cluster C_a is defined as the number of requests of pages in group z within C_a divided by the total number of requests of pages in group z within the data set. Tables of relative frequencies and exclusivities for groups of pages in data set 3 are given in appendix 4.

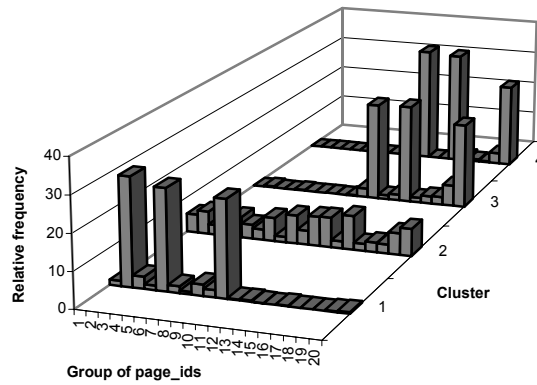


Figure 4.21: SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of groups of `page_ids` in four clusters.

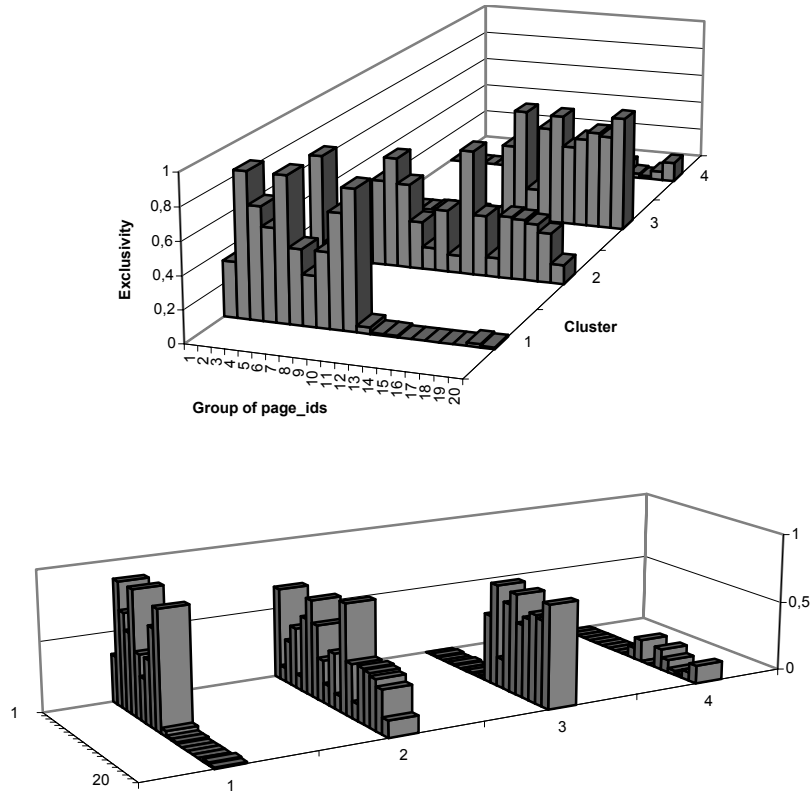


Figure 4.22: SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Exclusivity of groups of page_ids in four clusters.

Figure 4.21 shows, if SAM is applied to server sessions consisting of visited pages in the third data set, cluster *one* mainly represents pages of group 2 (i.e. page_ids of or between 51 and 75), 5 (i.e. page_ids of or between 101 and 125) and 10 (i.e. page_ids of or between 226 and 250). In cluster *two*, a more or less equal spread of the distribution of groups of page_ids is shown. Finally, cluster *three* and *four* show approximately the same distribution for groups of page_ids, concentrated on group 12 (i.e. page_ids of or between 276 and 300), 15 (i.e. page_ids of or between 351 and 375) and 20 (i.e. page_ids of or between 476 and 492).

Figure 4.22 shows, if SAM is applied to server sessions consisting of visited pages in the third data set, exclusivities for groups of page_ids in clusters one,

two and three are relatively high compared to cluster four. Yet, cluster four shows exclusivities equal to 1 for page 389 and 390. The reason why exclusivities for groups of page_ids may differ from exclusivities for individual pages is given in the previous section.

With regard to values of exclusivities for web pages, we must be careful with interpretations. Exclusivity measures for web pages provide an indication of the occurrence of page x in cluster a and the absence of page x in cluster b. Also, exclusivity measures do not provide any information about the order of pages. This means that, in the extreme situation of the next example of clusters holding the following sequences, low exclusivities do not necessarily mean that the data set is badly split. In terms of order-based information, the data set is split up perfectly.

Cluster a: 1, 2, 3

Cluster b: 3, 2, 1

Cluster c: 1, 3, 2

4.8.2 Examining clusters on the order of page_ids

In order to measure how well every cluster extracts order-based information from the data set, open sequences are calculated for every cluster using the software Capri (2001). For each data set and for each cluster, all possible open sequences with minimum support or confidence of 1% and minimum length of two elements are calculated. Open sequences having the *five highest support values* are selected for cluster description. Obviously, due to the high support demand, most of these selected sequences are quite short of length. Hence, to provide more order-based information, also open sequences having the *five highest confidence values* are selected for cluster description. If more than five open sequences were found, all of them showing the same high confidence values, two additional selection criteria were applied, based on the longest open sequences and on the highest support values. For example, after clustering server sessions, consisting of visited pages, based on SAM in the first data set, cluster five provided 119 open sequences having confidence values of 100% and length of three to eight elements. Some of them are given below:

- 1 (69, 2, 9) Support = 1.08; Confidence = 100.00
- ...
- 14 (58, 68, 2, 9) Support = 1.08; Confidence = 100.00
- ...
- 47 (67, 55, 57, 56, 58) Support = 1.08; Confidence = 100.00
- ...
- 95 (65, 57, 56, 58, 59, 60) Support = 1.44; Confidence = 100.00

- ...
- 112 (1, 68, 65, 55, 70, 9, 2) Support = 1.08; Confidence = 100.00
- 113 (65, 55, 57, 56, 58, 59, 60) Support = 1.08; Confidence = 100.00
- 114 (68, 55, 57, 56, 58, 59, 60) Support = 1.08; Confidence = 100.00
- 115 (68, 65, 55, 56, 58, 59, 60) Support = 1.08; Confidence = 100.00
- 116 (68, 65, 55, 57, 58, 59, 60) Support = 1.08; Confidence = 100.00
- 117 (68, 65, 55, 70, 57, 56, 58) Support = 1.81; Confidence = 100.00
- 118 (68, 65, 57, 56, 58, 59, 60) Support = 1.44; Confidence = 100.00
- 119 (68, 65, 55, 57, 56, 58, 59, 60) Support = 1.08; Confidence = 100.00

Instead of presenting all of these open sequences to show order-based information in cluster five, we first selected the longest one (i.e. 119 of eight elements long). Then, open sequences 117 and 118 are selected because they are seven elements long and represent the highest support. Following, because open sequences 112 till 116 have the same support value, all of them are selected for cluster description. For each data set and for each cluster, the selected open sequences along with support and confidence values are given below.

4.8.2.1 Data set 1 (<http://www.luc.ac.be/tew>)

In table 4.11, six visiting profiles, providing order-based information of visited pages at <http://www.luc.ac.be/tew>, are given by open sequences in clusters. Cluster *one* mainly represents visiting profiles to page 1, followed by pages 9, 68 or 2. Also, page 68 is followed by page 9 or 65. Cluster *two* represents visiting profiles to page 68 followed by pages 43, 49, 71 and 65. Also, page 49 is visited before page 47 in 5% of the server sessions. More than 90% of the server sessions in cluster *three* consist of only one page and represent visiting profiles to page 1. Other server sessions hold multiple consecutive values of the same element, page 1. For this reason, no open sequences with minimum two elements are found. Likewise, cluster *four* mainly groups server sessions consisting of one page. This page can be any page of the web site, for example page 11, 55, 47, 59, 60 or 63. Cluster *five* mainly represents visiting profiles to page 68, followed by pages 65 and/or 55 and/or 70. Cluster *six* provides information about visiting profiles starting with pages 59, 28, 26 and 40. Note that the profile of page 68 followed by page 65 is represented in cluster one (support = 25.27%), two (support = 6.31%) and five (support = 89.17%). Yet, each cluster represents a unique visiting profile because (68, 65) is related with pages 1 and 9 in cluster one, with pages 43, 49 and 71 in cluster two and with pages 55 and 70 in cluster three.

Besides open sequences with high support values, open sequences are also selected on high confidence values to provide more order-based information for every cluster. Such open sequences identify the probability that a particular page is visited following a certain visiting profile. For example, in data set 1, the probability that a visitor goes to page 65 after having followed the pattern, respecting the order, 69, 2, 66, 71, 12, 43 is 100% (cluster one).

For evaluation purposes, support and confidence values of the open sequences selected for describing order-based information within each cluster, are also given for the other clusters in table 4.12. The support and confidence values of the open sequences used to describe clusters in table 4.11 are written in bold and represent the cluster that is printed at the head of the columns. For example, page 1 followed by page 9 (1st row) represents cluster one. Page 68 followed by page 43 (14th row) represents cluster two. In general, we may state that the more zero support values at the non-diagonal places (or the more zero support values not printed in bold) in table 4.12, the better the model fits the data, i.e. the better open sequences printed in bold represent order-based information within clusters.

With regard to table 4.12 we give three remarks. First, no open sequences were found for describing order-based information in cluster three and four, since 73.59% and 73.61% of the server sessions are one page long. This information is presented in the table with ‘one-page sessions’ in the first column instead of open sequences. Also, to compare the percentage of one-page sessions in cluster three and four with other clusters, support values are given for server sessions of one page long. Second, at first sight, it might seem that open sequence (68, 65) does not strongly represent cluster one, since the support value is 25.27% for cluster one and 89.17% for cluster five. However, cluster one and five represent different navigation patterns related to pages 68 and 65. Cluster one represents navigations regarding pages 68 and 65 along with navigations to pages 1, 9 and 2. Cluster five represents navigations regarding pages 68 and 65 along with navigations to pages 55 and 70. This means that, server sessions in cluster one holding pages 68, followed by 65, also hold pages 1, 9 and/or 2. Likewise, server sessions in cluster five holding pages 68, followed by 65, also hold pages 55 and/or 70. Although not all of the numbers outside the diagonal have zero support values, we may say that most of them do (or are relatively small values) and that the model fits the data well. Third, four cells provide support values not printed in bold and higher than 10% (i.e. open sequence (68, 9) in cluster five, (68, 71) in cluster one and five, (33, 42) in cluster six). All other cells provide support values not printed in bold and zero or less than 10%, which indicates that order-based information is well represented by each cluster.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(1, 9)	50.11	68.62
	(1, 68)	49.68	68.04
	(1, 2)	38.54	52.79
	(68, 9)	29.98	43.61
	(68, 65)	25.27	36.76
	(69, 2, 66, 71, 12, 43, 65)	1.28	100.00
	(69, 2, 71, 12, 43, 65)	1.28	100.00
	(69, 2, 66, 12, 43, 65)	1.28	100.00
	(69, 2, 66, 71, 43, 65)	1.28	100.00
	(69, 2, 66, 71, 12, 65)	1.28	100.00
	(69, 2, 66, 71, 12, 43)	1.28	100.00
	(69, 66, 71, 12, 43, 65)	1.28	100.00
(2, 66, 71, 12, 43, 65)	1.28	100.00	
2	(68, 43)	19.54	23.65
	(68, 49)	14.15	17.13
	(68, 71)	11.85	14.34
	(68, 65)	6.31	7.64
	(49, 47)	5.38	28.46
	(1, 68, 43)	2.46	100.00
	(68, 43, 33, 42)	1.69	78.57
	(68, 33, 42)	1.69	78.57
	(43, 33, 42)	1.69	78.57
	(33, 42)	1.69	78.57
3	One-page sessions		
4	(8, 11)	1.39	26.92
	(1, 11)	1.29	48.15
5	(68, 65)	89.17	90.15
	(68, 55)	45.85	46.35
	(65, 55)	36.46	40.40
	(68, 65, 55)	36.10	40.49
	(68, 55, 70)	27.44	59.84
	(68, 65, 55, 57, 56, 58, 59, 60)	1.08	100.00
	(68, 65, 55, 70, 57, 56, 58)	1.81	100.00
	(68, 65, 57, 56, 58, 59, 60)	1.44	100.00
	(1, 68, 65, 55, 70, 9, 2)	1.08	100.00
	(65, 55, 57, 56, 58, 59, 60)	1.08	100.00
	(68, 55, 57, 56, 58, 59, 60)	1.08	100.00
	(68, 65, 55, 56, 58, 59, 60)	1.08	100.00
(68, 65, 55, 57, 58, 59, 60)	1.08	100.00	
6	(59, 60)	64.00	94.12
	(59, 63)	60.00	88.24
	(28, 35)	56.00	87.50
	(28, 33)	56.00	87.50
	(28, 30)	56.00	87.50
	(26, 35)	56.00	87.50
	(26, 33)	56.00	87.50
	(26, 38)	56.00	87.50
	(26, 30)	56.00	87.50
	(40, 35)	56.00	87.50
	(40, 33)	56.00	87.50
	(40, 30)	56.00	87.50
	(42, 14, 21, 13, 25, 64, 22, 18, 53, 54)	20.00	100.00
	(42, 14, 21, 13, 16, 64, 22, 18, 53, 54)	20.00	100.00
	(42, 14, 21, 20, 25, 64, 22, 18, 53, 54)	20.00	100.00
	(42, 14, 21, 20, 16, 64, 22, 18, 53, 54)	20.00	100.00
	(42, 14, 21, 13, 25, 64, 22, 18, 51)	20.00	100.00

Table 4.11: SAM applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Open sequences with high support or confidence values within six clusters.

Open sequences	1		2		3		4		5		6	
	S	C	S	C	S	C	S	C	S	C	S	C
(1, 9)	50.11	68.62	0.00	0.00	0.59	0.59	0.20	7.41	2.17	40.00	8.00	66.67
(1, 68)	49.68	68.04	2.46	50.00	0.59	0.59	0.50	18.52	3.97	73.33	4.00	33.33
(1, 2)	38.54	52.79	0.00	0.00	0.59	0.59	0.00	0.00	2.53	46.67	4.00	33.33
(68, 9)	29.98	43.61	2.62	3.17	0.00	0.00	0.10	2.86	16.97	17.15	4.00	50.00
(68, 65)	25.27	36.76	6.31	7.64	0.30	50.00	0.69	20.00	89.17	90.15	8.00	100.00
(69, 2, 66, 71, 12, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 2, 71, 12, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 2, 66, 12, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 2, 66, 71, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 2, 66, 71, 12, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 2, 66, 71, 12, 43)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(69, 66, 71, 12, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(2, 66, 71, 12, 43, 65)	1.28	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(68, 43)	7.28	10.59	19.54	23.65	0.00	0.00	0.69	20.00	14.08	14.23	4.00	50.00
(68, 49)	6.21	9.03	14.15	17.13	0.00	0.00	0.10	2.86	5.78	5.84	8.00	100.00
(68, 71)	10.28	14.95	11.85	14.34	0.00	0.00	0.40	11.43	18.77	18.98	0.00	0.00
(68, 65)	25.27	36.76	6.31	7.64	0.30	50.00	0.69	20.00	89.17	90.15	8.00	100.00
(49, 47)	0.86	11.76	5.38	28.46	0.00	0.00	0.00	0.00	1.81	31.25	0.00	0.00
(1, 68, 43)	4.93	9.91	2.46	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(68, 43, 33, 42)	0.00	0.00	1.69	78.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(68, 33, 42)	0.00	0.00	1.69	78.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(43, 33, 42)	0.00	0.00	1.69	78.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(33, 42)	0.00	0.00	1.69	78.57	0.00	0.00	0.00	0.00	0.00	0.00	12.00	18.75
One-page sessions	7.71	-	41.69	-	73.59	-	73.61	-	0.00	-	0.00	-
(8, 11)	0.00	0.00	0.15	100.00	0.00	0.00	1.39	26.92	0.00	0.00	0.00	0.00
(1, 11)	5.57	7.62	0.00	0.00	0.00	0.00	1.29	48.15	0.00	0.00	0.00	0.00
(68, 65)	25.27	36.76	6.31	7.64	0.30	50.00	0.69	20.00	89.17	90.15	8.00	100.00
(68, 55)	5.35	7.79	4.00	4.81	0.00	0.00	0.20	5.71	45.85	46.35	0.00	0.00
(65, 55)	4.28	15.62	1.08	14.58	0.00	0.00	0.10	3.23	36.46	40.40	0.00	0.00
(68, 65, 55)	4.07	16.10	0.92	14.63	0.00	0.00	0.10	14.29	36.10	40.49	0.00	0.00
(68, 55, 70)	2.14	40.00	1.69	42.31	0.00	0.00	0.00	0.00	27.44	59.84	0.00	0.00
(68, 65, 55, 57, 56, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(68, 65, 55, 70, 57, 56, 58)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.81	100.00	0.00	0.00

Open sequences	1		2		3		4		5		6	
	S	C	S	C	S	C	S	C	S	C	S	C
(68, 65, 57, 56, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.44	100.00	0.00	0.00
(1, 68, 65, 55, 70, 9, 2)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(65, 55, 57, 56, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(68, 55, 57, 56, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(68, 65, 55, 56, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(68, 65, 55, 57, 58, 59, 60)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(59, 60)	0.00	0.00	0.15	33.33	0.00	0.00	0.99	30.30	2.17	20.00	64.00	94.12
(59, 63)	0.00	0.00	0.15	33.33	0.00	0.00	0.79	24.24	1.81	16.67	60.00	88.24
(28, 35)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(28, 33)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(28, 30)	0.00	0.00	0.00	0.00	0.00	0.00	0.30	20.00	0.00	0.00	56.00	87.50
(26, 35)	0.00	0.00	0.15	5.88	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(26, 33)	0.00	0.00	0.00	0.00	0.00	0.00	0.10	5.00	0.00	0.00	56.00	87.50
(26, 38)	0.00	0.00	0.62	23.53	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(26, 30)	0.00	0.00	0.15	5.88	0.00	0.00	0.40	20.00	0.00	0.00	56.00	87.50
(40, 35)	0.00	0.00	0.46	9.68	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(40, 33)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(40, 30)	0.00	0.00	0.15	3.23	0.00	0.00	0.00	0.00	0.00	0.00	56.00	87.50
(42, 14, 21, 13, 25, 64, 22, 18, 53, 54)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	100.00
(42, 14, 21, 13, 16, 64, 22, 18, 53, 54)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	100.00
(42, 14, 21, 20, 25, 64, 22, 18, 53, 54)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	100.00
(42, 14, 21, 20, 16, 64, 22, 18, 53, 54)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	100.00
(42, 14, 21, 13, 25, 64, 22, 18, 51)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.00	100.00

Table 4.12: SAM applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.8.2.2 Data set 2 (<http://machines.hyperreal.org>)

In table 4.13, five visiting profiles, providing order-based information of visited pages at <http://machines.hyperreal.org>, are given by open sequences in clusters. Cluster one and two represent visiting profiles of mostly one-page sessions. For this reason, open sequences with low support measures are found for cluster description. In cluster *one* almost every web page is represented. For example, pages 657, 947, 804, 190, 338, 1153, 1082 or 933. Yet, in cluster *two*, not every web page is represented. Instead, pages 657 or 802 are visited. Cluster *three* represents visiting profiles to and from page 657. Cluster *four* represents visiting profiles from page 1129 to 657, 713 or 1026. Also page 657 is visited followed by page 713 and in reverse order. Note that page 713 also occurs in the profiles given by cluster three. Yet, profiles in cluster three show no relation with page 1129. In cluster *five*, visiting profiles regarding page 657 related with pages 815, 947 and 984 are presented. Note that server sessions in cluster three also contain pages 815 and 984. However, the difference between cluster five and cluster three is that profiles in cluster three are not related to page 947.

With regard to open sequences selected on high confidence values, we give some examples of page prediction. The probability that visitors proceed back to page 657 is 100% if they follow the pattern (and respect the order of pages) 713, 657, 713 (cluster three). Also, the probability of re-visiting page 1129 is 100% after pattern 1129, 996 or 1129, 947 is followed (cluster four).

For evaluation purposes, support and confidence values of the open sequences selected for describing order-based information within each cluster, are also given for the other clusters in table 4.14. The support and confidence values of the open sequences used to describe clusters in table 4.13 are written in bold and represent the cluster that is printed at the head of the columns. Support values of open sequences (657, 947), (657, 984), (657, 815) and (657, 713) are written in bold more than once, indicating that they are used for more than one cluster description. Yet, each cluster describes different navigations related with these open sequences. Generally, 7 cells are found with relatively high support values for open sequences, which are elsewhere (in other columns) not selected for cluster description (i.e. not printed in bold). For example, (657, 1082), (657, 933) and (933, 657) show support of 10.53%, 20.00% and 12.63% respectively in cluster five while being selected for describing cluster one. Unfortunately, support values in cluster one are very low (i.e. around 1.00%), which indicates that cluster one is not well represented by open sequences. The main reason is that almost every web page is represented and more than 50% of the server sessions are one-page sessions in cluster one. Finally, in general, we believe that most of the support values at

the non-diagonal places (or not printed in bold) in table 4.14 are zero values or relatively low values, indicating that, except for cluster one, order-based information within clusters is well represented by open sequences.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(657, 947)	1.79	10.03
	(804, 190)	1.69	82.05
	(338, 1153)	1.37	81.25
	(657, 1082)	1.16	6.49
	(657, 933)	1.16	6.49
	(933, 657)	1.00	26.76
2	(802, 657)	6.68	76.47
	(657, 802)	2.91	3.16
	(657, 802, 657)	1.71	58.82
3	(802, 657, 802)	1.54	38.08
	(657, 984)	18.51	25.48
	(657, 794)	12.43	17.11
	(657, 815)	11.60	15.97
	(984, 657)	11.33	37.61
	(657, 713)	10.22	14.07
	(657, 713, 657, 713, 657)	1.10	100.00
	(713, 657, 713, 657)	1.38	100.00
	(713, 815, 657)	1.10	100.00
	(990, 984)	1.38	100.00
	(1153, 698)	1.10	100.00
	4	(1129, 657)	13.98
(1129, 713)		9.68	14.40
(657, 713)		7.53	26.92
(713, 657)		5.38	17.86
(1129, 1026)		5.38	8.00
(1129, 947, 1129, 1103, 1129)		1.08	100.00
(947, 1129, 1103, 1129)		1.08	100.00
(1129, 947, 1103, 1129)		1.08	100.00
(1129, 1026, 996, 1129)		1.08	100.00
(1129, 996, 1129)		2.69	100.00
(1129, 947, 1129)		2.69	100.00
5	(815, 657)	33.68	96.97
	(657, 815)	33.68	34.41
	(657, 815, 657)	32.63	96.88
	(657, 947)	24.21	24.73
	(947, 657)	21.05	86.96
	(657, 984)	21.05	21.51
	(657, 947, 657)	21.05	86.96
	(815, 657, 813, 657)	5.26	100.00
	(933, 657, 933, 657)	5.26	100.00
	(657, 815, 984, 657)	5.26	100.00
	(815, 657, 984, 657)	6.32	100.00
	(815, 657, 815, 657)	11.58	100.00
	(657, 933, 657, 933, 657)	5.26	100.00
	(657, 815, 657, 815, 657)	10.53	100.00

Table 4.13: SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Open sequences with high support or confidence values within five clusters.

Open sequences	1		2		3		4		5	
	S	C	S	C	S	C	S	C	S	C
One-page sessions	53.87	-	83.90	-	17.68	-	59.14	-	3.16	-
(657, 947)	1.79	10.03	0.00	0.00	3.59	4.94	0.00	0.00	24.21	24.73
(804, 190)	1.69	82.05	0.17	100.00	0.55	50.00	0.00	0.00	1.05	33.33
(338, 1153)	1.37	81.25	0.34	100.00	0.55	100.00	0.00	0.00	0.00	0.00
(657, 1082)	1.16	6.49	0.00	0.00	1.66	2.28	0.00	0.00	10.53	10.75
(657, 933)	1.16	6.49	0.00	0.00	1.38	1.90	0.00	0.00	20.00	20.43
(933, 657)	1.00	26.76	0.34	66.67	0.55	22.22	0.54	25.00	12.63	63.16
(802, 657)	0.11	100.00	6.68	76.47	0.55	100.00	0.00	0.00	0.00	0.00
(657, 802)	0.00	0.00	2.91	3.16	0.00	0.00	0.00	0.00	0.00	0.00
(657, 802, 657)	0.00	0.00	1.71	58.82	0.00	0.00	0.00	0.00	0.00	0.00
(802, 657, 802)	0.00	0.00	1.54	38.08	0.00	0.00	0.00	0.00	0.00	0.00
(657, 984)	0.95	5.31	0.34	0.37	18.51	25.48	0.54	1.92	21.05	21.51
(657, 794)	0.21	1.18	0.17	0.19	12.43	17.11	1.08	3.85	9.47	9.68
(657, 815)	0.84	4.72	0.51	0.56	11.60	15.97	0.54	1.92	33.68	34.41
(984, 657)	0.58	22.00	0.34	66.67	11.33	37.61	0.00	0.00	20.00	90.48
(657, 713)	0.21	1.18	0.17	0.19	10.22	14.07	7.53	26.92	5.26	5.38
(657, 713, 657, 713, 657)	0.00	0.00	0.00	0.00	1.10	100.00	0.00	0.00	0.00	0.00
(713, 657, 713, 657)	0.00	0.00	0.00	0.00	1.38	100.00	0.00	0.00	0.00	0.00
(713, 815, 657)	0.00	0.00	0.00	0.00	1.10	100.00	0.00	0.00	0.00	0.00
(990, 984)	0.00	0.00	0.00	0.00	1.38	100.00	0.00	0.00	1.05	100.00
(1153, 698)	0.21	12.50	0.00	0.00	1.10	100.00	0.00	0.00	0.00	0.00
(1129, 657)	0.26	9.26	0.00	0.00	1.10	66.67	13.98	20.80	6.32	75.00
(1129, 713)	0.05	1.85	0.00	0.00	0.55	33.33	9.68	14.40	1.05	12.50
(657, 713)	0.21	1.18	0.17	0.19	10.22	14.07	7.53	26.92	5.26	5.38
(713, 657)	0.21	15.38	0.17	50.00	8.01	67.44	5.38	17.86	6.32	85.71
(1129, 1026)	0.11	3.70	0.17	50.00	0.00	0.00	5.38	8.00	2.11	25.00
(1129, 947, 1129, 1103, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(947, 1129, 1103, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(1129, 947, 1103, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(1129, 1026, 996, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	1.08	100.00	0.00	0.00
(1129, 996, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	2.69	100.00	0.00	0.00
(1129, 947, 1129)	0.00	0.00	0.00	0.00	0.00	0.00	2.69	100.00	0.00	0.00
(815, 657)	0.74	48.28	0.34	66.67	9.39	58.62	0.00	0.00	33.68	96.97

Open sequences	1		2		3		4		5	
	S	C	S	C	S	C	S	C	S	C
(657, 815)	0.84	4.72	0.51	0.56	11.60	15.97	0.54	1.92	33.68	34.41
(657, 815, 657)	0.42	50.00	0.34	66.67	7.73	66.67	0.00	0.00	32.63	96.88
(657, 947)	1.79	10.03	0.00	0.00	3.59	4.94	0.00	0.00	24.21	24.73
(947, 657)	0.84	13.91	0.34	40.00	2.76	62.50	0.00	0.00	21.05	86.96
(657, 984)	0.95	5.31	0.34	0.37	18.51	25.48	0.54	1.92	21.05	21.51
(657, 947, 657)	0.42	23.53	0.00	0.00	1.93	53.85	0.00	0.00	21.05	86.96
(815, 657, 813, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	100.00
(933, 657, 933, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	100.00
(657, 815, 984, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	100.00
(815, 657, 984, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.32	100.00
(815, 657, 815, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	11.58	100.00
(657, 933, 657, 933, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.26	100.00
(657, 815, 657, 815, 657)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.53	100.00

Table 4.14: Data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.8.2.3 Data set 3 (Belgian telecom provider)

In table 4.15, four visiting profiles, providing order-based information of visited pages at the web site of a telecom provider, are given by open sequences in clusters. Cluster *one* represents visiting profiles starting with pages 28 and 27. Cluster *two* shows visiting profiles with regard to pages 281, 286, 305, 317, 355, 368 and 372. Surprising is the relatively strong distinction of order based relations between pages 281 and 280 in cluster three and four. In cluster *three*, 97.39% of the server sessions provide pattern 281, followed by 280. Yet, in cluster *four*, 100% of the server sessions provide pattern 280, followed by 281. With regard to open sequences selected by high confidence values, every cluster provides information for page predication with probability of 100%.

In table 4.16, generally, most of the support values at the non-diagonal places (or not printed in bold) are zero values or relatively low values, indicating that, order-based information within clusters is well represented by open sequences. However, one remark is given. Open sequences (281, 355) and (280, 355) are well represented in cluster three and four. However, different navigations are presented in each cluster. Cluster three groups server sessions holding page 281 followed by 280, followed by 355. Yet, cluster four groups server sessions holding page 280, followed by 281, followed by 355.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(28, 109)	97.39	99.12
	(27, 109)	85.65	87.17
	(27, 250)	84.78	86.28
	(109, 250)	84.78	85.53
	(28, 250)	83.91	85.40
	(28, 27, 250, 113, 123, 138, 181, 126, 161, 176)	1.30	100.00
	(28, 27, 250, 113, 127, 138, 181, 126, 161, 176)	1.30	100.00
	(28, 27, 109, 250, 234, 227, 237, 244, 230, 239)	1.30	100.00
	(28, 109, 250, 113, 127, 123, 138, 181, 126, 161)	1.30	100.00
	(27, 109, 250, 113, 127, 123, 138, 181, 126, 161)	1.30	100.00
2	(305, 317)	6.49	89.47
	(286, 317)	5.73	83.33
	(305, 286)	4.58	63.16
	(372, 368)	3.82	76.92
	(281, 355)	3.82	47.62
	(281, 280, 372, 386)	1.15	100.00
	(286, 305, 317)	1.91	100.00
	(21, 22, 17, 19)	1.53	100.00
	(21, 18, 17, 19)	1.53	100.00
	(18, 17, 19)	1.53	100.00
	(22, 17, 19)	1.53	100.00
	(21, 17, 19)	1.53	100.00
	(196, 186, 194)	1.53	100.00
3	(281, 355)	97.39	98.68
	(281, 280)	97.39	98.68
	(280, 355)	85.22	85.96
	(281, 280, 355)	83.91	86.16
	(281, 492)	83.04	84.14
	(280, 492)	83.04	83.77
	(280, 355, 492, 491, 372, 368, 386, 425, 371, 404)	2.61	100.00
	(280, 355, 492, 358, 372, 368, 386, 425, 371, 404)	2.61	100.00
	(281, 355, 492, 491, 372, 368, 386, 425, 371, 404)	2.61	100.00
	(281, 355, 492, 358, 372, 368, 386, 425, 371, 404)	2.61	100.00
	(281, 280, 492, 491, 372, 368, 386, 425, 371, 404)	2.61	100.00
4	(280, 281)	100.00	100.00
	(281, 355)	98.04	98.04
	(280, 355)	98.04	98.04
	(280, 281, 355)	98.04	98.04
	(280, 281, 355, 358)	90.20	92.00
	(280, 281, 358)	90.20	90.20
	(280, 355, 358)	90.20	92.00
	(281, 355, 358)	90.20	92.00
	(280, 358)	90.20	90.20
	(281, 358)	90.20	90.20
	(355, 358)	90.20	90.20
	(280, 281, 355, 358, 491, 492, 272, 275, 276)	3.92	100.00
	(280, 281, 355, 491, 492, 460, 461, 462, 402)	3.92	100.00
	(280, 281, 355, 491, 492, 460, 462, 402)	3.92	100.00
	(280, 281, 355, 491, 492, 460, 461, 402)	3.92	100.00
	(280, 281, 355, 491, 492, 460, 461, 462)	3.92	100.00

Table 4.15: SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Open sequences with high support or confidence values within four clusters.

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(28, 109)	97.39	99.12	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109)	85.65	87.17	2.67	63.64	0.87	100.00	0.00	0.00
(27, 250)	84.78	86.28	0.76	18.18	0.87	100.00	0.00	0.00
(109, 250)	84.78	85.53	0.76	11.76	0.87	100.00	0.00	0.00
(28, 250)	83.91	85.40	1.15	16.67	0.87	50.00	0.00	0.00
(28, 27, 250, 113, 123, 138, 181, 126, 161, 176)	1.30	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 27, 250, 113, 127, 138, 181, 126, 161, 176)	1.30	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 27, 109, 250, 234, 227, 237, 244, 230, 239)	1.30	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 109, 250, 113, 127, 123, 138, 181, 126, 161)	1.30	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109, 250, 113, 127, 123, 138, 181, 126, 161)	1.30	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(305, 317)	0.00	0.00	6.49	89.47	0.00	0.00	0.00	0.00
(286, 317)	0.00	0.00	5.73	83.33	0.00	0.00	0.00	0.00
(305, 286)	0.00	0.00	4.58	63.16	0.00	0.00	0.00	0.00
(372, 368)	0.00	0.00	3.82	76.92	9.13	75.00	1.96	50.00
(281, 355)	0.00	0.00	3.82	47.62	97.39	98.68	98.04	98.04
(281, 280, 372, 386)	0.00	0.00	1.15	100.00	9.57	81.48	0.00	0.00
(286, 305, 317)	0.00	0.00	1.91	100.00	0.00	0.00	0.00	0.00
(21, 22, 17, 19)	0.43	13.00	1.53	100.00	0.00	0.00	0.00	0.00
(21, 18, 17, 19)	0.43	11.00	1.53	100.00	0.00	0.00	0.00	0.00
(18, 17, 19)	0.43	8.00	1.53	100.00	0.00	0.00	0.00	0.00
(22, 17, 19)	0.43	5.00	1.53	100.00	0.00	0.00	0.00	0.00
(21, 17, 19)	0.43	2.00	1.53	100.00	0.00	0.00	0.00	0.00
(196, 186, 194)	0.00	0.00	1.53	100.00	0.00	0.00	0.00	0.00
(281, 355)	0.00	0.00	3.82	47.62	97.39	98.68	98.04	98.04
(281, 280)	0.00	0.00	3.05	38.10	97.39	98.68	0.00	0.00
(280, 355)	0.00	0.00	1.53	25.00	85.22	85.96	98.04	98.04
(281, 280, 355)	0.00	0.00	0.76	25.00	83.91	86.16	0.00	0.00
(281, 492)	0.00	0.00	1.15	14.29	83.04	84.14	74.51	74.51
(280, 492)	0.00	0.00	1.53	25.00	83.04	83.77	74.51	74.51
(280, 355, 492, 491, 372, 368, 386, 425, 371, 404)	0.00	0.00	0.00	0.00	2.61	100.00	0.00	0.00
(280, 355, 492, 358, 372, 368, 386, 425, 371, 404)	0.00	0.00	0.00	0.00	2.61	100.00	0.00	0.00
(281, 355, 492, 491, 372, 368, 386, 425, 371, 404)	0.00	0.00	0.00	0.00	2.61	100.00	0.00	0.00
(281, 355, 492, 358, 372, 368, 386, 425, 371, 404)	0.00	0.00	0.00	0.00	2.61	100.00	0.00	0.00

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(281, 280, 492, 491, 372, 368, 386, 425, 371, 404)	0.00	0.00	0.00	0.00	2.61	100.00	0.00	0.00
(280, 281)	0.00	0.00	1.15	18.75	1.74	1.75	100.00	100.00
(281, 355)	0.00	0.00	3.82	47.62	97.39	98.68	98.04	98.04
(280, 355)	0.00	0.00	1.53	25.00	85.22	85.96	98.04	98.04
(280, 281, 355)	0.00	0.00	0.76	66.67	1.30	75.00	98.04	98.04
(280, 281, 355, 358)	0.00	0.00	0.00	0.00	0.87	66.67	90.20	92.00
(280, 281, 358)	0.00	0.00	0.00	0.00	0.87	50.00	90.20	90.20
(280, 355, 358)	0.00	0.00	0.00	0.00	59.13	69.39	90.20	92.00
(281, 355, 358)	0.00	0.00	0.00	0.00	68.26	70.09	90.20	92.00
(280, 358)	0.00	0.00	0.76	12.50	69.57	70.18	90.20	90.20
(281, 358)	0.00	0.00	0.00	0.00	69.57	70.48	90.20	90.20
(355, 358)	0.00	0.00	0.00	0.00	68.70	69.91	90.20	90.20
(280, 281, 355, 358, 491, 492, 272, 275, 276)	0.00	0.00	0.00	0.00	0.00	0.00	3.92	100.00
(280, 281, 355, 491, 492, 460, 461, 462, 402)	0.00	0.00	0.00	0.00	0.00	0.00	3.92	100.00
(280, 281, 355, 491, 492, 460, 462, 402)	0.00	0.00	0.00	0.00	0.00	0.00	3.92	100.00
(280, 281, 355, 491, 492, 460, 461, 402)	0.00	0.00	0.00	0.00	0.00	0.00	3.92	100.00
(280, 281, 355, 491, 492, 460, 461, 462)	0.00	0.00	0.00	0.00	0.00	0.00	3.92	100.00

Table 4.16: SAM applied to dataset 3 (Belgian telecom provider), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.8.3 Examining clusters on the length of server sessions

A third way of cluster examination is based on the length of server sessions. In order to analyse how sensitive SAM is to the length of the server sessions, figures 4.23 to 4.25 provide, for each data set and SAM application, the distribution of the length of server sessions within each cluster. On the horizontal axis, the length of the server session is given. On the vertical axis, the relative frequency is presented. *Relative frequency* of server sessions' length x ($x = 1, 2, \dots, 55$) in cluster C_a is equal to the number of server sessions of x elements (pages) long in cluster C_a divided by the total number of server sessions in cluster C_a , multiplied by 100.

4.8.3.1 Data set 1 (<http://www.luc.ac.be>)

In figure 4.23, cluster one and five group server sessions, which are mainly of or between one and thirteen elements long. In cluster two, three and four, 41.69%, 73.59% and 73.61% of the server sessions are one element long. In cluster six, most of the server sessions are between 20 and 35 elements long. Compared with the distribution of server sessions' length in the first data set, presented in figure 4.4, cluster two shows approximately the same distribution.

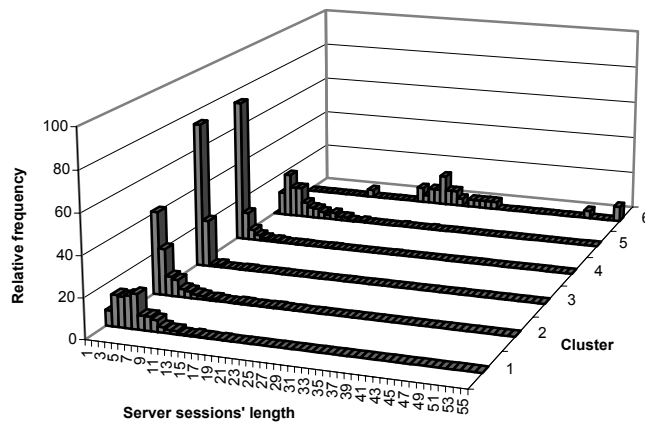


Figure 4.23: SAM applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Distribution of server sessions' length in six clusters.

4.8.3.2 Data set 2 (<http://machines.hyperreal.org>)

In figure 3.24, cluster one, two and four show similar distributions for server sessions of or between one to three elements long, with highest relative frequencies for one-page sessions of respectively 53.87%, 83.9% and 59.14%. In cluster three, 94.75% of the server sessions are of or between one and ten elements long. Cluster five represents server sessions, which are of or between one and twenty pages long, showing a maximum relative frequency of 13.68% for server sessions of seven elements long. Compared with the distribution of server sessions' length in the second data set, presented in figure 4.4, cluster one shows the most similar distribution.

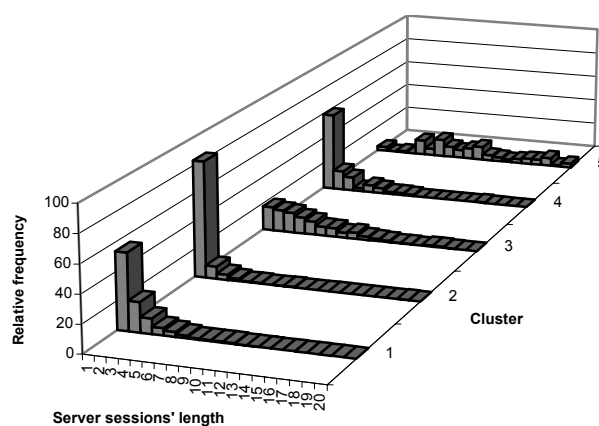


Figure 4.24: SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of server sessions' length in five clusters.

4.8.3.3 Data set 3 (Belgian telecom provider)

In figure 4.25, cluster one, three and four provide a distribution of server sessions' length with the same characteristics as the one given in figure 4.4 for dataset 3. Most of the server sessions are five or six elements long. Finally, in cluster two, another distribution is provided, showing maximum relative frequencies for one-page sessions (29.39%), two-page sessions (19.85%), five-page sessions (15.27%) and three-page sessions (14.12%).

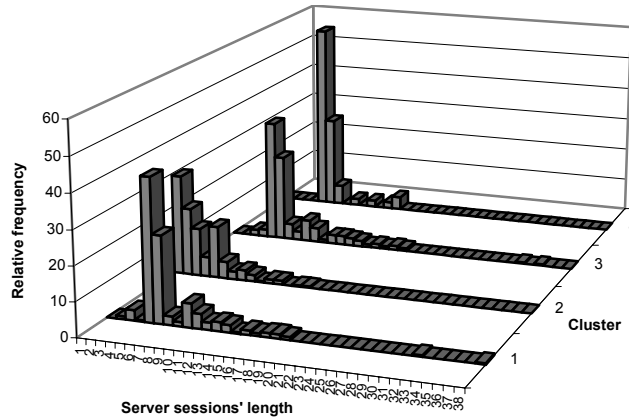


Figure 4.25: SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of server sessions' length in four clusters.

4.9 Results

In order to provide a general overview how people visit a web site, clustering server sessions based on SAM distance extracts *groups of surfing behaviour*, also called *visiting profiles*, providing not only visited pages, but also the order in which pages are visited. For each data set, graphical presentations of the visiting profiles are given.

In figures 4.26 to 4.28, visiting profiles are represented by means of clusters showing the web site structure with regard to the `page_ids` of open sequences with high support values given in tables 4.11, 4.13 and 4.15. For each page, the `page_id` is given along with (a part of) the URL address of this particular page, which is written under the `page_id` inside the rectangular. Regarding the privacy agreements of the web site of a Belgian telecom provider, a description of the pages in figure 4.28 is given instead of the complete URL addresses. Yet, in figures 4.26 and 4.27, the complete URL address of each page can be read taking into account the level in the web site structure along with the links. For example, in figure 4.26, page 1 is the main page with URL address <http://www.luc.ac.be/tew>. Going one level downwards, three web pages appear. The complete URL address of page 2 is http://www.luc.ac.be/tew/diensten/diensten_voor_students.htm. Another example of how to read a URL address is http://www.luc.ac.be/tew/opleidingen/basisopleiding/opbouw_tew.htm for page

65. In figure 4.27, page 1026 can be reached through the URL address <http://machines.hyperreal.org/manufacturers/Roland/TR-909>.

In figures 4.26 to 4.28, links between pages are drawn by thin black solid arrows, while navigations, including order-based information, are given by the bigger solid arrows. For example, in figure 4.26, page 68 can be accessed through page 1 and vice versa. A visitor can also go from page 43 to page 55. However, the other way around is not possible here. *Open sequences* having the *five highest support values*, illustrated in tables 4.11, 4.13 and 4.15, are used to represent navigation patterns within each cluster. Support (s) and confidence (c) values are written next to or above the navigation arrow. For example, in cluster two of the first data set, in 20% of the server sessions page 68 is visited before page 43. The confidence value indicates that, if people visit page 68, the probability that they will proceed to page 43 afterwards is 24%. Nevertheless, we could also use open sequences selected on high confidence values, presented in tables 4.11, 4.13 and 4.15. Yet, the open sequences selected on high support values present order-based information for most of the server sessions, are short of length and therefore more efficient to provide a clear, realistic and graphical view of the results.

For evaluation purposes, distribution of server sessions is given in the upper left corner of every cluster in figures 4.26 to 4.28. For example, 23.51% of the server sessions analysed in the first data set are grouped in cluster two. Practically, this means that 650 out of the 2764 server sessions are grouped in cluster two.

In order to avoid complex drawings of arrows making the figures unclear, some modifications are made in figures 4.26 to 4.28. First, regarding the links between pages, arrows pointing to a `page_id` may appear. For example, in figure 4.26, from pages 43, 49, 65 and 71 one may proceed to pages 1, 2 and 9. Likewise, from pages 26, 28, 30, 33, 35, 38, 40, 55, 47, 59, 60 and 63 one may proceed to pages 1, 2, 9 and 68. The dashed parts of the links indicate that there is no intersection with other links. If there were no dashed parts, the links could be misinterpreted, saying, for example, that from page 71 a link goes to page 63. Second, with regard to navigations presented by open sequences, lines showing arrows in the middle of navigations, instead of at the beginning or at the end, may appear. For example, in cluster two of figure 4.27, when navigating from page 657 to page 802 and from page 802 to page 657, somewhere in the middle of both navigations, an arrow is drawn. These arrows are used for interpreting open sequences having more than two elements. Support and confidence values are given next to or above the arrow of the last navigation within the open sequence. In cluster two of figure 4.27, in 2% of the server sessions navigations appear in the following order: 657, 802, 657. Furthermore, if people visit page 802 after page 657, the probability is 59%

that they will return to page 657. Also, in 2% of the server sessions navigations appear in the following order: 802, 657, 802. Furthermore, if people visit page 657 after page 802, the probability is 38% that they will return to page 802.

Finally, an overview of how people visit the web sites <http://www.luc.ac.be/tew>, <http://machines.hyperreal.org> and a Belgian telecom provider are given below. For the three data sets of our analyses, the largest clusters are graphically presented below. Graphical presentations of the remaining clusters are given in appendix 4.

4.9.1 Surfing behaviour at <http://www.luc.ac.be/tew>

In figure 4.26, cluster two and four are given, representing respectively 23.51% and 36.47% of the server sessions in dataset 1. Note that web pages at the lowest level of the web site structure show only the page_id, without a URL address, due to space limitations. Besides, all of these pages represent curriculum pages for each educational degree, showing courses and teaching subjects with regard to each year of study and specialisation. For example, page 26 is the curriculum page for the second year education in economic engineering, specialisation accountancy and finance. Page 28 is the curriculum page for the second year education in economic engineering, specialisation international business. And page 30 is the curriculum page for the second year education in economic engineering, specialisation small and medium enterprises.

Cluster two represents visiting profiles from the main education page (68/opleidingen) followed by underlying pages in the structure, which are education pages with regard to a particular degree, i.e. education in economic engineering (43 /basisopl./opbouw_hi), education in economic engineering computer sciences (49 /basisopl./opbouw_hibin) and education in applied economic sciences (65 /basisopl./opbouw_tew). Navigation patterns also show visits to the main education page followed by an information page about new education and examination systems (71 /V_onderwijs_en_examensysteem). 5% of the server sessions in cluster two requested education in economic engineering computer sciences, followed by the web page of the second year of their curriculum (47).

In cluster four, visiting profiles mostly consist of only one page, which is indicated in figure 4.26 by a frame around web pages. For example, 5.58% of the server sessions in cluster four are one-page visits to the web page giving information about study evenings, where presentations of political or economical subjects take place (11 /studie-avonden). Other examples of one-page visits are the curriculum web pages at the lowest level of the web site structure: first year education in applied economic sciences (55), second year

education in economic engineering computer sciences (47), education in applied economic sciences, specialisation accountancy and finance (59), specialisation service management (60) and specialisation marketing (63).

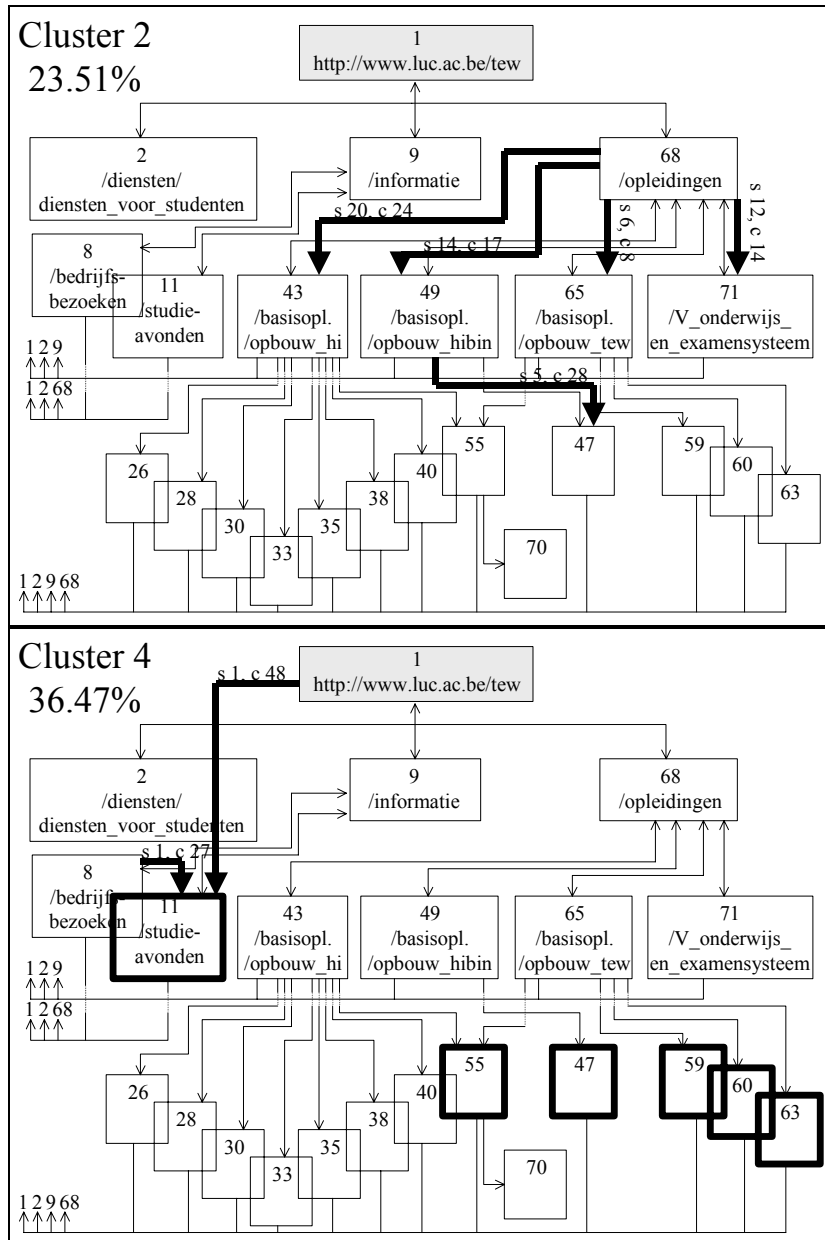


Figure 4.26: SAM applied to data set 1: Surfing behaviour on <http://www.luc.ac.be/tew>: navigation patterns, providing page and order-based information.

4.9.2 Surfing behaviour at <http://machines.hyperreal.org>

In figure 4.27, cluster one and two are the two largest clusters representing respectively 60.50% and 18.65% of the server sessions in the second data set. Note that, in figure 4.27, the web pages /manufacturers, /categories, /drum-machines, /Korg and /software do not show a page_id because they do not appear within the open sequences selected by high support values, given in table 4.13. As such, they are only presented for structural reasons. Furthermore, frames drawn with dashed rectangles represent web pages that originated from a different logged URL address in the files. Further analysis revealed that the log files also stored information of people who used the URL address <http://www.hyperreal.org> and navigations from this main page on. For example, page 933, having URL address <http://www.hyperreal.org/manufacturers/Korg>, appears to be exactly the same as <http://machines.hyperreal.org/manufacturers/Korg>.

In cluster one and cluster two, visiting profiles, shown by the navigations drawn by the big arrows, do not provide high support values because 53.87% and 83.90% of the server sessions in cluster one and two are one-page sessions. With regard to one-page sessions in cluster one, the highest number of direct accesses are found for the web page with the alternative URL address <http://www.hyperreal.org> (163), for <http://www.hyperreal.org/manufacturers/Moog> (947) and for the home page <http://machines.hyperreal.org> (657). This is indicated in the figure by a frame around these pages. With regard to one-page sessions in cluster two, instead of several pages, only the home page <http://machines.hyperreal.org> (657) is considered.

Although the support values for navigations in cluster one and two are low, some of the confidence values are relatively high. This means that information for page prediction may be provided about the probability that people will visit particular pages after or before other pages. For example, the probability that <http://machines.hyperreal.org/ecards> (190) is visited after http://machines.hyperreal.org/the_Roland_TB-303 (804) is 82% (cluster one). Interesting to know is that a direct link from page 804 to 190 is not present in the web site structure, which means that people have follow a workaround procedure to fulfil this pattern. Also, the probability that people will visit <http://machines.hyperreal.org/schematics> (1153) after <http://machines.hyperreal.org/manufacturers/Moog/schematics> (338) is 81% (cluster one). The probability that people will proceed to the home page <http://machines.hyperreal.org> (657) after having visited <http://machines.hyperreal.org/manufacturers/categories/DR-660> (802) is 76%

(cluster two). Finally, if people visit page 657, followed by page 802, the probability that page 657 is re-visited becomes 59% (cluster two).

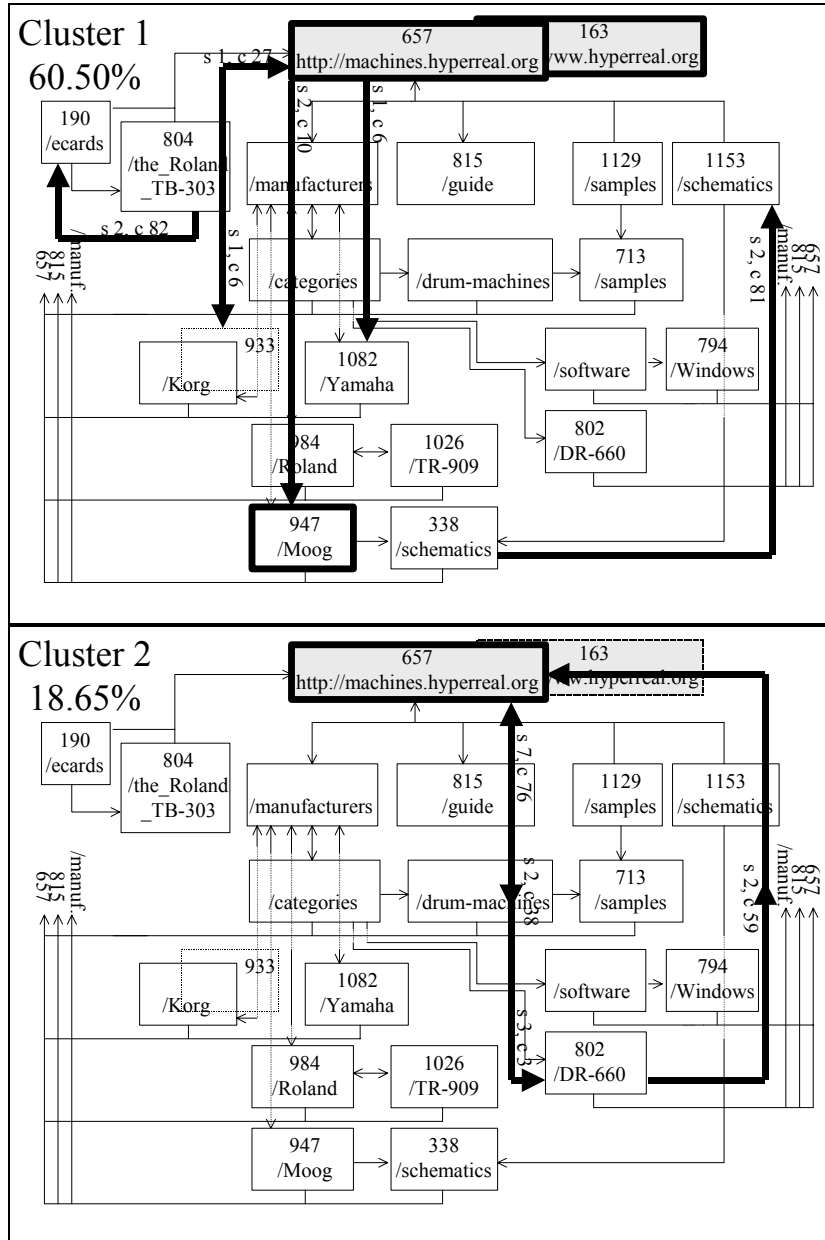


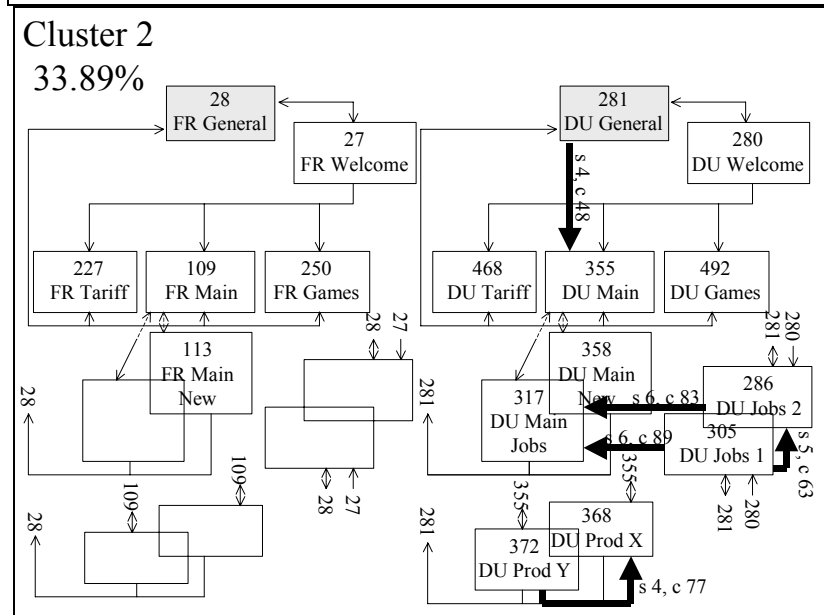
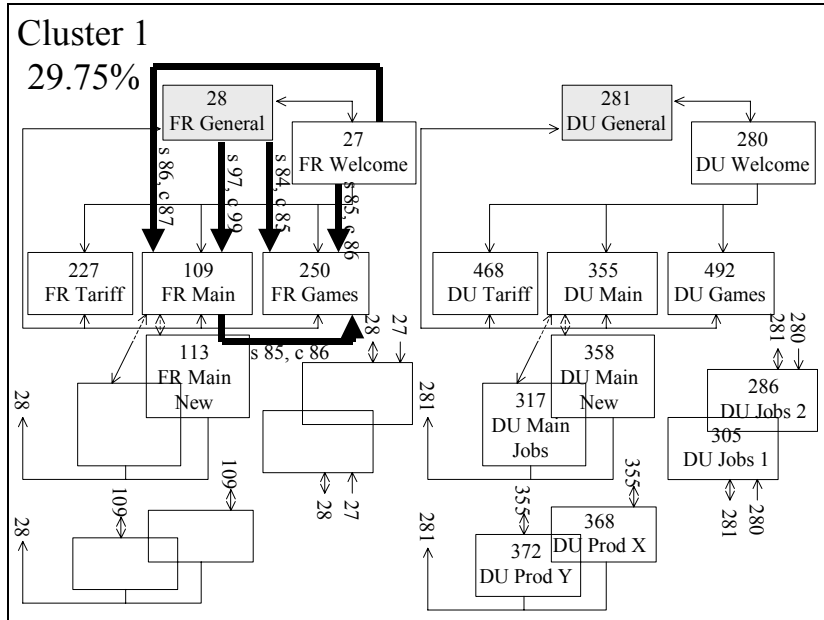
Figure 4.27: SAM applied to data set 2: Surfing behaviour at <http://machines.hyperreal.org>: navigation patterns, providing page and order-based information.

4.9.3 *Surfing behaviour at the web site of a Belgian telecom provider*

In figure 4.28, cluster one, two and three representing respectively 29.75%, 33.89% and 29.75% of the server sessions in data set 3, are graphically presented. On the web site of a Belgian telecom provider, every web page is given in two languages and for each language a different URL address is stored in the log files. Therefore, in figure 4.28, the structure of the web site is build for two languages: French, indicated by FR and Dutch, indicated by DU in the description of each page. In every cluster, web pages in French language and their structure are printed left; web pages in Dutch language and their structure are printed right. Furthermore, rectangles that are drawn without page_id and description of the web page do not appear in table 4.15, where clusters are described by means of open sequences selected by high support values. In order to show that the structure of French pages is equal to Dutch pages, the empty rectangles are given for structural reasons only.

Generally, visiting profiles are categorized in two groups. On the one hand, cluster one represents visiting profiles towards pages in French language. On the other, clusters two and three (and four, re. appendix 4) represent visiting profiles towards pages in Dutch language. The main difference between French and Dutch profiles is that web pages with regard to jobs, or employment in general at the company, as well as web pages with regard to products 'X' and 'Y' are not visited by French speaking people.

Particularly, cluster one represents visiting profiles from the French general (28) and welcome (27) pages, which are considered as two home pages in French language. Also, 85% of the server sessions in cluster one visit the main page before the games page. Furthermore, if people have visited the main page in French language, the probability is 86% that they will visit thereafter the games page in French language. Cluster two represents visiting profiles in Dutch language with regard to pages offering jobs at the company. Also, pages of products X and Y are visited. If people have visited the first page about jobs at the company, the probability is 63% that they will visit thereafter the second page. If people have visited the web page giving information about product Y, the probability is 77% that they will visit thereafter the web page giving information about product X. Cluster three is, despite the difference in languages, comparable with cluster one. Visiting profiles from the Dutch general (281) and welcome (280) pages, which are considered as two home pages in Dutch language, to the Dutch main page (355) and the Dutch games page (492), are represented. However, the French speaking people usually visit the games page after the main page whereas the Dutch speaking people visit the games page after one of the home pages.



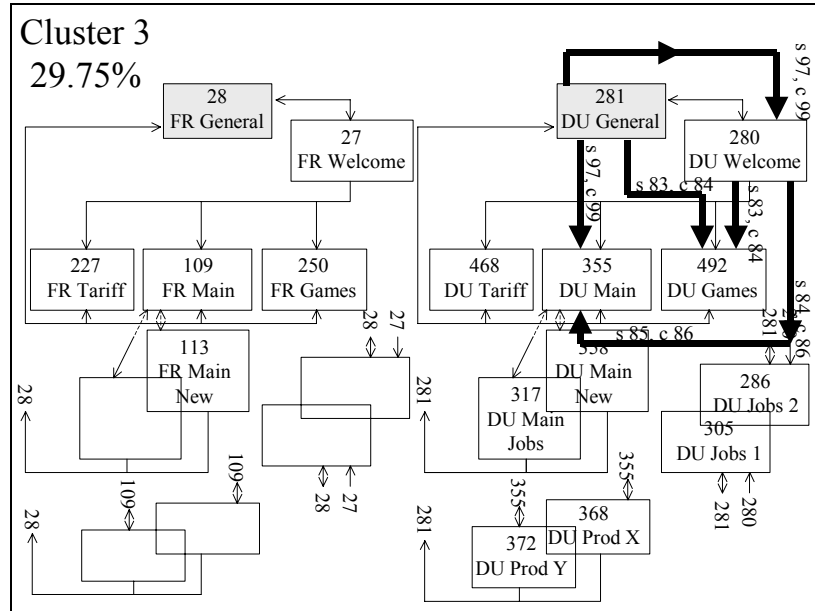


Figure 4.28: SAM applied to data set 3: Surfing behaviour at the web site of a Belgian telecom provider: navigation patterns, providing page and order-based information.

4.10 Deploying the results

Visiting profiles, graphically presented by clusters in the previous section, provide information, which may be incorporated into two optimisation tasks within web development (re. 4.3 Objectives). First, *the structure of the web site* may be adjusted conform to visiting profiles. Practically, this means that, for the convenience of the visitor, information of visiting profiles is used by link optimisation studies. Through deleting and inserting direct hyperlinks between web pages, the web site may be transformed for the benefit of the largest group of visitors. Second, proxy servers or web personalization techniques may incorporate general visiting profiles in order to *predict page requests*. As such, better and faster services to web visitors are provided. For each data set, we give some suggestions for re-structuring the web site by means of direct hyperlinks between pages. We also give some examples of page prediction. Finally, for each data set, some suggestions are given for other topics like *cross-link examination, learning curve* and *psychology studies*.

4.10.1 *Suggestions for structure and service improvement of the web site* <http://www.luc.ac.be/tew>

In cluster two, low confidence values of navigations between pages with direct hyperlinks indicate that the links are not optimally used. We therefore suggest moving pages 43, 49, 65 and 71 to a higher level in the web site structure. For example, the main education page (68) may be replaced by pages 43, 49, 65 and 71 since the direct hyperlinks from the main education page (68) to the education pages of specific degrees in economic sciences are only used in 24%, 17%, 8% and 14% of the cases. We may also suggest deleting the following direct hyperlinks:

- From main education (68) to education in economic engineering (43)
- From main education (68) to education in economic engineering computer sciences (49)
- From main education (68) to education in applied economic sciences (65)
- From main education (68) to new education and examination system (71)

In cluster four, profiles of one-page sessions are identified, particularly for the following pages: information about study evenings (11), curriculum first year applied economic sciences and economic engineering (55), curriculum second year economic engineering computer sciences (47), curriculum applied economic sciences specialization accountancy and finance (59), service

management (60) and marketing (63). This means that pages 11, 55, 47, 59, 60 and 63 are directly accessed without using the direct hyperlinks. We therefore suggest moving these pages one or two levels higher in the web site structure. For example, information about study evenings (11) may replace the general information page (9) since the direct hyperlink from page 9 to page 11 is not optimally used. Moreover, 48% of the cases that visited the home page (1) visited thereafter page 11, indicating that a direct hyperlink from (1) to (11) may be used more efficiently. Also, the curriculum pages may be re-directed to the second level in the web site structure, just below the home page. We may also suggest deleting the following direct hyperlinks:

- From general information (9) to information about study evenings (11)
- From education in applied economic sciences (65) to curriculum applied economic sciences specialization accountancy and finance (59)
- From education in applied economic sciences (65) to curriculum applied economic sciences specialization service management (60)
- From education in applied economic sciences (65) to curriculum applied economic sciences specialization marketing (63)
- From education in economic engineering computer sciences (49) to curriculum second year economic engineering computer sciences (47)
- From education in economic engineering (43) to curriculum first year applied economic sciences and economic engineering (55)

4.10.2 Suggestions for structure and service improvement of the web site <http://machines.hyperreal.org>

Cluster one identifies mostly one-page sessions to the home pages with URL addresses <http://machines.hyperreal.org> (657), <http://www.hyperreal.org> (163) and to other pages like for example the manufacturers label Moog page, which is reached at <http://machines.hyperreal.org/manufacturers/Moog> (947). Since page 947 is a popular one-page visit, we suggest moving this page to a higher level in the web site structure. For example, instead of structuring page 947 at the third level in the hierarchy under the manufacturers page, page 947 may be re-directed to the second level, directly under the home page.

High confidence values between pages without direct hyperlinks suggest inserting direct hyperlinks. For example, in cluster one, the probability that <http://machines.hyperreal.org/ecards> (190) is visited after http://machines.hyperreal.org/the_Roland_TB-303 (804) is 82%. Also, the probability that <http://machines.hyperreal.org/schematics> (1153) is visited after <http://machines.hyperreal.org/manufacturers/Moog/schematics> (338) is 81%. We therefore suggest inserting the following direct hyperlinks:

- From http://machines.hyperreal.org/the_Roland_TB-303 (804) to <http://machines.hyperreal.org/ecards> (190)
- From <http://machines.hyperreal.org/manufacturers/Moog/schematics> (338) to <http://machines.hyperreal.org/schematics> (1153)

Although the support values are not very high, high confidence values may provide information for page prediction. For example, in cluster one and two:

- If people visit http://machines.hyperreal.org/the_Roland_TB-303 (804), the probability is 82% that they will visit thereafter <http://machines.hyperreal.org/ecards> (190)
- If people visit <http://machines.hyperreal.org/manufacturers/Moog/schematics> (338), the probability is 81% that they will visit thereafter <http://machines.hyperreal.org/schematics> (1153)
- If people visit <http://machines.hyperreal.org/manufacturers/categories/DR-660> (802), the probability is 76% that they will proceed thereafter to the home page <http://machines.hyperreal.org> (657)

Finally, relatively low confidence values between pages without direct hyperlinks indicate that the site is being used conform to the structure. For example, in cluster one, if page 657 is visited, the probability is only 6% that page 1082 is visited since no direct hyperlink exist from page 657 to 1082. Likewise, relatively high confidence values between pages with direct hyperlinks indicate that the site is being used conform to the structure. For example, in cluster two, if page 802 is visited, the probability is 76% that page 657 is visited after page 802 since a direct hyperlink exists from page 802 to 657.

4.10.3 *Suggestions for structure and service improvement of the web site of a Belgian telecom provider*

In cluster one, visiting profiles indicate that people visiting the French web pages are using the web site conform to the intentions of the web developer i.e. navigating from the general home (28) and welcome (27) page to pages like the main page (109) and games page (250), following the direct hyperlinks. The high confidence values may be used for page predictions with regard to pages presented in French language as follows:

- If people visit the general home page (28), the probability is 99% that they will visit thereafter the main page (109)
- If people visit the general home page (28), the probability is 85% that they will visit thereafter the games page (250)
- If people visit the welcome page (27), the probability is 87% % that they will visit thereafter the main page (109)
- If people visit the welcome page (27), the probability is 86% that they will visit thereafter the games page (250)
- If people visit the main page (109), the probability is 86% that they will visit thereafter the games page (250)

In cluster two, high confidence values between pages presented in Dutch language without direct hyperlinks suggest inserting direct hyperlinks:

- From Jobs 2 (286) to Main Jobs (317)
- From Jobs 1 (305) to Main Jobs (317)
- From Jobs 1 (305) to Jobs 2 (286)
- From Prod Y (372) to Prod X (368)

Generally, the results of analysing visiting behaviour on the web site of a Belgian telecom provider suggest different structures between pages in French and pages in Dutch language, in order to allow for the visitor to directly follow common patterns, without having to click on other pages that are not typically visited. The main difference between visiting behaviour of French and Dutch speaking people is that web pages with regard to employment or jobs at the company as well as web pages presenting information about products X and Y are not visited by French speaking people. Therefore, we may suggest moving less important pages, such as ‘FR Main Jobs’, ‘FR Jobs 1’, ‘FR Jobs 2’, ‘FR Prod X’, ‘FR Prod Y’ to a lower level in the structure of the web site.

4.10.4 Suggestions for other topics

Besides link optimisation and page prediction studies, the information provided by SAM-based clustering may be used in other studies as well. *Cross-links* are references to other documents based on common features. For example, in the first data set of our university web site, two groups of visitors may be distinguished: staff and students. For each group different cross-links are applied in order to provide different guided tours for staff members and for students. On the one hand, a guided tour for staff members may create a path through the curriculum web pages (cluster 4) in order to visit and change pages of interest. On the other, guided tours for students to curriculum web pages

(cluster 4) do not provide cross-links to documents for changing information. Moreover, order-based associations provide the ability to deliver common paths to the visitor. For example, in the second data set, cross-links may be inserted between pages in order to present the following common path (re. cluster 4; support = 1.08%, confidence = 100%):

- <http://machines.hyperreal.org/samples> (1129) followed by
- <http://machines.hyperreal.org/manufacturers/Moog> (947) followed by
- <http://machines.hyperreal.org/samples> (1129) followed by
- <http://machines.hyperreal.org/manufacturers/Yamaha/TX-81z> (1103) followed by
- <http://machines.hyperreal.org/samples> (1129)

Finally, an example of suggestions for cross-links in the third data set may be as follows. If visitors use the Dutch pages, two different guided tours may be offered. The first provides cross-links between jobs and products pages; the second between main, welcome and games pages. If visitors use the French pages, one guided tour providing cross-links between main, welcome and games pages covers most common paths.

Results of SAM-based clustering may also be deployed in *learning curves* and *psychology studies*. Through frequent visits, visitors become more experienced users of web sites. This means that visitors become more confident with particular web pages and, as such, the function of these web pages may change from content (after first time visit) to navigation (after frequent visits). For example, in the first data set, a first time visitor to <http://www.luc.ac.be/tew> may use the home page as content page and stays more than 12.89 seconds on this page to read the information that is available on the web site. After frequent visits, the user interprets the home page no longer as content but as navigational page in order to directly proceed to other content pages of her/his interest.

4.11 Comparing SAM with Association

In this section, SAM is compared with Association distance (re. section 4.1 and equation (4.1)) in order to examine how well SAM measures order-based information within server sessions. In order to proof the surplus value of SAM, described in section 4.1, we also calculate Association distance measures between server sessions of each data set in table 4.4. The same procedure, given in figure 4.3, is used, which means that hierarchical clustering is invoked on the distance matrices holding the pair wise Association distance measures between server sessions. Criteria for defining the number of clusters for each data set are given in figures 4.29 to 4.31. For the first and second data set, three clusters are chosen. For the third data set, four clusters are chosen. Then, like in section 4.8, clusters are analysed on `page_ids`, the order of `page_ids` and on the length of server sessions. For each data set, the results are given below.

The following remarks are given with regard to figures 4.29 and 4.30. The PSF in figures 4.29 and 4.30 augments with rising number of clusters and starts to decrease at solutions of a large number of clusters. In the first and second data set, if Association distance is applied to server sessions, PSF statistic keeps on rising with the number of clusters and finally reaches a peak at respectively 422 and 519 clusters. Compared with PSF in figure 4.31 and in figures 4.14, 4.15 and 4.16 of the SAM application, PSF reached a relatively high level at or near the chosen cluster solution and starts to decrease relatively early after 6, 12, 11 or 8 clusters. The reason why PSF keeps on rising with the number of clusters in figures 4.29 and 4.30 is due to incremental differences between dispersion of server sessions in the data set and the sum of dispersions of server sessions within clusters (re. equation PSF in table 4.9). For each additional cluster, the sum of dispersions of server sessions within clusters decreases, indicating that the clustering solution improves. Unfortunately, if we follow the information provided by PSF in figures 4.29 and 4.30 for defining the number of clusters we will end up with 422 and 519 clusters, which obviously will not provide visiting profiles of large groups of visitors. The PSF statistic in figures 4.29 and 4.30 might indicate that Association distance is not an appropriate method for measuring distances between server sessions.

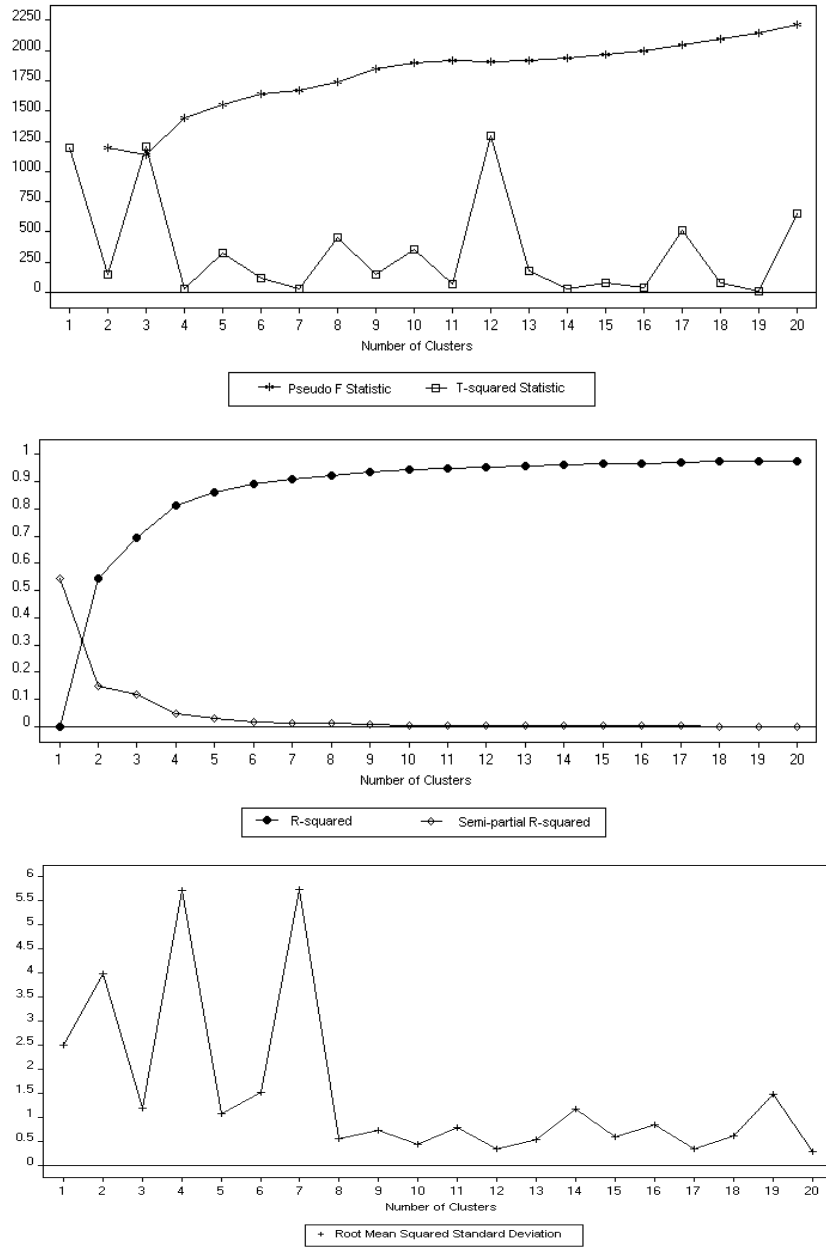


Figure 4.29: Applying Association distance to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

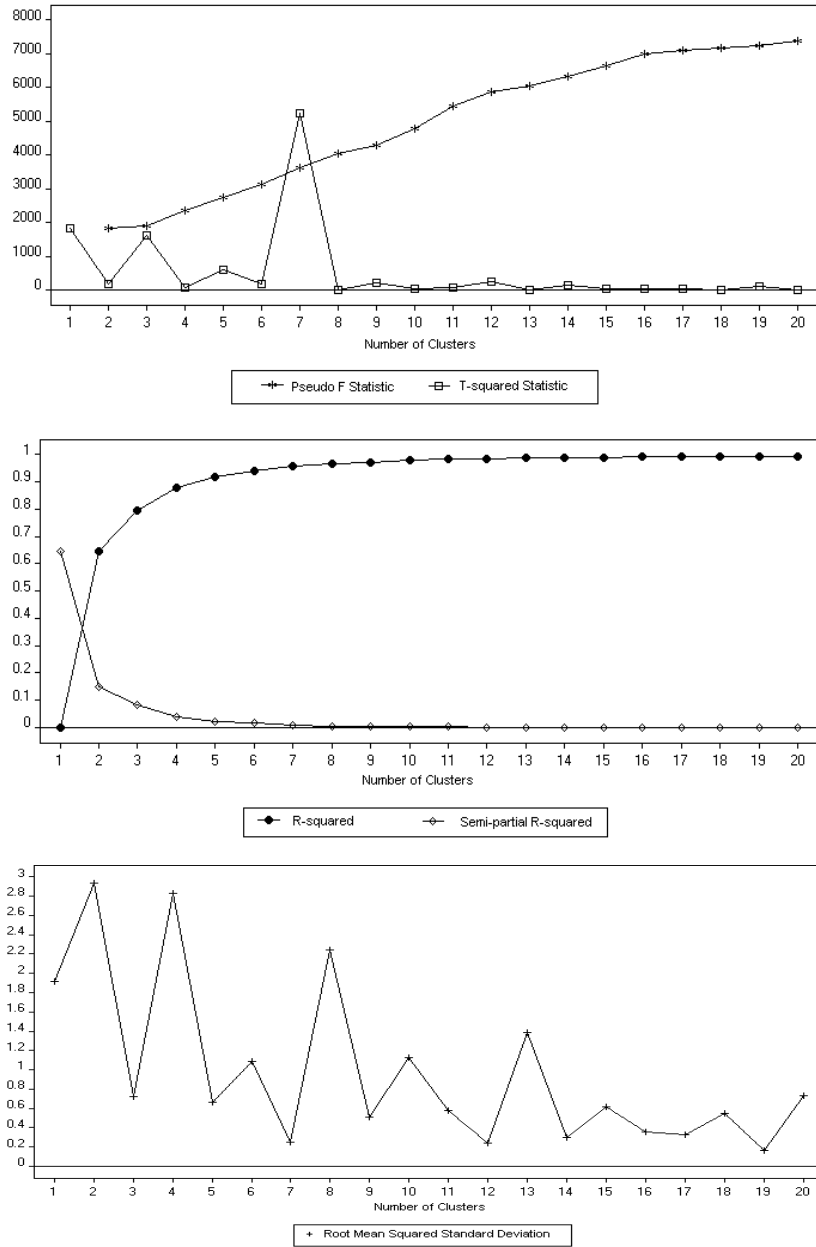


Figure 4.30: Applying Association distance to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

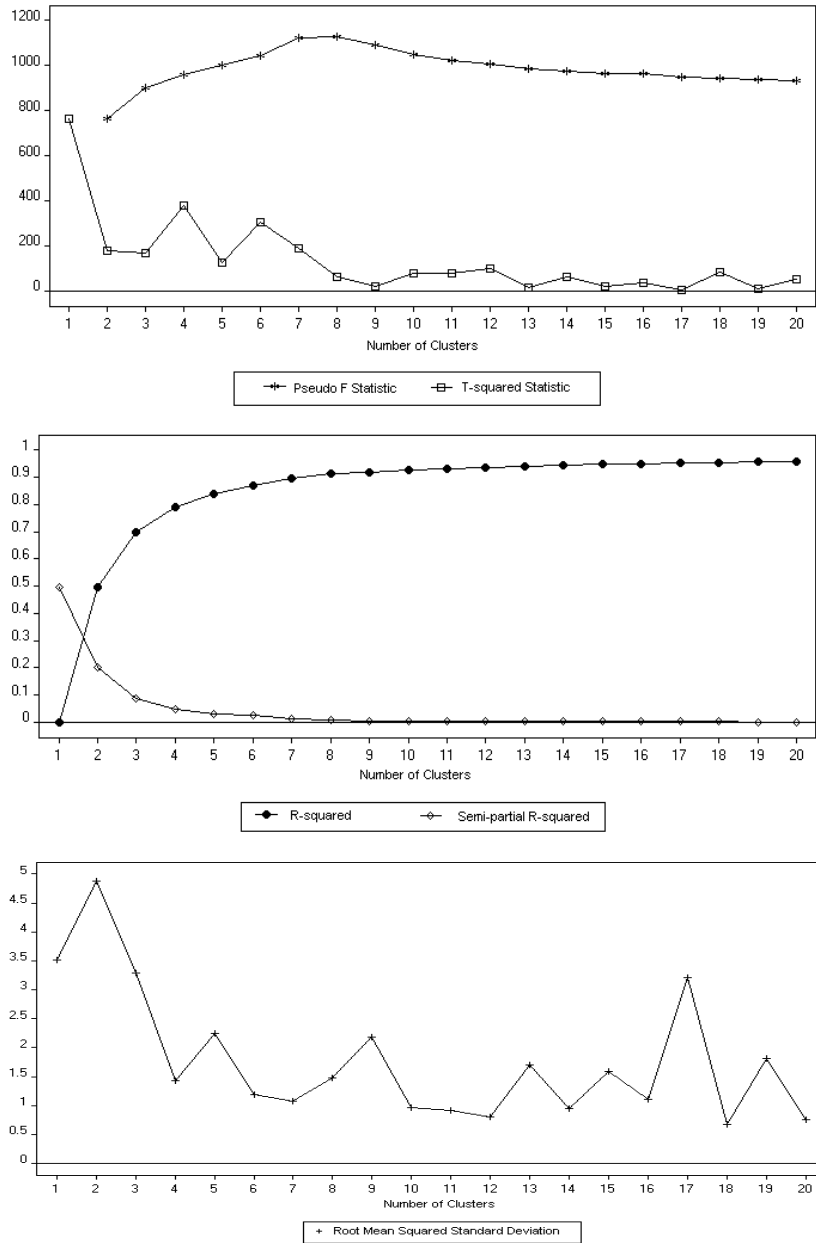


Figure 4.31: Applying Association distance to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Information criteria for defining the number of clusters.

4.11.1 Examining clusters on page_ids

For comparison reasons, the same scales are used as in the previous sections, when clusters resulting from the SAM application are examined. Definitions of relative frequencies and exclusivities are given in section 4.8.1. Tables of relative frequencies and exclusivities with regard to Association distance applied to each data set presented in table 4.4 are given in appendix 4.

4.11.1.1 Data set 1 (<http://www.luc.ac.be/tew>)

In figure 4.32, the distributions of page_ids in different clusters are very much alike. Compared with figure 4.17 in section 4.8.1, where SAM is applied to the same data set, each cluster presents a different distribution of page_ids. Exclusivities in figure 4.33 are not better/worse than exclusivities in figure 4.18. The difference between exclusivities of clusters of page_ids based on SAM and Association is that clusters resulting from the SAM application show a considerable number of zero exclusivity values, which may indicate that clusters are better separated.

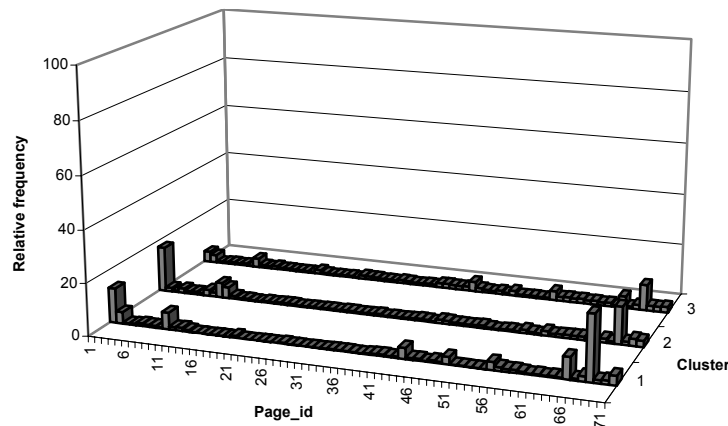


Figure 4.32: Association distance applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Distribution of web pages in three clusters.

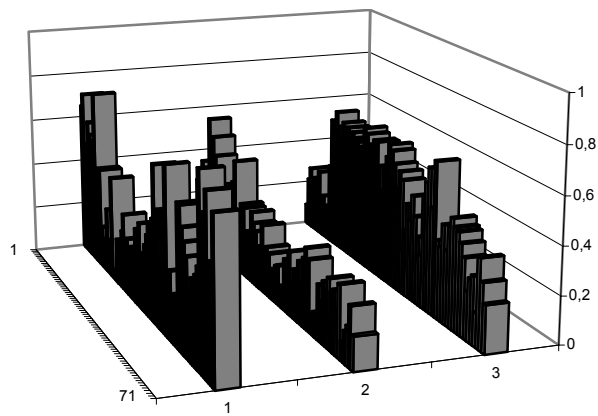
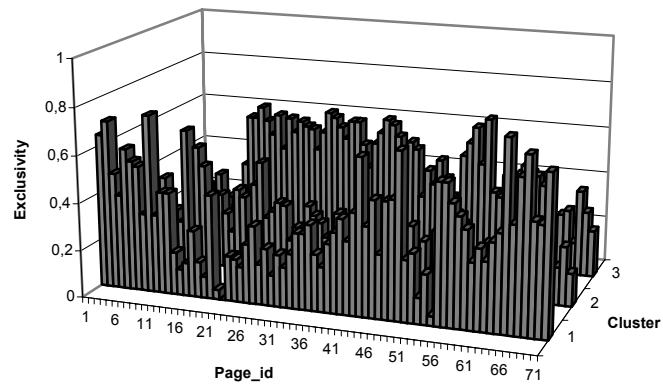


Figure 4.33: Association distance applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Exclusivity of web pages in three clusters.

4.11.1.2 Data set 2 (<http://machines.hyperreal.org>)

Distributions of groups of page_ids in different clusters look very much alike in figure 4.34. Compared with figure 4.19 in section 4.8.1, where SAM is applied to data set 2, each cluster presents a different distribution of groups of page_ids. Proceeding to figure 4.35, exclusivities of groups of page_ids within clusters are not worse or better compared with figure 4.20.

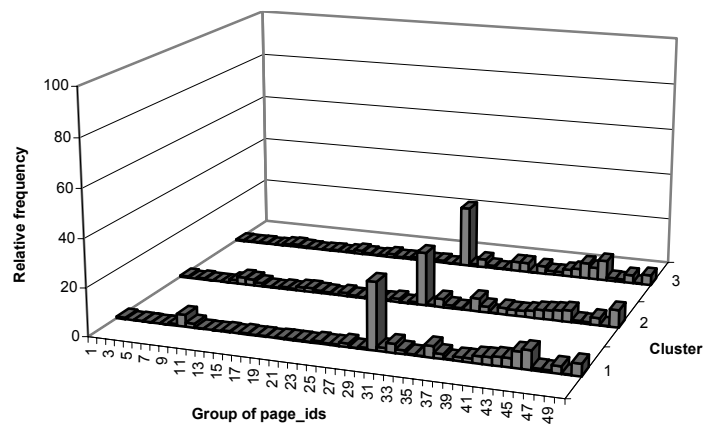


Figure 4.34: Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of groups of page_ids in three clusters.

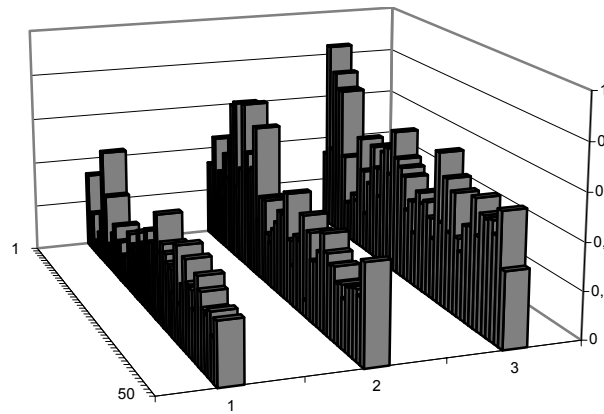
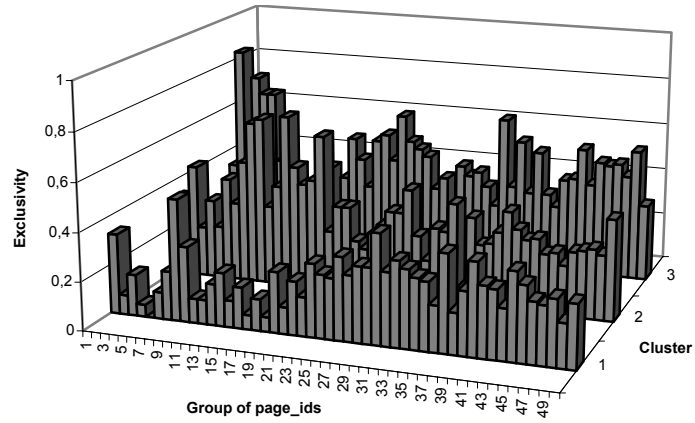


Figure 4.35: Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Exclusivity of groups of page_ids in three clusters.

4.11.1.3 Data set 3 (Belgian telecom provider)

Comparing figure 4.36 with figure 4.21 in section 4.8.1, where SAM is applied to data set 3, the distribution of page_ids in cluster 4 of figure 4.36 looks very much like cluster three and four of figure 4.21. With regard to exclusivities of page_ids represented by clusters, more zero values are found in clusters of figure 4.22, which may indicate that clusters are better separated based on SAM distance. Generally, relatively high exclusivities are found in three clusters in figure 4.22 whereas in figure 4.37 only two of the four clusters provide high exclusivities.

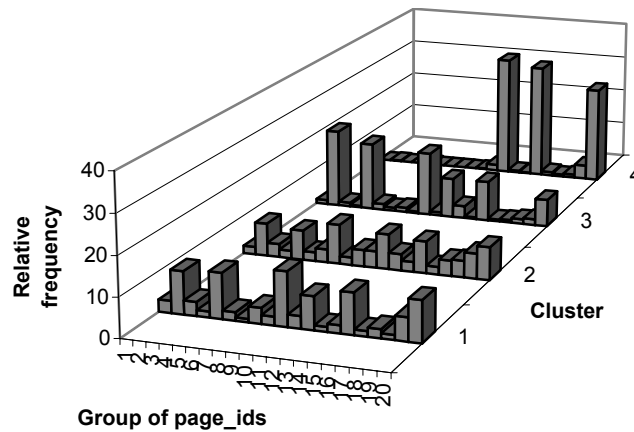


Figure 4.36: Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of groups of page_ids in four clusters.

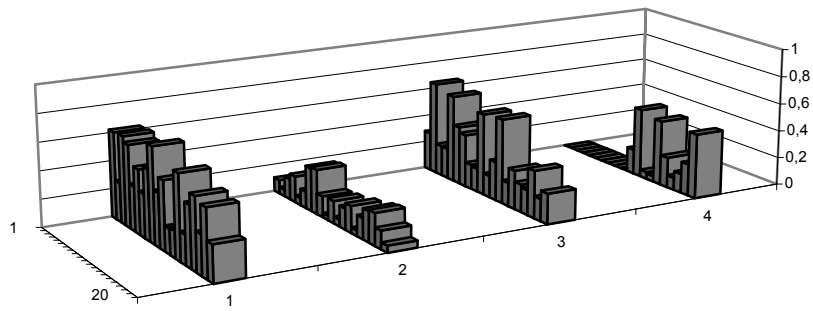
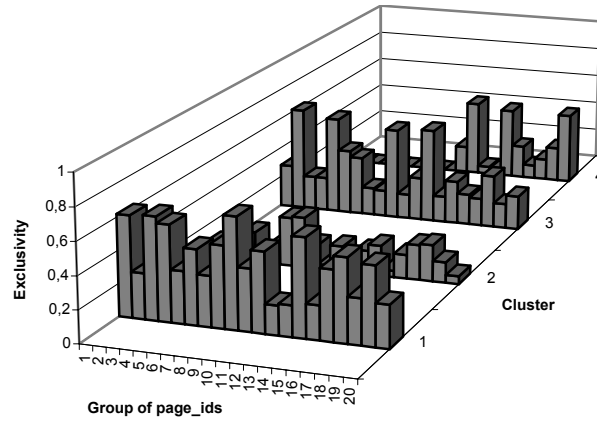


Figure 4.37: Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Exclusivity of groups of page_ids in four clusters.

4.11.2 Examining clusters on the order of page_ids

4.11.2.1 Data set 1 (<http://www.luc.ac.be/tew>)

Table 4.17 presents order-based information of visited pages within three clusters, based on Association distance measures, for data set 1. Cluster two consists of one-page sessions. Clusters one and three are very much alike. Four out of five open sequences selected on high support values are the same in cluster one and three. Note that, with regard to open sequences selected on high confidence values, more than thirty combinations were found of open sequences with four pages long and support of 5.84% and confidence of 100%. This is indicated in the last row of cluster three with ‘...’. Looking at table 4.18, we notice that the main difference between cluster one and three is indicated by the open sequences selected on high confidence values for cluster three, which are not found in cluster one. All of the remaining open sequences are found in cluster one and three.

Comparing table 4.18 with table 4.12, Association distance is much more sensitive to one-page sessions than SAM distance. 95.05% of the server sessions in cluster two, based on Association distance, are one-page sessions, while the other two clusters do not contain one-page sessions. When SAM distance is used between server sessions, four out of six clusters hold one-page sessions. Also, taking into account that it is very obvious to have zero support and confidence values for open sequences in clusters of one-page sessions, table 4.18 contains far less zero values for support values at the non-diagonal places (or support values not written in bold) than table 4.12. This may indicate that order-based information of visited pages is better represented in clusters based on SAM than on Association distance.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(68, 65)	27.39	40.33
	(1, 68)	17.28	48.65
	(1, 9)	16.08	45.29
	(68, 43)	12.50	18.41
	(68, 9)	12.10	17.82
	(8, 11)	1.11	73.68
	(68, 49, 47, 48)	1.35	60.71
	(68, 47, 48)	1.35	60.71
	(1, 68, 2, 9)	2.63	56.90
(49, 47, 48)	1.43	54.55	
2	One-page sessions		
3	(68, 65)	46.10	62.83
	(68, 9)	34.42	46.90
	(68, 43)	33.12	45.13
	(68, 55)	32.47	44.25
	(1, 68)	31.82	81.67
	(42, 14, 25, 17, 51)	5.19	100.00
	(42, 14, 25, 17, 53)	5.19	100.00
	(19, 27, 63, 55)	5.84	100.00
	(19, 28, 63, 55)	5.84	100.00
	(19, 59, 37, 55)	5.84	100.00
...	

Table 4.17: Association distance applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Open sequences with high support or confidence values within three clusters.

Open sequences	1		2		3	
	S	C	S	C	S	C
(68, 65)	27.39	40.33	0.00	0.00	46.10	62.83
(1, 68)	17.28	48.65	0.00	0.00	31.82	81.67
(1, 9)	16.08	45.29	0.00	0.00	29.22	75.00
(68, 43)	12.50	18.41	0.00	0.00	33.12	45.13
(68, 9)	12.10	17.82	0.00	0.00	34.42	46.90
(8, 11)	1.11	73.68	0.07	2.94	0.65	20.00
(68, 49, 47, 48)	1.35	60.71	0.00	0.00	1.95	27.27
(68, 47, 48)	1.35	60.71	0.00	0.00	1.95	27.27
(1, 68, 2, 9)	2.63	56.90	0.00	0.00	12.34	65.52
(49, 47, 48)	1.43	54.55	0.00	0.00	2.60	30.77
One-page sessions	0.80	-	95.05	-	0.00	-
(68, 65)	27.39	40.33	0.00	0.00	46.10	62.83
(68, 9)	12.10	17.82	0.00	0.00	34.42	46.90
(68, 43)	12.50	18.41	0.00	0.00	33.12	45.13
(68, 55)	10.35	15.24	0.00	0.00	32.47	44.25
(1, 68)	17.28	48.65	0.00	0.00	31.82	81.67
(42, 14, 25, 17, 51)	0.00	0.00	0.00	0.00	5.19	100.00
(42, 14, 25, 17, 53)	0.00	0.00	0.00	0.00	5.19	100.00
(19, 27, 63, 55)	0.00	0.00	0.00	0.00	5.84	100.00
(19, 28, 63, 55)	0.00	0.00	0.00	0.00	5.84	100.00
(19, 59, 37, 55)	0.00	0.00	0.00	0.00	5.84	100.00

Table 4.18: Association distance applied to dataset 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.11.2.2 Data set 2 (<http://machines.hyperreal.org>)

In table 4.19 and 4.20, cluster one represents mainly one-page sessions. Both cluster two and three group server sessions that are related with pages 657, 984 and 815. Generally, support measures for open sequences are higher in cluster three, indicating that cluster three represents more order-based information related with pages 657, 984 and 815.

Comparing table 4.20 with table 4.14, Association distance is much more sensitive to one-page sessions than SAM distance. 88.33% of the server sessions in cluster one, based on Association distance, are one-page sessions. In cluster two and three, 8.06% and 4.84% of the server sessions are one-page sessions. When SAM distance is used between server sessions, one-page sessions are distributed more equally across clusters. For example, in cluster one and four, 53.87% and 59.14% of the server sessions are one-page sessions. Although, at a general level, we cannot say that table 4.20 contains far less zero values for support values at the non-diagonal places (or support values not written in bold) than table 4.12, we state an important remark about open sequences selected on high confidence values. In table 4.19, open sequence (657, 815, 810, 657) with confidence of 100% is selected for describing order-based information within cluster three. Unfortunately, the same open sequence provides a confidence value of 100% in table 4.20 for cluster two as well. Since support is below 1%, it is not provided in cluster two of table 4.19. Yet, with regard to cluster description by means of open sequences selected on high confidence values, if high confidence is shown in different clusters, server sessions are not well clustered with regard to order-based information of visited pages. No such information is shown when SAM is used as distance measure for clustering. Practically, open sequence (657, 815, 810, 657) from table 4.20 indicates that, if visitors go to page 657, followed by pages 815 and 810, the probability is 100% that they will proceed back to page 657 (cluster two and three). This means that server sessions holding the same sequential relationships are clustered differently. This may indicate that order-based information of visited pages is better represented in clusters based on SAM than on Association distance.

Cluster	Open sequences	Support (%)	Confidence (%)
1	One-page sessions		
2	(802, 657)	4.19	90.48
	(657, 947)	3.31	6.82
	(804, 190)	3.20	96.67
	(657, 984)	3.20	6.59
	(657, 815)	3.09	6.36
	(338, 1153)	2.10	100.00
	(283, 576)	1.66	93.75
3	(337, 1152)	1.32	85.71
	(657, 984)	17.19	24.15
	(815, 657)	14.29	88.06
	(657, 815)	13.56	19.05
	(984, 657)	12.59	59.77
	(657, 815, 657)	12.35	91.07
	(657, 933, 657, 933, 657)	1.21	100.00
	(657, 713, 657, 713, 657)	1.21	100.00
	(657, 947, 657, 1026, 657)	1.21	100.00
	(657, 815, 810, 657)	1.45	100.00
	(657, 984, 993, 657)	1.45	100.00

Table 4.19: Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Open sequences with high support or confidence values within three clusters.

Open sequences	1		2		3	
	S	C	S	C	S	C
One-page sessions	88.33	-	8.06	-	4.84	-
(802, 657)	0.06	11.11	4.19	90.48	0.97	100.00
(657, 947)	0.11	0.36	3.31	6.82	9.20	12.93
(804, 190)	0.06	11.11	3.20	96.67	1.45	66.67
(657, 984)	0.44	1.45	3.20	6.59	17.19	24.15
(657, 815)	0.55	1.81	3.09	6.36	13.56	19.05
(338, 1153)	0.17	37.50	2.10	100.00	1.94	88.89
(283, 576)	0.17	42.86	1.66	93.75	0.48	66.67
(337, 1152)	0.00	0.00	1.32	85.71	0.97	66.67
(657, 984)	0.44	1.45	3.20	6.59	17.19	24.15
(815, 657)	0.28	26.32	1.99	47.37	14.29	88.06
(657, 815)	0.55	1.81	3.09	6.36	13.56	19.05
(984, 657)	0.44	16.67	1.43	26.53	12.59	59.77
(657, 815, 657)	0.28	50.00	1.43	46.43	12.35	91.07
(657, 933, 657, 933, 657)	0.00	0.00	0.00	0.00	1.21	100.00
(657, 713, 657, 713, 657)	0.00	0.00	0.00	0.00	1.21	100.00
(657, 947, 657, 1026, 657)	0.00	0.00	0.00	0.00	1.21	100.00
(657, 815, 810, 657)	0.00	0.00	0.11	100.00	1.45	100.00
(657, 984, 993, 657)	0.00	0.00	0.00	0.00	1.45	100.00

Table 4.20: Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.11.2.3 Data set 3 (Belgian telecom provider)

Table 4.21 presents order-based information of visited pages within four clusters, based on Association distance measures, for data set 3. Cluster four represents server sessions related with pages 281, 280, 355 and 492 while clusters one, two and three mainly represent server sessions related with pages 28, 27, 109 and 250. Note that, with regard to open sequences selected on high confidence values for describing clusters one, two and four, lots of combinations were found with respectively ten, two and eight pages long and support of 1.94%, 1.10% and 1.18%. This is indicated in the last row of clusters one, two and four with ‘...’. Comparing table 4.21 with table 4.15 it becomes very clear that clusters are better separated with regard to sequential information if SAM distance is used between server sessions.

If we compare table 4.21 with table 4.16, the same remark is given about open sequences selected on high confidence values. In table 4.21, several open sequences selected for cluster description on high confidence values show 100% confidence for different clusters. For example, open sequence (196, 186, 194) shows confidence of 100% for clusters two and three. Likewise, open sequence (471, 480) shows confidence of 100% for clusters one, two and four. Other examples are open sequences (365, 369), (52, 64), (461, 462) and (281, 280, 355, 492, 358, 491, 461, 462). Since support is below 1%, they are not provided for cluster description in table 4.21. This means that server sessions are not well clustered with regard to order-based information if Association distance is used between server sessions. Otherwise stated, server sessions holding the same sequential relationships are clustered differently based on Association distance. No such information is shown when SAM is used as distance measure for clustering. Finally, table 4.22 contains far less zero values for support values at the non-diagonal places (or support values not written in bold) than table 4.16. This may also indicate that order-based information of visited pages is better represented in clusters based on SAM than on Association distance.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(28, 109)	54.37	100.00
	(27, 109)	51.46	94.64
	(28, 109, 250)	50.49	92.86
	(109, 250)	50.49	92.86
	(28, 250)	50.49	92.86
	(27, 109, 250, 113, 249, 123, 181, 126, 161, 176)	1.94	100.00
	(27, 109, 250, 113, 249, 123, 181, 126, 161, 176)	1.94	100.00
	(27, 109, 250, 113, 249, 123, 138, 126, 161, 176)	1.94	100.00
	(27, 109, 250, 113, 249, 123, 138, 181, 161, 176)	1.94	100.00
	(27, 109, 250, 113, 249, 123, 138, 181, 126, 176)	1.94	100.00
...	
2	(28, 109)	4.95	64.29
	(28, 27)	4.40	57.14
	(28, 27, 109)	2.75	62.50
	(27, 109)	2.75	62.50
	(471, 480)	2.20	100.00
	(281, 355)	2.20	30.77
	(305, 317, 286)	1.10	100.00
	(196, 186, 194)	1.10	100.00
	(471, 485, 480)	1.10	100.00
	(471, 480)	2.20	100.00
(365, 369)	1.10	100.00	
...	
3	(28, 109)	52.66	94.92
	(27, 109)	46.08	84.48
	(27, 250)	45.77	83.91
	(109, 250)	45.77	83.91
	(28, 250)	45.45	81.92
	(280, 281, 355)	15.05	100.00
	(286, 305, 317)	1.57	100.00
	(52, 64)	1.25	100.00
	(460, 462)	1.25	100.00
	(461, 462)	1.25	100.00
(245, 231)	1.25	100.00	
4	(281, 280)	98.92	99.40
	(281, 355)	97.63	98.21
	(281, 280, 355)	95.27	96.41
	(280, 355)	95.27	95.83
	(281, 492)	84.62	85.12
	(280, 492)	84.62	85.12
	(281, 280, 355, 492, 358, 491, 460, 461, 462)	1.18	100.00
	(281, 280, 355, 492, 358, 491, 461, 462)	2.37	100.00
	(281, 280, 355, 492, 358, 460, 461, 462)	1.78	100.00
	(281, 280, 355, 492, 491, 358, 368, 386)	1.78	100.00
(281, 280, 355, 492, 491, 460, 461, 462)	1.18	100.00	
...	

Table 4.21: Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Open sequences with high support or confidence values within four clusters.

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(28, 109)	54.37	100.00	4.95	64.29	52.66	94.92	0.59	100.00
(27, 109)	51.46	94.64	2.75	62.50	46.08	84.48	0.59	100.00
(28, 109, 250)	50.49	92.86	0.00	0.00	44.51	84.52	0.59	100.00
(109, 250)	50.49	92.86	0.00	0.00	45.77	83.91	0.59	100.00
(28, 250)	50.49	92.86	0.00	0.00	45.45	81.92	0.59	100.00
(27, 109, 250, 113, 249, 138, 181, 126, 161, 176)	1.94	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109, 250, 113, 249, 123, 138, 181, 126, 161, 176)	1.94	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109, 250, 113, 249, 123, 138, 126, 161, 176)	1.94	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109, 250, 113, 249, 123, 138, 181, 161, 176)	1.94	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109, 250, 113, 249, 123, 138, 181, 126, 176)	1.94	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 109)	54.37	100.00	4.95	64.29	52.66	94.92	0.59	100.00
(28, 27)	42.72	78.57	4.40	57.14	43.26	77.97	0.00	0.00
(28, 27, 109)	39.81	93.18	2.75	62.50	34.80	80.43	0.00	0.00
(27, 109)	51.46	94.64	2.75	62.50	46.08	84.48	0.59	100.00
(471, 480)	5.83	100.00	2.20	100.00	0.00	0.00	1.18	100.00
(281, 355)	36.89	95.00	2.20	30.77	24.76	98.75	97.63	98.21
(305, 317, 286)	0.00	0.00	1.10	100.00	0.31	66.67	0.00	0.00
(196, 186, 194)	0.97	50.00	1.10	100.00	0.31	100.00	0.00	0.00
(471, 485, 480)	1.94	66.67	1.10	100.00	0.00	0.00	0.00	0.00
(471, 480)	5.83	100.00	2.20	100.00	0.00	0.00	1.18	100.00
(365, 369)	0.97	100.00	1.10	100.00	0.63	100.00	0.00	0.00
(28, 109)	54.37	100.00	4.95	64.29	52.66	94.92	0.59	100.00
(27, 109)	51.46	94.64	2.75	62.50	46.08	84.48	0.59	100.00
(27, 250)	50.49	92.86	0.00	0.00	45.77	83.91	0.59	100.00
(109, 250)	50.49	92.86	0.00	0.00	45.77	83.91	0.59	100.00
(28, 250)	50.49	92.86	0.00	0.00	45.45	81.92	0.59	100.00
(280, 281, 355)	5.83	85.71	0.55	50.00	15.05	100.00	0.00	0.00
(286, 305, 317)	0.00	0.00	0.00	0.00	1.57	100.00	0.00	0.00
(52, 64)	10.68	100.00	0.00	0.00	1.25	100.00	0.00	0.00
(460, 462)	4.85	83.33	0.00	0.00	1.25	100.00	2.96	62.50
(461, 462)	5.83	100.00	0.55	25.00	1.25	100.00	4.14	77.78
(245, 231)	3.88	66.67	0.00	0.00	1.25	100.00	0.00	0.00
(281, 280)	33.98	87.50	1.10	15.38	9.40	37.50	98.92	99.40

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(281, 355)	36.89	95.00	2.20	30.77	24.76	98.75	97.63	98.21
(281, 280, 355)	27.18	80.00	0.00	0.00	2.19	23.33	95.27	96.41
(280, 355)	32.04	82.50	0.55	14.29	17.55	68.29	95.27	95.83
(281, 492)	35.92	92.50	1.10	15.38	15.99	63.75	84.62	85.12
(280, 492)	35.92	92.50	0.00	0.00	16.93	65.85	84.62	85.12
(281, 280, 355, 492, 358, 491, 460, 461, 462)	0.00	0.00	0.00	0.00	0.00	0.00	1.18	100.00
(281, 280, 355, 492, 358, 491, 461, 462)	0.97	100.00	0.00	0.00	0.00	0.00	2.37	100.00
(281, 280, 355, 492, 358, 460, 461, 462)	0.00	0.00	0.00	0.00	0.00	0.00	1.78	100.00
(281, 280, 355, 492, 491, 358, 368, 386)	4.85	83.33	0.00	0.00	0.00	0.00	1.78	100.00
(281, 280, 355, 492, 491, 460, 461, 462)	0.00	0.00	0.00	0.00	0.00	0.00	1.18	100.00

Table 4.22: Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Evaluating open sequences in other clusters.

4.11.3 Examining clusters on the length of server sessions

4.11.3.1 Data set 1 (<http://www.luc.ac.be/tew>)

If server sessions of data set 1 are clustered based on Association distance, one-page sessions are clustered almost exclusively in cluster two. Moreover, server sessions in cluster one are shorter than server sessions in cluster three. Looking at the distribution of server sessions based on their length when SAM is used as distance measure for clustering (re. figure 4.23), one-page sessions are distributed in four out of six clusters. Besides one-page sessions, clusters three and four merely group server sessions which are two pages long. In general, clusters one, two and five group short as well as longer server sessions. Cluster six, representing only 0.94% of the server sessions in data set 2, holds server sessions, which mostly are longer than 20 elements.

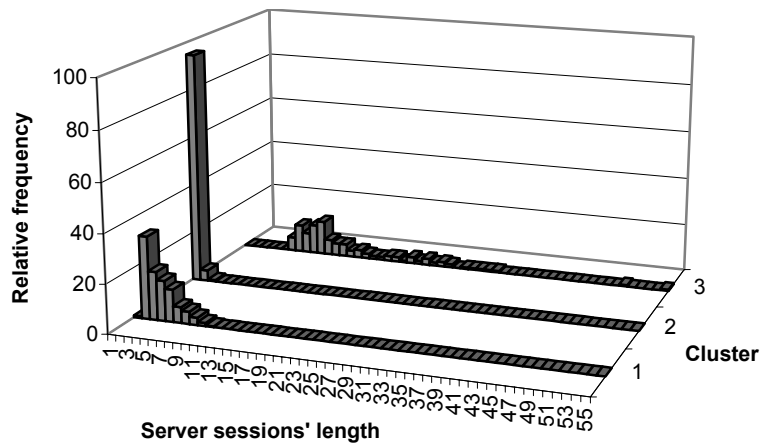


Figure 4.38: Association distance applied to data set 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages: Distribution of server sessions' length in three clusters.

4.11.3.2 Data set 2 (<http://machines.hyperreal.org>)

Comparing figure 4.39 with figure 4.24, one-page sessions are concentrated in one cluster, which is not shown in figure 4.24. Generally, clustering server sessions based on Association distance groups relatively short sessions (i.e. two to four pages long) in cluster two and relatively long sessions (i.e. five pages and more) in cluster three. When server sessions are clustered based on SAM distance measures, clusters are less sensitive to the length of server sessions, given the distribution of server sessions' length in data set 2 (re. figure 4.4).

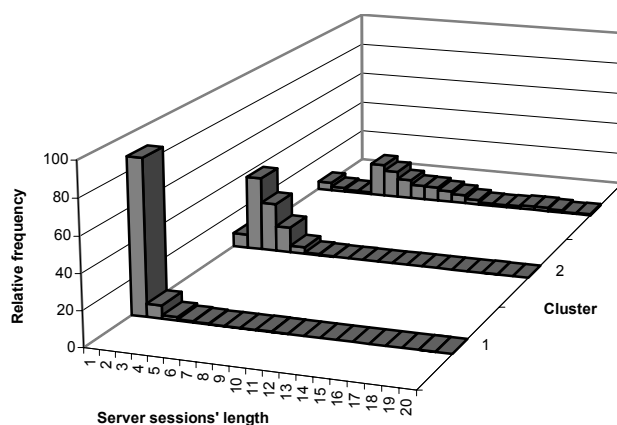


Figure 4.39: Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of server sessions' length in three clusters.

4.11.3.3 Data set 3 (Belgian telecom provider)

Comparing the distribution of the length of server sessions in clusters based on Association distance, presented in figure 4.40, with clusters based on SAM distance, presented in figure 4.25, we remark that, given the distribution in data set 3 (re. figure 4.4), SAM distance measure is less sensitive to the length of server sessions. In figure 4.40, cluster one groups server sessions of nine pages and longer. Cluster two groups server sessions, which are one to four pages long. Cluster three and four respectively group server sessions of minimum five pages and maximum ten pages long. This means that, server sessions of, for example five pages long, are not found in cluster one or two. Yet, in figure 4.25, server sessions of five pages long are found in all of the four clusters.

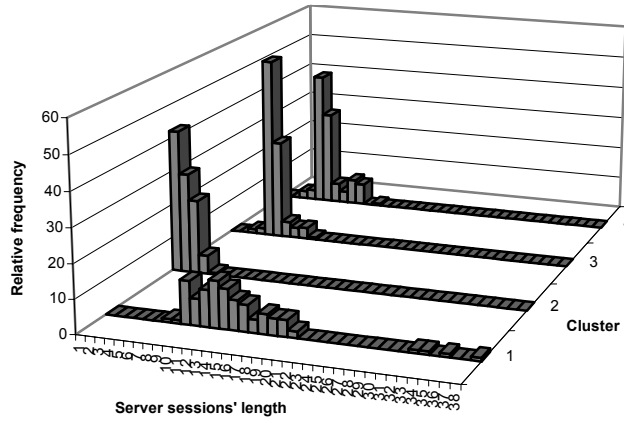


Figure 4.40: Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of server sessions' length in four clusters.

4.12 Comparing SAM with 2-dim SAM

In order to proof the surplus value of 2-dim SAM over SAM, described in section 4.2 of this chapter, we also calculate 2-dim SAM distance measures between server sessions. Instead of using SAM on only one attribute, which is visited pages (re. sections 4.7 to 4.10), in this section 2-dim SAM and SAM are calculated between server sessions consisting of two attributes: *visited pages* and *categories of visiting page time*. The Web Usage Mining process given in figure 4.3 is applied to the data sets that were pre-processed in section 4.5 and described in section 4.6. Examples of how 2-dim SAM and SAM calculate distances between sequences consisting of two attributes, as well as the equations that are used, are given in section 4.2. Criteria for defining the number of clusters for each data set are given in figures 4.41 to 4.46.

If 2-dim SAM is applied to dataset 1, 2 and 3, the number of clusters is respectively three, four and four (re. figures 4.41, 4.42, 4.43). In figure 4.42, the T-squared statistic might indicate that two clusters are a good clustering solution as well. However, we chose for a solution of four clusters because the root mean squared standard deviation at this point is 0.90 compared to 1.37 at two clusters. In figure 4.43, the T-squared statistic might advise a solution of two clusters. However, we chose for a solution of four clusters because R-squared is far too low at two clusters. Only 35.18% of the variance in the data is explained with two clusters, whereas 68.08% of the variance in the data is explained with four clusters.

If SAM is applied to dataset 1, 2 and 3, the number of clusters chosen is respectively three, two and four (re. figures 4.44, 4.45 and 4.46). In figure 4.45, two clusters are chosen because 63.47% of the variance in the data is explained, which is higher than the minimum standard variance explanation of 60% (Hair et al, 1998). In figure 4.46, the T-squared statistic might indicate that three clusters are a good solution as well. However, only 56.18% of the variance in the data is explained, which is below the minimum standard of 60% (Hair et al, 1998), whereas four clusters explain 68.17% of the variance in the data. Besides, the homogeneity of the server sessions in four clusters improves from 4.62 to 3.82.

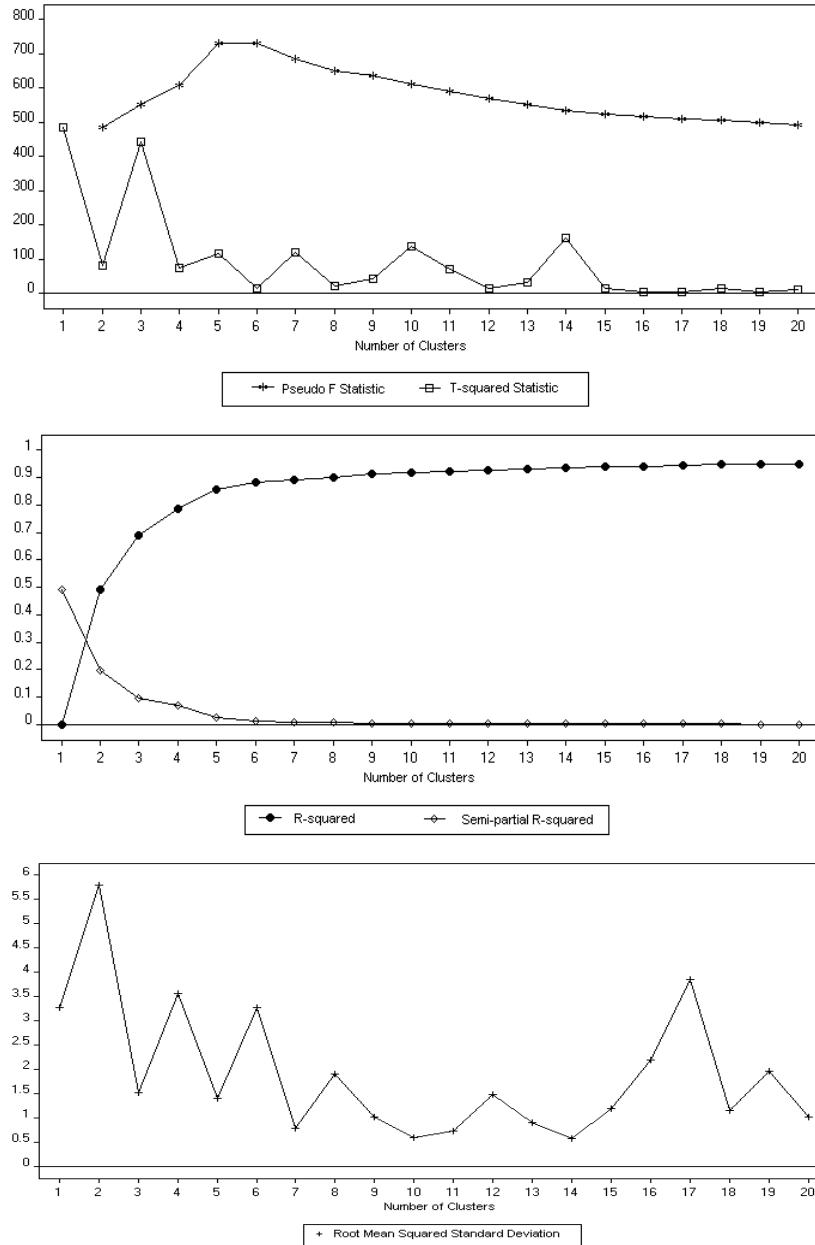


Figure 4.41: Applying 2-dim SAM to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

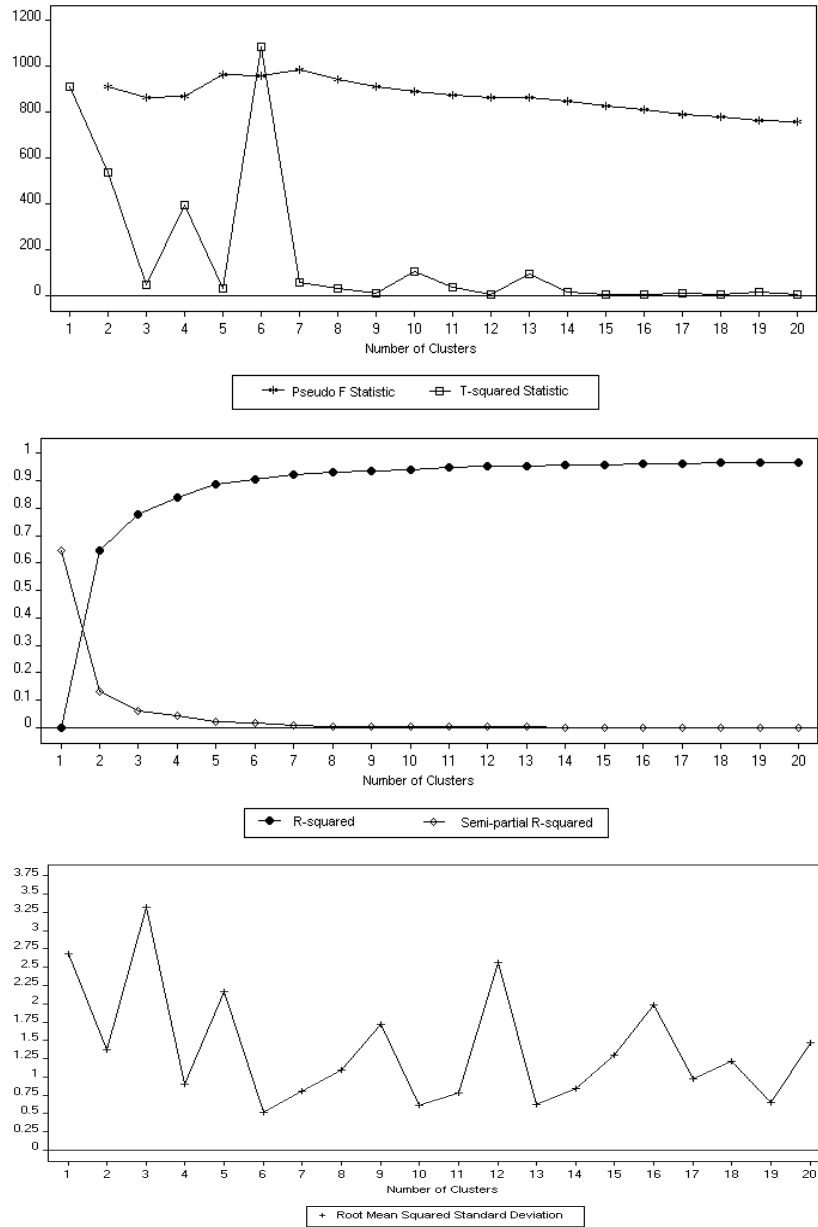


Figure 4.42: Applying 2-dim SAM to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

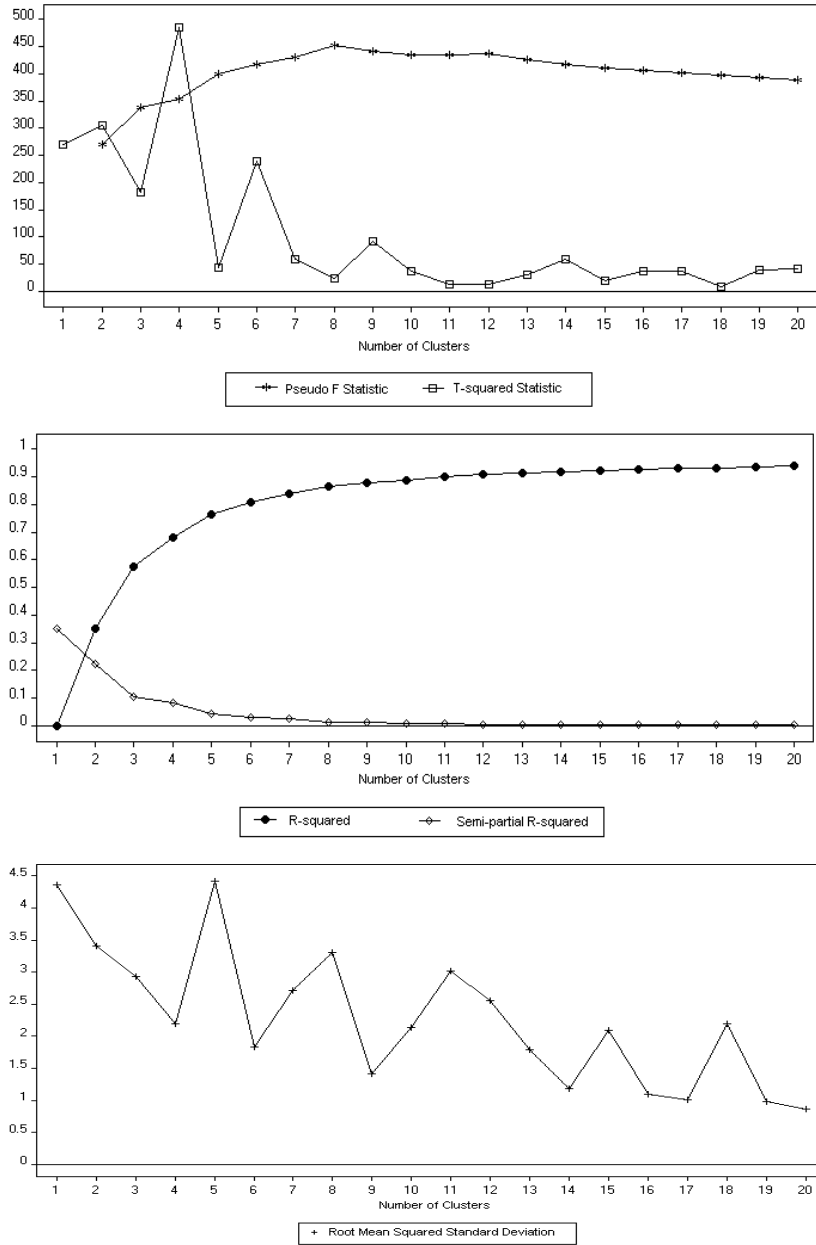


Figure 4.43: Applying 2-dim SAM to data set 3 (Belgian telecom provider), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

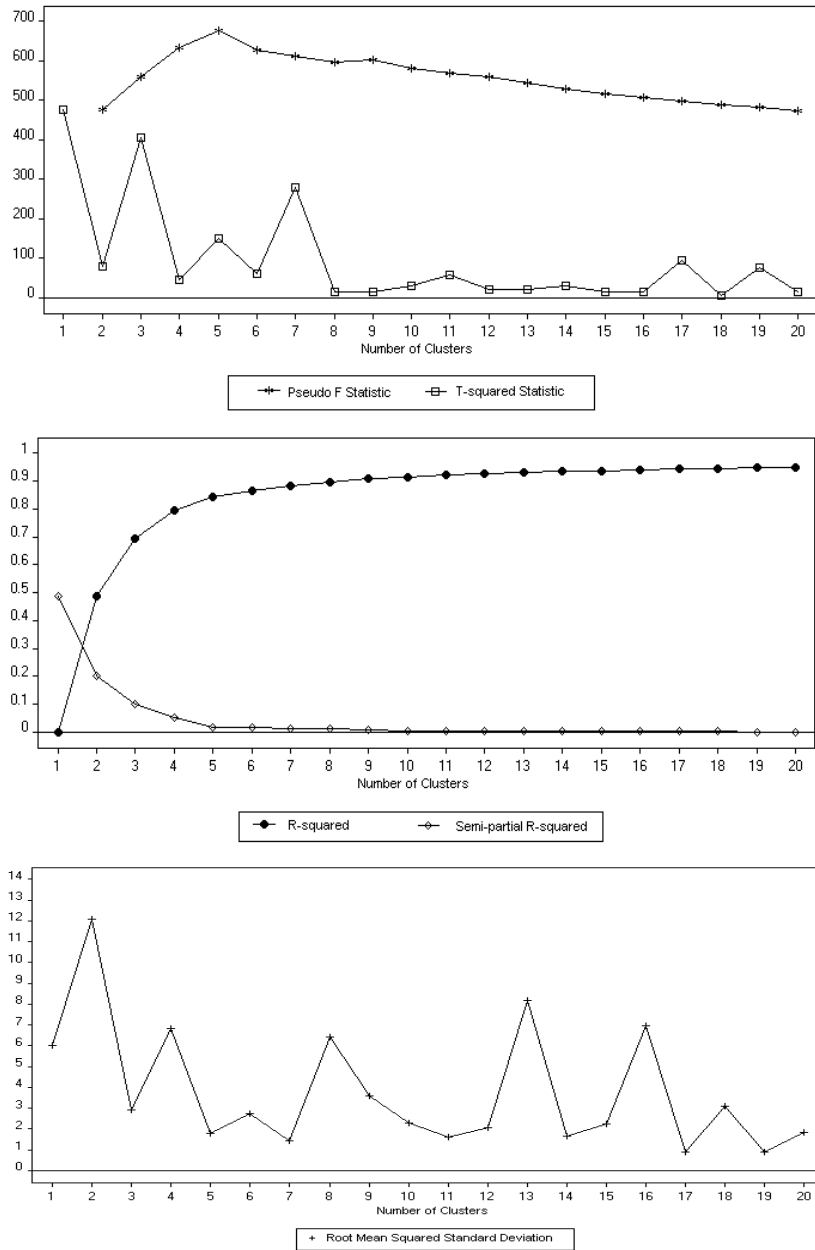


Figure 4.44: Applying SAM to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

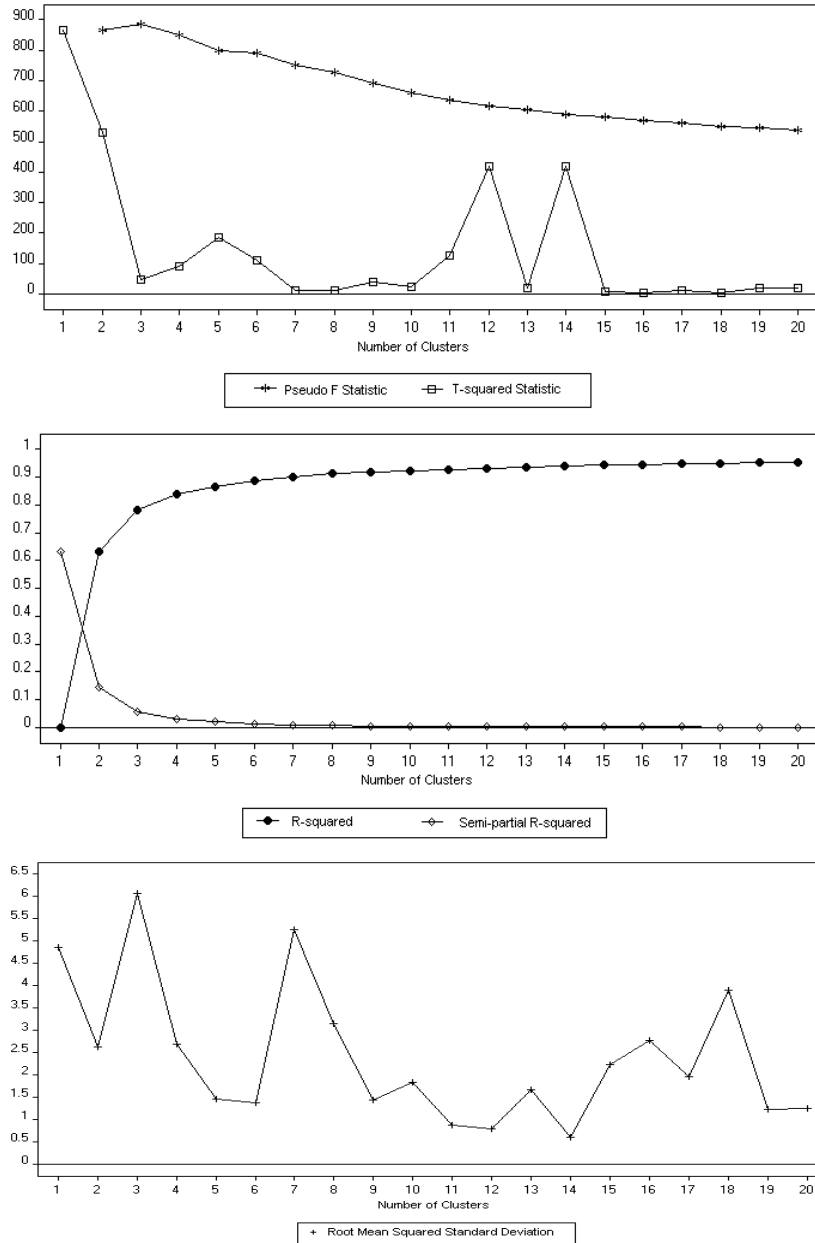


Figure 4.45: Applying SAM to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

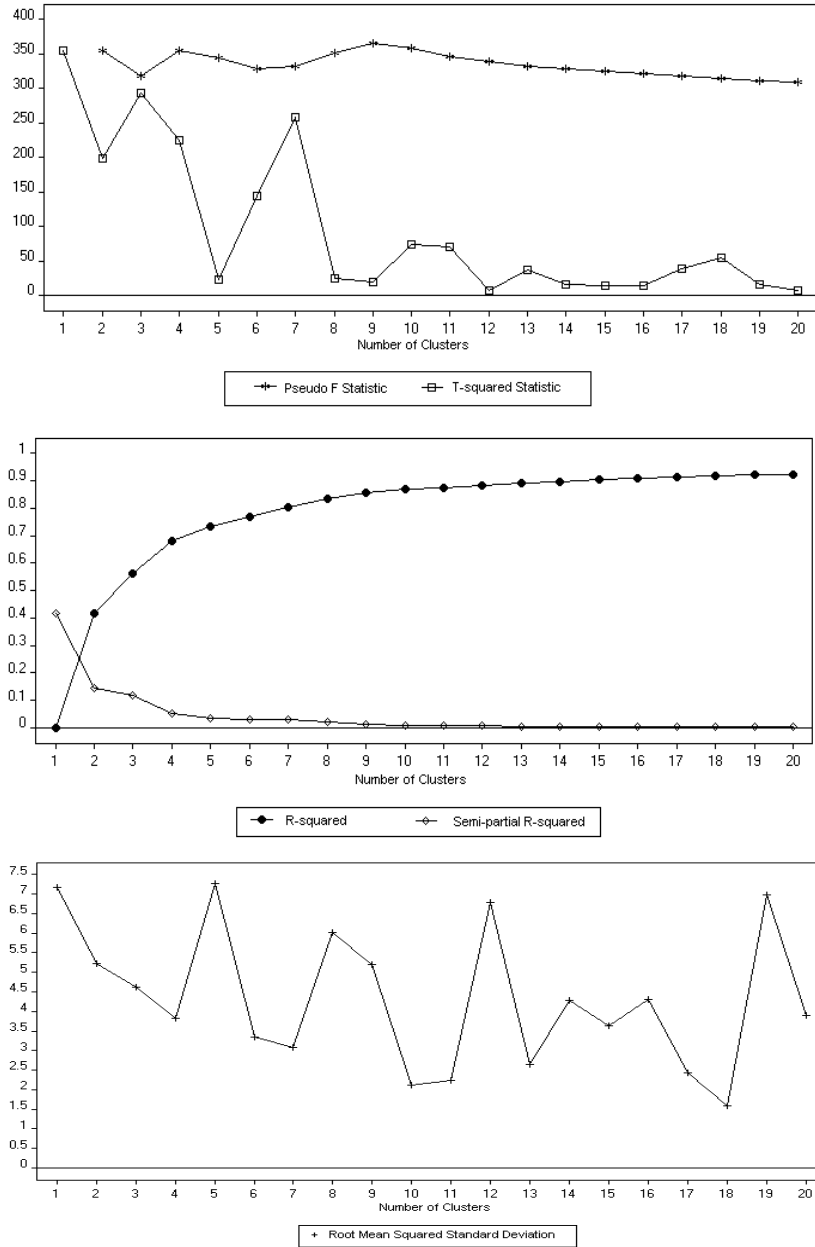


Figure 4.46: Applying SAM to data set 3 (Belgian telecom provider), server sessions consisting of visited pages and categories of visiting page time: Information criteria for defining the number of clusters.

4.12.1 *Examining clusters on order-based information and on relations between page_ids and categories of visiting page time*

Tables 4.23 to 4.28 provide, for each cluster, open sequences consisting of visited pages as well as categories of visiting page time, describing order-based information and relations between pages and times. Open sequences are described in section 4.8 and selected on high support or high confidence values. Definitions and calculations of support and confidence for open sequences are provided in section 4.8 as well. Tables 4.23, 4.25 and 4.27 provide open sequences for the three data sets presented in section 4.6, consisting of visited pages and categories of visiting page time, using 2-dim SAM as distance measure between server sessions. Tables 4.24, 4.26 and 4.28 provide open sequences for the same data sets, using SAM as distance measure between server sessions. The distribution of server sessions across clusters is given at the head of each table. For example, 47% of the server sessions in data set 1 are grouped in cluster 1 (re. table 4.23), based on 2-dim SAM distance.

Comparing table 4.23 (2-dim SAM and data set 1) with table 4.24 (SAM and data set 1), a general difference is that support values for open sequences consisting of two attributes and written in bold are higher in table 4.23. This indicates that two-dimensional server sessions are better clustered with regard to relations between pages and times as well as sequential information if 2-dim SAM is used as distance measure. For example, in table 4.23, 2-dim SAM distinguishes server sessions holding page 1 (with category of visiting page time equal to 0 or 1) followed by page 68 (with category of visiting page time equal to 1) in cluster 1 from server sessions holding page 1 (with category of visiting page time equal to 0) followed by page 68 (with category of visiting page time equal to 0) in cluster 2. Also, in cluster 3, server sessions holding re-visits to page 68 (with category of visiting page time equal to 0 or 1) and related with pages 65 and 71 are distinguished from those related with pages 1, 43 and 69 in cluster 1. This means that 2-dim SAM discovers three different profiles related with page 68. For presentational reasons we will write t_0 , t_1 , t_2 or t_3 between brackets to refer to visiting page time category:

- Page 68 (t_1) is visited after page 1 (t_0 or t_1) and/or before pages 43 and 69 (cluster 1)
- Page 68 (t_0) is visited after page 1 (t_0) (cluster 2)
- Page 68 (t_0) is re-visited (t_1) and/or visited before pages 65 and 71 (cluster 3)

The profiles discovered by 2-dim SAM might reveal important information since page 68 is considered as the most requested page in data set 1 (re. figure 4.5).

In table 4.24, two-dimensional server sessions are clustered differently. SAM is not able to distinguish different profiles with regard to sequential page and related time information for page 68. Instead, each cluster represents unique page combinations. For example, SAM discovers the following visiting profiles:

- Page 1 (t0 or t1) is visited before pages 9 and/or 2. Page 1 (t0 or t3) is also re-visited (t2) (cluster 1).
- Page 68 (t0) is re-visited (t1) and/or visited before pages 65 and 71 and/or after page 1 (t0) (cluster 2).
- Page 11 (t1) is visited before page 8 (t3). Page 6 (t3) is visited after page 5 (t3), followed by pages 7 (t3) and 3 (t3) (cluster 3).

Open sequences	1 (47.00%)		2 (26.80%)		3 (26.20%)	
	S	C	S	C	S	C
(1, 68), (0, 1)	2.98	87.50	5.22	11.48	0.00	0.00
(68, 43), (0, 1)	2.98	38.89	0.00	0.00	6.11	6.78
(1, 68), (1, 1)	1.70	66.67	2.24	16.67	0.00	0.00
(68, 43), (0, 2)	1.70	22.22	1.49	8.00	0.00	0.00
(68, 68), (0, 1)	1.70	22.22	0.00	0.00	21.37	23.73
(68, 69), (0, 2)	1.70	22.22	0.00	0.00	0.00	0.00
(61, 63, 60), (1, 1, 1)	1.28	100.00	0.00	0.00	0.00	0.00
(59, 63, 60), (1, 1, 1)	1.28	100.00	0.00	0.00	0.00	0.00
(61, 60), (1, 1)	1.28	100.00	0.00	0.00	0.00	0.00
(61, 63), (1, 1)	1.28	100.00	0.00	0.00	0.00	0.00
(1, 1), (0, 2)	0.00	0.00	14.93	32.79	1.53	15.38
(1, 68), (0, 0)	0.00	0.00	14.18	31.15	7.63	76.92
(1, 2), (0, 0)	0.00	0.00	14.18	31.15	0.00	0.00
(1, 9), (0, 2)	0.00	0.00	13.43	29.51	1.53	15.38
(1, 9), (0, 0)	0.00	0.00	10.45	22.95	0.00	0.00
(1, 68, 65, 2, 9), (0, 0, 1, 0, 2)	0.00	0.00	1.49	100.00	0.00	0.00
(1, 65, 2, 9), (0, 1, 0, 2)	0.00	0.00	1.49	100.00	0.00	0.00
(68, 65, 2, 9), (0, 1, 0, 2)	0.00	0.00	1.49	100.00	0.00	0.00
(65, 2, 9), (1, 0, 2)	0.00	0.00	1.49	100.00	0.00	0.00
(1, 1), (3, 2)	0.00	0.00	2.99	80.00	0.00	0.00
(68, 68), (0, 1)	1.70	22.22	0.00	0.00	21.37	23.73
(68, 65), (0, 0)	0.00	0.00	5.22	28.00	17.56	19.49
(68, 65), (0, 1)	0.00	0.00	3.73	20.00	16.79	18.64
(68, 71), (0, 3)	0.00	0.00	0.00	0.00	15.27	16.95
(68, 65), (0, 2)	0.00	0.00	1.49	8.00	13.74	15.25
(68, 57, 65, 2, 9), (0, 1, 1, 0, 0)	0.00	0.00	0.00	0.00	1.53	100.00
(68, 33, 42, 43, 38), (0, 0, 3, 0, 3)	0.00	0.00	0.00	0.00	1.53	100.00
(68, 43, 33, 42, 38), (0, 1, 0, 3, 3)	0.00	0.00	0.00	0.00	1.53	100.00
(68, 65, 58, 9, 71), (0, 1, 1, 1, 3)	0.00	0.00	0.00	0.00	1.53	100.00
(68, 65, 55, 65, 58), (0, 1, 1, 0, 1)	0.00	0.00	0.00	0.00	1.53	100.00
(68, 65, 55, 70, 2), (0, 1, 1, 1, 0)	0.00	0.00	0.00	0.00	1.53	100.00

Table 4.23: 2-dim SAM applied to dataset 1 (<http://www.luc.ac.be/tew>), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in three clusters.

Open sequences	1 (28.80%)		2 (51.00%)		3 (20.20%)	
	S	C	S	C	S	C
(1, 1), (0, 2)	13.19	76.00	1.18	5.36	0.00	0.00
(1, 9), (0, 2)	4.17	24.00	5.49	25.00	0.00	0.00
(1, 1), (3, 2)	3.47	62.50	0.00	0.00	0.00	0.00
(1, 9), (1, 2)	2.78	44.44	1.57	21.05	0.00	0.00
(1, 2), (1, 2)	2.78	44.44	0.00	0.00	0.00	0.00
(1, 9, 69, 66, 12), (3, 3, 2, 3, 2)	1.39	100.00	0.00	0.00	0.00	0.00
(1, 9, 69, 66, 71), (3, 3, 2, 3, 3)	1.39	100.00	0.00	0.00	0.00	0.00
(1, 9, 69, 2, 12), (3, 3, 2, 3, 2)	1.39	100.00	0.00	0.00	0.00	0.00
(1, 9, 69, 2, 71), (3, 3, 2, 3, 3)	1.39	100.00	0.00	0.00	0.00	0.00
(1, 9, 69, 2, 66), (3, 3, 2, 3, 3)	1.39	100.00	0.00	0.00	0.00	0.00
(68, 68), (0, 1)	0.00	0.00	12.94	20.75	0.00	0.00
(68, 65), (0, 0)	0.00	0.00	11.76	18.87	0.00	0.00
(1, 68), (0, 0)	0.00	0.00	11.37	51.79	0.00	0.00
(68, 65), (0, 1)	0.00	0.00	10.98	17.61	0.00	0.00
(68, 71), (0, 3)	0.00	0.00	8.63	13.84	0.00	0.00
(68, 33, 42, 38), (0, 0, 3, 3)	0.00	0.00	1.18	100.00	0.00	0.00
(61, 59, 63), (1, 1, 1)	0.00	0.00	1.18	100.00	0.00	0.00
(33, 42, 38), (0, 3, 3)	0.00	0.00	1.18	100.00	0.00	0.00
(68, 42, 38), (0, 3, 3)	0.00	0.00	1.18	100.00	0.00	0.00
(68, 33, 38), (0, 0, 3)	0.00	0.00	1.18	100.00	0.00	0.00
(11, 8), (1, 3)	0.00	0.00	0.00	0.00	2.97	100.00
(5, 6, 7, 3), (3, 3, 3, 3)	0.00	0.00	0.00	0.00	1.98	100.00
(5, 7, 3), (3, 3, 3)	0.00	0.00	0.00	0.00	1.98	100.00
(5, 6, 3), (3, 3, 3)	0.00	0.00	0.00	0.00	1.98	100.00
(5, 6, 7), (3, 3, 3)	0.00	0.00	0.00	0.00	1.98	100.00
(6, 7, 3), (3, 3, 3)	0.00	0.00	0.00	0.00	1.98	100.00

Table 4.24: SAM applied to dataset 1 (<http://www.luc.ac.be/te/w>), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in three clusters.

In data set 2, page 657 is considered to be the most frequent requested page (23.86% of the requested pages in data set 2 is page 657, re. figure 4.6). 2-dim SAM discovers the following profiles (re. table 4.25) with regard to page 657:

- Page 657 (t0) is visited before page 947 (t0) (cluster 1).
- Page 657 is visited after (t2) / before (t0) page 802 and/or visited before (t0) page 984 and/or visited before (t0) page 815. Page 657 (t0) is also re-visited (t2) (cluster 2).
- Page 657 (t2) presents one-page sessions (support of one-page sessions for page 657 is 88.50%) (cluster 3).
- Page 657 (t0) is visited before / after page 984 and/or after page 1026. Page 657 (t0) is also re-visited (t2) (cluster 4).

If two-dimensional server sessions are clustered based on SAM distance measures (re. table 4.26), server sessions are not clustered based on sequential relationships with regard to the most requested page (657) and relations between pages and times. Instead, the following profiles are discovered:

- Page 1129 (t0) is visited before page 713 (t0) (cluster 1).
- Page 657 (t0) is visited before / after page 984 and/or before page 713 and/or before page 815. Page 657 (t0) is also re-visited (t2) (cluster 2).

Open sequences	1 (54.40%)		2 (23.40%)		3 (17.40%)		4 (4.80%)	
	S	C	S	C	S	C	S	C
(657, 947), (0, 0)	1.84	35.71	0.00	0.00	0.00	0.00	12.50	12.50
(1129, 713), (0, 0)	1.84	17.86	0.00	0.00	0.00	0.00	0.00	0.00
(804, 190), (1, 1)	1.10	60.00	0.00	0.00	0.00	0.00	0.00	0.00
(947, 996), (0, 1)	1.10	16.67	0.00	0.00	0.00	0.00	0.00	0.00
(657, 657), (0, 2)	0.00	0.00	17.95	30.00	0.00	0.00	0.00	0.00
(802, 657), (1, 2)	0.00	0.00	7.69	64.29	0.00	0.00	37.50	37.50
(657, 802), (0, 1)	0.00	0.00	6.84	11.43	0.00	0.00	0.00	0.00
(657, 984), (0, 0)	0.00	0.00	5.98	10.00	0.00	0.00	54.17	54.17
(657, 815), (0, 0)	0.00	0.00	5.13	8.57	0.00	0.00	20.83	20.83
(657, 786, 794, 786), (0, 1, 1, 1)	0.00	0.00	1.71	100.00	0.00	0.00	0.00	0.00
(657, 786, 794), (0, 1, 1)	0.00	0.00	1.71	100.00	0.00	0.00	0.00	0.00
(657, 1089, 713), (0, 3, 0)	0.00	0.00	1.71	100.00	0.00	0.00	0.00	0.00
(657, 1089, 657), (0, 3, 0)	0.00	0.00	1.71	100.00	0.00	0.00	0.00	0.00
(1082, 657, 1082), (0, 0, 0)	0.00	0.00	1.71	100.00	0.00	0.00	0.00	0.00
One-page sessions	0.77	-	0.05	-	0.98	-	0.00	-
(657, 984), (0, 0)	0.00	0.00	5.98	10.00	0.00	0.00	54.17	54.17
(657, 984, 657), (0, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	37.50	69.23
(984, 657), (0, 0)	0.00	0.00	2.56	16.67	0.00	0.00	37.50	69.23
(657, 657), (0, 2)	0.00	0.00	17.95	30.00	0.00	0.00	37.50	37.50
(1026, 657), (1, 0)	0.00	0.00	0.00	0.00	0.00	0.00	29.17	87.50
(657, 1026, 657, 984), (0, 1, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	12.50	100.00
(657, 815, 657), (0, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	20.83	100.00
(657, 713, 657), (0, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	16.67	100.00
(657, 947, 657), (0, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	12.50	100.00
(1026, 713, 657), (1, 0, 0)	0.00	0.00	0.00	0.00	0.00	0.00	12.50	100.00

Table 4.25: 2-dim SAM applied to dataset 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in four clusters.

Open sequences	1 (76.00%)		2 (24.00%)	
	S	C	S	C
(1129, 713), (0, 0)	1.32	19.23	0.00	0.00
(657, 657), (0, 2)	0.00	0.00	25.00	29.70
(657, 984), (0, 0)	0.00	0.00	16.67	19.80
(984, 657), (0, 0)	0.00	0.00	10.00	48.00
(657, 713), (0, 0)	0.00	0.00	9.17	10.89
(657, 815), (0, 0)	0.00	0.00	9.17	10.89
(657, 984, 1006, 984), (0, 0, 2, 0)	0.00	0.00	1.67	100.00
(657, 984, 1006, 657), (0, 0, 2, 0)	0.00	0.00	1.67	100.00
(657, 984, 663, 657), (0, 0, 1, 0)	0.00	0.00	1.67	100.00
(657, 815, 1082, 657), (0, 0, 0, 0)	0.00	0.00	1.67	100.00
(657, 1025, 984, 1025), (0, 1, 0, 1)	0.00	0.00	1.67	100.00

Table 4.26: SAM applied to dataset 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in two clusters.

In data set 3, most requested pages are page 28 (5.32%), 27 (5.56%), 109 (5.45%), 250 (4.56%), 280 (6.47%), 281 (6.69%), 355 (6.51%) and 358 (4.91%). This is also shown in figure 4.7. A general difference between data sets 1, 2 and 3 is that in the first two data sets one high relative frequency value for one page is shown (re. figures 4.5 and 4.6) whereas in data set 3, several high relative frequency values for several pages are shown. Therefore, for data set 3, profiles discovered by 2-dim SAM distance measures provide information for several pages whereas for data set 1 and 2, profiles provide information that is concentrated on one page.

Generally, if we compare table 4.27 with 4.28, server sessions are more equally spread across clusters based on 2-dim SAM. For example, in table 4.27, cluster 1, 2, 3 and 4 represent respectively 31.40%, 29.20%, 22.60% and 16.80% of the server sessions whereas in table 4.28, cluster 1, 2, 3 and 4 represent respectively 32.40%, 31.80%, 8.20% and 27.60% of the server sessions.

Particularly, if we compare table 4.27 with 4.28, cluster one based on 2-dim SAM is very much alike with cluster one based on SAM. The same is shown for cluster two. The main difference between clustering 2-dimensional server sessions of data set 3 based on 2-dim SAM and SAM is shown by clusters three and four. The profiles presented by clusters three and four in table 4.27 are as follows:

- Page 281 (t0) is visited before page 280 (t0) and/or page 355 (t1) and/or page 492 (t0). Also, page 280 (t0) is visited before page 355 (t1) (cluster 3).
- Page 281 (t0) is visited before page 355 (t1) and/or page 358 (t0). Also, page 280 (t0) is visited before page 355 (t1) and/or page 358 (t0) (cluster 4).

The profiles presented by clusters three and four in table 4.28 are as follows:

- Page 281 (t0) is visited before page 280 (t0) and/or page 355 (t1) and/or page 358 (t0) and/or page 491 (t0). Also, page 280 (t0) is visited before page 355 (t1) and/or page 358 (t0) and/or page 491 (t0). Also, page 355 (t1) is visited before page 492 (t0) (cluster 3).
- Page 281 (t0) is visited before page 280 (t0) and/or page 355 (t1) and/or page 492 (t0). Also, page 280 (t0) is visited before page 355 (t1) and/or page 492 (t0) (cluster 4).

The main difference is that in table 4.27, cluster three and four distinguish server sessions excluding page 358 (cluster 3) and including page 358 (cluster 4). In table 4.28, cluster three and four distinguish server sessions including pages 358 and 491 (cluster 3) and excluding pages 358 and 491 (cluster 4).

Open sequences	1 (31.40%)		2 (29.20%)		3 (22.60%)		4 (16.80%)	
	S	C	S	C	S	C	S	C
(28, 109), (0, 0)	98.09	98.09	2.05	33.33	1.77	100.00	0.00	0.00
(27, 109), (0, 0)	87.26	87.26	0.00	0.00	1.77	100.00	0.00	0.00
(109, 250), (0, 0)	86.62	87.18	0.00	0.00	1.77	100.00	0.00	0.00
(28, 250), (0, 0)	86.62	86.62	1.37	22.22	1.77	100.00	0.00	0.00
(27, 250), (0, 0)	86.62	86.62	0.00	0.00	1.77	100.00	0.00	0.00
(28, 109, 250), (0, 0, 0)	85.35	87.01	0.00	0.00	1.77	100.00	0.00	0.00
(28, 109, 250, 113, 52, 64), (0, 0, 0, 0, 2, 1)	5.10	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 250, 113, 52, 64), (0, 0, 0, 2, 1)	5.10	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 109, 250, 52, 64), (0, 0, 0, 2, 1)	5.10	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 109, 113, 52, 64), (0, 0, 0, 2, 1)	5.10	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(28, 109, 249, 52, 64), (0, 0, 0, 2, 1)	5.10	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(305, 317), (2, 2)	0.00	0.00	9.59	87.50	0.00	0.00	0.00	0.00
(286, 317), (2, 2)	0.00	0.00	8.90	86.67	0.00	0.00	0.00	0.00
(305, 286), (2, 2)	0.00	0.00	6.85	62.50	0.00	0.00	0.00	0.00
(305, 286, 317), (2, 2, 2)	0.00	0.00	5.48	80.00	0.00	0.00	0.00	0.00
(286, 305, 317), (2, 2, 2)	0.00	0.00	2.74	100.00	0.00	0.00	0.00	0.00
(305, 286, 317, 307, 311), (2, 2, 2, 2, 3)	0.00	0.00	1.37	100.00	0.00	0.00	0.00	0.00
(234, 227, 237, 244, 230), (3, 2, 3, 3, 3)	0.00	0.00	1.37	100.00	0.00	0.00	0.00	0.00
(286, 317, 307, 311), (2, 2, 2, 3)	0.00	0.00	2.05	100.00	0.00	0.00	0.00	0.00
(305, 286, 307, 311), (2, 2, 2, 3)	0.00	0.00	1.37	100.00	0.00	0.00	0.00	0.00
(305, 317, 307, 311), (2, 2, 2, 3)	0.00	0.00	1.37	100.00	0.00	0.00	0.00	0.00
(281, 280), (0, 0)	0.00	0.00	0.00	0.00	100.00	100.00	44.05	52.11
(281, 280, 355), (0, 0, 1)	0.00	0.00	0.00	0.00	99.12	99.12	0.00	0.00
(281, 355), (0, 1)	0.00	0.00	0.00	0.00	99.12	99.12	75.00	88.73
(280, 355), (0, 1)	0.00	0.00	0.00	0.00	99.12	99.12	53.57	66.18
(281, 280, 492), (0, 0, 0)	0.00	0.00	0.00	0.00	92.04	92.04	0.00	0.00
(281, 280, 355, 491, 297, 328), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	2.65	100.00	0.00	0.00
(281, 280, 355, 491, 340, 328), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	2.65	100.00	0.00	0.00
(281, 280, 355, 491, 340, 297), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	2.65	100.00	0.00	0.00
(281, 280, 355, 491, 295, 328), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	2.65	100.00	0.00	0.00
(281, 280, 355, 491, 295, 297), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	2.65	100.00	0.00	0.00
(281, 355), (0, 1)	0.00	0.00	0.00	0.00	99.12	99.12	75.00	88.73
(280, 355), (0, 1)	0.00	0.00	0.00	0.00	99.12	99.12	53.57	66.18

Open sequences	1 (31.40%)		2 (29.20%)		3 (22.60%)		4 (16.80%)	
	S	C	S	C	S	C	S	C
(280, 358), (0, 0)	0.00	0.00	0.00	0.00	74.34	74.34	52.38	64.71
(281, 358), (0, 0)	0.00	0.00	0.00	0.00	74.34	74.34	50.00	59.15
(281, 355, 358), (0, 1, 0)	0.00	0.00	0.00	0.00	73.35	74.11	47.62	63.49
(281, 280, 355, 358, 483), (0, 0, 1, 0, 3)	0.00	0.00	0.00	0.00	0.00	0.00	5.95	100.00
(281, 355, 475, 474, 483), (0, 1, 3, 3, 3)	0.00	0.00	0.00	0.00	0.00	0.00	5.95	100.00
(281, 355, 468, 474, 483), (0, 1, 3, 3, 3)	0.00	0.00	0.00	0.00	0.00	0.00	5.95	100.00
(281, 280, 355, 492, 483), (0, 0, 1, 0, 3)	0.00	0.00	0.00	0.00	0.00	0.00	5.95	100.00
(281, 280, 468, 474, 483), (0, 0, 3, 3, 3)	0.00	0.00	0.00	0.00	0.00	0.00	5.95	100.00

Table 4.25: 2-dim SAM applied to dataset 3 (Belgian telecom provider), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in four clusters.

Open sequences	1 (32.40%)		2 (31.80%)		3 (8.20%)		4 (27.60%)	
	S	C	S	C	S	C	S	C
(28, 109), (0, 0)	96.91	96.91	1.26	33.33	0.00	0.00	0.00	0.00
(28, 250), (0, 0)	86.42	86.42	0.00	0.00	0.00	0.00	0.00	0.00
(109, 250), (0, 0)	85.80	87.42	0.00	0.00	0.00	0.00	0.00	0.00
(27, 250), (0, 0)	85.80	86.34	0.00	0.00	0.00	0.00	0.00	0.00
(27, 109), (0, 0)	85.80	86.34	0.00	0.00	0.00	0.00	0.00	0.00
(27, 28, 109, 250, 116, 161), (0, 0, 0, 0, 2, 1)	1.23	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 28, 109, 113, 52, 33), (0, 0, 0, 0, 2, 3)	1.23	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 28, 109, 113, 116, 161), (0, 0, 0, 0, 2, 1)	1.23	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 28, 109, 250, 52, 64), (0, 0, 0, 0, 2, 1)	1.23	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(27, 28, 109, 250, 52, 33), (0, 0, 0, 0, 2, 3)	1.23	100.00	0.00	0.00	0.00	0.00	0.00	0.00
(305, 317), (2, 2)	0.00	0.00	8.18	86.67	0.00	0.00	1.45	100.00
(286, 317), (2, 2)	0.00	0.00	7.55	85.71	0.00	0.00	1.45	100.00
(305, 286), (2, 2)	0.00	0.00	6.29	66.67	0.00	0.00	0.00	0.00
(305, 286, 317), (2, 2, 2)	0.00	0.00	5.03	80.00	0.00	0.00	0.00	0.00
(45, 76), (2, 2)	0.00	0.00	2.52	100.00	0.00	0.00	0.00	0.00
(305, 286, 317, 307, 311), (2, 2, 2, 2, 3)	0.00	0.00	1.26	100.00	0.00	0.00	0.00	0.00
(286, 317, 307, 311), (2, 2, 2, 3)	0.00	0.00	1.89	100.00	0.00	0.00	0.00	0.00
(492, 358, 485, 480), (0, 0, 3, 3)	0.00	0.00	1.26	100.00	0.00	0.00	0.00	0.00
(305, 286, 307, 311), (2, 2, 2, 3)	0.00	0.00	1.26	100.00	0.00	0.00	0.00	0.00
(305, 317, 307, 311), (2, 2, 2, 3)	0.00	0.00	1.26	100.00	0.00	0.00	0.00	0.00
(281, 280), (0, 0)	1.85	100.00	0.00	0.00	95.12	97.50	79.71	80.29
(281, 355), (0, 1)	1.85	100.00	0.00	0.00	92.68	95.00	97.83	98.54
(355, 492), (1, 0)	1.23	66.67	0.00	0.00	82.93	85.00	75.36	77.04
(280, 358), (0, 0)	1.85	100.00	0.00	0.00	82.93	85.00	66.67	67.15
(280, 491), (0, 0)	1.85	100.00	0.00	0.00	82.93	85.00	64.49	64.96
(280, 355), (0, 1)	1.85	100.00	0.00	0.00	82.93	85.00	86.96	87.59
(281, 358), (0, 0)	1.85	100.00	0.00	0.00	82.93	85.00	65.22	65.69
(281, 491), (0, 0)	1.85	100.00	0.00	0.00	82.93	85.00	63.04	63.50
(281, 280, 355, 491, 449, 452), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	4.88	100.00	0.00	0.00
(281, 280, 355, 491, 326, 339), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	4.88	100.00	0.00	0.00
(281, 280, 355, 491, 326, 328), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	4.88	100.00	0.00	0.00
(281, 280, 355, 491, 326, 297), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	4.88	100.00	0.00	0.00
(281, 280, 355, 491, 326, 340), (0, 0, 1, 0, 2, 2)	0.00	0.00	0.00	0.00	4.88	100.00	0.00	0.00

Open sequences	1 (32.40%)		2 (31.80%)		3 (8.20%)		4 (27.60%)	
	S	C	S	C	S	C	S	C
(281, 355), (0, 1)	1.85	100.00	0.00	0.00	92.68	95.00	97.83	98.54
(280, 355), (0, 1)	1.85	100.00	0.00	0.00	82.93	85.00	86.96	87.59
(281, 280), (0, 0)	1.85	100.00	0.00	0.00	95.12	97.50	79.71	80.29
(280, 492), (0, 0)	1.85	100.00	0.00	0.00	80.49	82.50	77.54	78.10
(281, 492), (0, 0)	1.85	100.00	0.00	0.00	80.49	82.50	76.09	76.64
(281, 280, 355, 492, 363, 356), (0, 0, 1, 0, 2, 1)	0.00	0.00	0.00	0.00	0.00	0.00	1.45	100.00
(281, 280, 492, 363, 356), (0, 0, 0, 2, 1)	0.00	0.00	0.00	0.00	0.00	0.00	1.45	100.00
(281, 280, 355, 275, 272), (0, 0, 1, 3, 3)	0.00	0.00	0.00	0.00	0.00	0.00	1.45	100.00
(281, 355, 492, 363, 356), (0, 1, 0, 2, 1)	0.00	0.00	0.00	0.00	0.00	0.00	1.45	100.00
(280, 355, 392, 363, 356), (0, 1, 0, 2, 1)	0.00	0.00	0.00	0.00	0.00	0.00	1.45	100.00

Table 4.28: SAM applied to dataset 3 (Belgian telecom provider), server sessions consisting of visited pages and categories of visiting page time: Evaluating open sequences in four clusters.

4.12.2 Deploying the results

4.12.2.1 Suggestions for structure and service improvement of the web site <http://www.luc.ac.be/tew>

If 2-dim SAM is applied to server sessions consisting of visited pages and categories of visiting page time, information provided by profiles related with page 68 or <http://www.luc.ac.be/tew/opleidingen> (which is the most requested page) may be deployed as follows. First, web personalization systems may provide better and faster services to the web visitor and predict the urgency of page deliveries. Some examples are given below. Note that for interpretations of categories of visiting page time table 4.7 is used:

- If visitors go to the home page <http://www.luc.ac.be/tew> and stay on this page for less than 12.89 seconds (t_0), the probability is 87.50% that they will visit thereafter <http://www.luc.ac.be/tew/opleidingen> and stay on this page between 12.89 and 56 seconds (t_1) (re. open sequence (1, 68), (0, 1) in cluster 1).
- If visitors go to the home page <http://www.luc.ac.be/tew> and stay on this page for less than 12.89 seconds (t_0), the probability is 31.15% that they will visit thereafter <http://www.luc.ac.be/tew/opleidingen> and stay on this page for less than 12.89 seconds (re. open sequence (1, 68), (0, 0) in cluster 2).
- The probability is 100% that <http://www.luc.ac.be/tew/informatie> is requested and visited between 56 and 166 seconds if a visitor follows the pattern and respects sequential order and visiting page times:
<http://www.luc.ac.be/tew> (visiting page time less than 12.89 seconds), followed by <http://www.luc.ac.be/tew/opleidingen> (visiting page time less than 12.89 seconds), followed by http://www.luc.ac.be/tew/opleidingen/basisopleiding/opbouw_tew (visiting page time between 12.89 and 56 seconds), followed by http://www.luc.ac.be/tew/diensten/diensten_voor_students (visiting page time less than 12.89 seconds) (re. open sequence (1, 68, 65, 2, 9), (0, 0, 1, 0, 2) in cluster 2).
- The probability is 100% that <http://www.luc.ac.be/tew/informatie> is requested and visited for less than 12.89 seconds if a visitor follows the pattern and respects sequential order and visiting page times:

<http://www.luc.ac.be/tew/opleidingen> (visiting page time less than 12.89 seconds), followed by http://www.luc.ac.be/tew/opleidingen/basisopleidingen/opbouw_tew/curriculum/cur2kantew.html (visiting page time between 12.89 and 56 seconds), followed by http://www.luc.ac.be/tew/opleidingen/basisopleiding/opbouw_tew (visiting page time between 12.89 and 56 seconds), followed by http://www.luc.ac.be/tew/diensten/diensten_voor_students (visiting page time less than 12.89 seconds) (re. open sequence (68, 57, 65, 2, 9), (0, 1, 1, 0, 0) in cluster 3).

Second, navigational pages and content pages related with <http://www.luc.ac.be/tew/opleidingen> may be verified with regard to their actual use by the web visitors. If, in data set 1, navigational pages with visiting page time of more than 12.89 seconds are found, they are not used as intended and structured by the web developer. Likewise, if content pages with visiting page time of less than 12.89 seconds are found, they are not used the way they should be used. Some examples are given below:

- Generally, <http://www.luc.ac.be/tew/opleidingen> (68) is actually used as a navigation page (cluster 2 and 3). Yet, <http://www.luc.ac.be/tew/opleidingen> is related with other navigational pages that are actually used as content pages. Examples are http://www.luc.ac.be/tew/opleidingen/basisopleiding/opbouw_tew (65) and <http://www.luc.ac.be/tew/informatie> (9) (cluster 2).

If SAM is used as distance measure between two-dimensional server sessions, not every cluster is concentrated on profile information with regard to the most requested page <http://www.luc.ac.be/tew/opleidingen> (68) in data set 1. Instead, each cluster represents unique page combinations. Instead of extracting information that is concentrated on the most requested web page, now profiles provide more general information, which may also be used to predict page requests and to verify whether navigational and content pages are used as intended and structured in the web site.

4.12.2.2 Suggestions for structure and service improvement of the web site <http://machines.hyperreal.org>

If 2-dim SAM is applied to server sessions consisting of visited pages and categories of visiting page time, information provided by profiles related with page 657 or **<http://machines.hyperreal.org>** (which is the most requested page) may be deployed as follows. First, web personalization systems may provide

better and faster services to the web visitor and predict the urgency of page deliveries. Some examples are given below. Note that for interpretations of categories of visiting page time table 4.7 is used:

- If visitors go to the home page <http://machines.hyperreal.org> and stay on this page for less than 9.63 seconds (t_0), the probability is 35.71% that they will visit thereafter <http://machines.hyperreal.org/manufacturers/Moog> and stay on this page for less than 9.63 seconds (re. open sequence (657, 947), (0, 0) in cluster 1).
- If visitors go to <http://machines.hyperreal.org/manufacturers/categories/DR-660> and stay on this page between 9.63 and 68 seconds (t_1), the probability is 64.29% that they will visit thereafter <http://machines.hyperreal.org> and stay on this page between 68 and 204 seconds (t_2) (re. open sequence (802, 657), (1, 2) in cluster 2).
- <http://machines.hyperreal.org> is the only page that is visited. The average visiting page time of all the requests of <http://machiens.hyperreal.org> in data set 2 lies between 68 and 204 seconds (cluster 3).
- If visitors go to <http://machines.hyperreal.org> and stay on this page for less than 9.63 seconds, followed by <http://machines.hyperreal.org/manufacturers/Roland> with visiting page time less than 9.63 seconds, the probability is 69.23% that they will re-visit thereafter <http://machines.hyperreal.org> and stay on this page for less than 9.63 seconds (re. open sequence (657, 984, 657), (0, 0, 0) in cluster 4).

Second, navigational pages and content pages related with <http://machines.hyperreal.org> may be verified with regard to their actual use by the web visitors. If, in data set 2, navigational pages with visiting page time of more than 9.63 seconds are found, they are not used as intended and structured by the web developer. Likewise, if content pages with visiting page time of less than 9.63 seconds are found, they are not used the way they should be used. Some examples are given below:

- Generally, <http://machines.hyperreal.org> is actually used as a navigation page except when the home page is re-visited (cluster 2 and 4) and when <http://machines.hyperreal.org/manufacturers/categories/DR-660> (802) precedes the home page (cluster 2).

If SAM is used as distance measure between two-dimensional server sessions, not every cluster is concentrated on profile information with regard to the most requested page, which is <http://machines.hyperreal.org> in data set 2. Instead, each cluster distinguishes unique page combinations. Instead of extracting information that is concentrated on the most requested web page, now profiles provide more general information, which may also be used to predict page requests and to verify whether navigational and content pages are used as intended and structured in the web site.

4.12.2.3 Suggestions for structure and service improvement of the web site of a Belgian telecom provider

If 2-dim SAM is applied to server sessions consisting of visited pages and categories of visiting page time, information provided by profiles related with several highly requested pages, such as **FR General (28)**, **FR Main (109)**, **FR Games (250)**, **DU Welcome (280)**, **DU General (281)**, **DU Main (355)** and **DU Main New (358)** may be deployed as follows. First, web personalization systems may provide better and faster services to the web visitor and predict the urgency of page deliveries. Some examples are given below. Note that for interpretations of categories of visiting page time table 4.7 is used:

- If visitors go to the home page FR General and stay on this page for less than 5.86 seconds (t_0), the probability is 98.09% that they will visit thereafter FR Main and stay on this page for less than 5.86 seconds (re. open sequence (28, 109), (0, 0) in cluster 1).
- If visitors go to DU General and stay on this page for less than 5.86 seconds, the probability is 99.12% that they will visit thereafter DU Main and stay on this page between 5.86 and 13 seconds (t_1) (re. open sequence (281, 355), (0, 1) in cluster 3).

Second, navigational pages and content pages may be verified with regard to their actual use by the web visitors. If, in data set 3, navigational pages with visiting page time of more than 5.86 seconds are found, they are not used as intended and structured by the web developer. Likewise, if content pages with visiting page time of less than 5.86 seconds are found, they are not used the way they should be used. Some examples are given below:

- The home pages in French and Dutch language, FR General (28), FR Welcome (27), DU General (281) and DU Welcome (280) are all visited

for less than 5.86 seconds, which is conform the intentions of the web developer (cluster 1, 2, 3, 4).

- Page DU Main (355), which is structured as navigation page, is actually used as a content page and visited between 5.86 and 13 seconds (cluster 3, 4).

If SAM is used as distance measure between two-dimensional server sessions in data set 3, small differences are shown in clustering results compared with those based on 2-dim SAM distance measures.

4.13 Conclusion and Future Research

4.13.1 Conclusion

In this chapter, the surplus value of clustering server sessions based on SAM and 2-dim SAM distance measures is demonstrated on real log file data registering visiting behaviour on three different web sites. Experimental tests analyse Web Usage Data from <http://www.luc.ac.be/tew>, <http://machines.hyperreal.org> and the web site of a Belgian telecom provider. For privacy reasons we omit the URL address of the last web site.

SAM measures distances between server sessions taking into account *equalities of visited pages* and the *order of visited pages*. Clustering based on SAM groups server sessions with equal visited pages and similar order of visited pages together. Generally, SAM is used for server sessions consisting of one attribute or dimension, which are visited pages. SAM is compared with a method that does not incorporate order-based information, called Association distance. Given the results of our experimental tests, SAM-based clustering performs better than Association-based clustering for the following reasons:

- Different statistical measures for defining the number of clusters behave towards the same level of solutions if SAM is used as distance measure. Yet, if Association distance is used, for two of the three data sets statistical measures lead to different levels of clustering solutions.
- The distribution of (groups of) page_ids in different clusters is generally better represented when SAM is used as distance measure between server sessions. Also, at a general level, if page x is highly represented in cluster

a, it is (nearly) not represented in cluster b following SAM-based clustering.

- Order-based information is better given by clusters based on SAM. This might indicate that the model based on SAM fits the data better.
- SAM is less sensitive to the length of server sessions when compared to Association.

2-dim SAM measures distances between server sessions taking into account the same features as SAM and additionally measures *equalities of categories of visiting page time* as well as *relations between visited pages and categories of visiting page time*. Clustering based on 2-dim SAM groups server sessions together based on four characteristics: equal visited pages, equal categories of visiting page time, similar order of visited pages and similar relations between visited pages and categories of visiting page time. 2-dim SAM is used for 2-dimensional server sessions or server sessions consisting of two attributes or dimensions, which are visited pages and categories of visiting page time. In order to show that 2-dim SAM is more able to distinguish server sessions with regard to relations between visited pages and categories of visiting page time, the data sets of the three different web sites (providing server sessions consisting of two attributes) are also analysed by means of SAM. Given the results of our experimental tests, the main difference between clusters based on 2-dim SAM and SAM are the following:

- Generally, support values of two-dimensional open sequences selected for cluster description are higher for clusters based on 2-dim SAM. This might indicate that two-dimensional server sessions are better clustered with regard to relations between visited pages and categories of visiting page time.
- Generally, 2-dimensional server sessions are more equally distributed across clusters based on 2-dim SAM. This means that relatively large groups of visiting behaviour are found based on 2-dim SAM. Since we are interested in adjusting the web site conform to large groups of visiting behaviour, clustering 2-dimensional server sessions based on 2-dim SAM might provide better results.
- If a particular web page is highly requested relative to other requests, clustering based on 2-dim SAM tends to group 2-dimensional server sessions together based on equal visiting behaviour related to the web page

that is highly requested. Practically, this means that different profiles with regard to the most requested web page are discovered.

- If a particular web page is highly requested relative to other requests, clustering based on SAM tends to group 2-dimensional server sessions together irrelevant of their relation to the web page that is highly requested. Practically, this means that one profile is discovered with regard to the highest requested web page. The remaining profiles are not related with the highest requested web page. Yet, every profile is different and provides information about different visiting behaviour.
- If several web pages are highly requested relative to others, clustering based on 2-dim SAM as well as SAM tend to group 2-dimensional server sessions together based on equal visiting behaviour related to the web pages that are highly requested. Practically, this means that different profiles, mostly with regard to the high requested web pages, are discovered.
- Clustering based on 2-dim SAM groups 2-dimensional server sessions together based on equalities, order-based information *and* relations between visited pages and categories of visiting page time. Clustering based on SAM groups 2-dimensional server sessions together based on equalities and order-based information of visited pages and categories of visiting page time, without considering the relations between the two attributes.

4.13.2 Future research

Future research should employ other heuristics or algorithms for identifying users and server sessions. The influence of user and session identification should be measured on the final results. In Berend et al (2001), the accuracy of sessionizers for web usage analysis is measured. Also, the impact of caching must be examined. Analysis on sequences of requests may be affected by cached pages since backward moves to previously seen pages should be part of the analysis (Berend et al, 2001). Likewise, cached pages affect the visiting page times if looking at cached pages takes so long that sessions are erroneously split (Berend et al, 2001). The only way to effectively study the impact of caching is by deploying client-side agents that keep track of users' actions (Berend et al, 2001).

Algorithms should be developed for analysing continuous, numerical values for visiting page time information, instead of using categories. As a general rule, neural networks (Craven and Shavlik, 1998) work best on data sets with a large number of numeric attributes (Mena, 1999).

Another topic for future research is examining whether time differences between logged page requests on the server and real page delivery to the user are significant. If so, the visiting page times that are based on the time of page requests logged on the server must be reconsidered.

Another topic for future research concerns the boundary calculation for t_1 , t_2 and t_3 . We defined boundaries based on equal number of requests. Instead, boundaries may be defined based on equal total visiting page time. For example, the total visiting page time of all the requests above 12.89 seconds in the first data set amounts up to 1,027,500 seconds. Following, actual visiting page times are ordered ascending starting with 12.891 up to 1799 seconds. Each time one actual visiting page time is assigned to t_1 until the total visiting page time within time category t_1 is equal to 342,500 (i.e. $1,027,500 / 3$) seconds. The boundary of t_1 is equal to the last visiting page time that was added to t_1 before total visiting page time exceeded 342,500 seconds. Then, visiting page times are assigned to t_2 and t_3 accordingly. Table 4.29 provides the boundaries based on equal total visiting page time within each time category for the datasets used in our analyses. More research is necessary to evaluate whether other methods for calculating boundaries for t_1 , t_2 and t_3 provide significant differences in the results.

Dataset		
1	2	3
$0 < t_0 \leq 12.89$	$0 < t_0 \leq 9.63$	$0 < t_0 \leq 5.86$
$12.89 < t_1 \leq 203$	$9.63 < t_1 \leq 209$	$5.86 < t_1 \leq 208$
$203 < t_2 \leq 562$	$209 < t_2 \leq 322$	$208 < t_2 \leq 335$
$562 < t_3$	$322 < t_3$	$335 < t_3$

Table 4.29: Visiting page time categories based on equal total visiting page time.

Future research should also examine how sensitive the final results are to (small) changes in the value of t_{cutoff} , particularly regarding γ (the number of navigation pages divided by the total number of pages in the analysis) in equation (4.2). For example, with regard to analysing visiting behaviour on <http://www.luc.ac.be/tew>, 7 out of 71 pages are navigation pages, which means that γ equals 0.09859. Yet, further examination is necessary in order to know how increases/decreases in the number of navigational pages might affect the final clustering results.

Besides Ward clustering, other clustering methods may be invoked on the distance matrices as well. In table 4.8 of section 4.7, the dissimilarities between clusters are calculated using Single linkage, Complete linkage, Average linkage and Centroid linkage.

Future research should also explore how open sequences might be used before the mining process takes place. In our experiments, open sequences are used after the mining process or in the post-processing step. The reasons why open sequences are searched after the mining process are given in section 4.8. Nevertheless, in order to employ open sequences before the mining process, a cutoff value needs to be defined to extract open sequences with high support and/or confidence values, while, at the same time, employing an algorithm which notifies when valuable information embedded in server sessions might be lost.

Throughout the analysis of invoking hierarchical clustering algorithms on distance matrices including SAM or 2-DIM SAM distance measures between server sessions, we used an agglomerative clustering method (i.e. Ward). Future research should examine whether divisive hierarchical clustering techniques are suitable as well. Divisive methods proceed by splitting the data set into smaller and smaller clusters until each object belongs to a separate cluster (Kaufman and Rousseeuw, 1990). Generally, divisive methods are computationally more complex than agglomerative methods. Nevertheless, it is possible to construct divisive methods that do not consider *all* divisions, because many of them would be totally inappropriate anyway. Therefore, Kaufman and Rousseeuw (1990) suggest an algorithm based on the proposal of Macnaughton-Smith et al (1964).

For each of the three data sets that we examined, one-page sessions are not omitted from the analysis because we believe that they might provide real visiting behaviour. However, care must be taken for interpreting visiting page times of one-page sessions, since time difference between request and subsequent request cannot be calculated. Therefore, within our analyses, we used the average visiting page time for one-page sessions. Future research should examine whether omitting one-page sessions from the data leads to different results.

SAM and 2-dim SAM analyses are executed on individual page_ids and not on groups of page_ids. Future research should examine whether analysing web surfing behaviour at a higher hierarchical level, using groups of pages (also known as classes of pages), provides meaningful results. This way, suggestions for links between classes of pages instead of individual pages may be provided.

Since web sites currently evolve into dynamic data repositories, further research is necessary in applying SAM-based clustering to dynamic data sets. In order to examine order-based information of dynamic pages (instead of static pages), other pre-processing techniques for constructing server sessions are necessary. Instead of using only the information recorded in log files, extra information of CGI scripts, SSI and/or cookies may construct sequences consisting of page_id and text. SAM is able to calculate distances between

sequences consisting of both numeric and alphabetic information. Finally, clustering based on SAM may provide large groups of visiting patterns offering order-based information of dynamic web pages.

CHAPTER 5

SAM AND INTERESTINGNESS

In this chapter, a new algorithm called Sequence Alignment Method integrated with an Interestingness Measure (SAM^I) (Hay et al, 2003a) is illustrated for mining navigation patterns on a web site. Through log file analysis, SAM^I distinguishes *interesting patterns* (i.e. unexpected, surprising patterns contradicting with the structure of the web site or direct hyperlinks between web pages) from *uninteresting patterns* (i.e. expected, known, obvious patterns resulting from the structure of the web site or direct hyperlinks between web pages) and provides information about the order of visited web pages. The algorithm is validated using real data sets of the Music Machines web site <http://machines.hyperreal.org>, home of musical electronics on the web. Empirical results show that SAM^I identifies profiles of visiting behaviour, which may be used for web personalization techniques and for optimising the layout of the web site through structuring of page-links.

5.1 Motivation

In the previous chapter, we discovered knowledge about visiting patterns on three different web sites. The extracted knowledge from log file analysis is presented by means of visiting profiles showing visited pages (and visiting times) along with order-based information. Clustering server sessions by means of SAM or two-dimensional SAM provides a *general* overview of how people visit a web site, without distinguishing interesting patterns from uninteresting patterns. *Interesting patterns* are unexpected, surprising patterns contradicting with the structure of the web site or direct hyperlinks between web pages. *Uninteresting patterns* are expected, known, obvious patterns resulting from the structure of the web site or direct hyperlinks between web pages. For example, with regard to the analysis of the log files of the web site <http://www.luc.ac.be/tew> in chapter four (re. section 4.9.1), uninteresting patterns are navigations from page 1 to page 9 and from page 1 to page 68. The relatively strong support measures of these patterns result from direct hyperlinks between page 1 and 9 and page 1 and 68. Also, navigating from page 2 to page 40 might not be interesting since the support is expected to be relatively low, due to the absence of links between pages 2 and 40. However, navigating from page 68 to page 55 might be surprising and therefore interesting due to a relatively high support without any direct hyperlinks between pages 68 and 55. In other words, if two different web pages (i.e. web pages having different URL addresses) A and B are connected by a direct hyperlink from A to B, visiting pattern AB may be interesting if the support is relatively low. Yet, if A and B are not connected by a direct hyperlink, visiting pattern AB may be interesting if the support is relatively high.

If the web administrator wants to automatically discover only interesting patterns, instead of a general overview providing both uninteresting and interesting patterns, it is necessary to extend SAM with a measure that distinguishes interesting patterns from uninteresting patterns by looking at support measures and the structure of the web site. Besides, as stated in chapter two, section 2.3, successful data mining projects extract previously unknown and useful information while searching beyond obvious correlations. To address this problem, we integrated SAM with an interestingness measure, based on the structure of the web site. The new method is called SAM¹ and discovers interesting patterns providing visited web pages as well as the order of visited web pages.

5.2 Interestingness measures

Researchers have been working on defining various measures of interestingness for patterns (Liu et al, 1997; Padmanabhan and Tuzhilin, 1998; Piatetsky-Shapiro and Matheus, 1994; Silberschatz and Tuzhilin, 1996). Generally, a common theme among the various criteria for interestingness is *novelty* or *unexpectedness* of a rule. This means that, results that were previously known by the data analyst before the mining process took place, are not considered interesting.

Silberschatz and Tuzhilin (1996) describe two types of interestingness measures. *Objective* measures rate rules based on the data in the analysis. Often thresholds for values of *support*, *confidence* or *chi-square* are used to search for interesting information (Brin et al, 1997; Cooley et al, 1999b). However, high thresholds rarely discover knowledge that was not previously known and low thresholds usually result in an explosion and therefore unmanageable number of rules (Cooley et al, 1999b). *Subjective measures* depend on the class of users who examine the pattern and use two criteria to define whether a rule is interesting: *unexpectedness* and *action ability* or ease of integration within existing processes.

In Liu et al (1997) and Padmanabhan and Tuzhilin (1998), *sets of beliefs* or *general impressions* are used as a filter when searching for interesting rules. Sets of beliefs or general impressions are a-priori information or knowledge about a particular domain. For example, in Web Usage Mining, the web administrator knows that people commonly use the root (home) page of the web site to proceed to other pages within that site. Eventually, rules contradicting the set of beliefs or general impressions are considered interesting. The drawback of this approach is that beliefs and impressions are manually created and, unless a comprehensive set is defined, many interesting results may be lost. Hence, to deal with problems of imprecise or incomplete sets of beliefs, Cooley et al (1999b) developed a method for *automatically* discovering interesting frequent item sets within Web Usage Mining studies using a *framework based on Baldwin's support logic* (Baldwin, 1987), which is specifically designed to handle reasoning about multiple sources of evidence. The algorithm is incorporated into the Web Site Information Filter (WebSIFT) system and evaluated using real web data.

5.3 Approach

In this study, SAM is integrated with the results of the new algorithm, based on Baldwin's support logic and developed by Cooley et al (1999b), in order to *automatically discover interesting visiting patterns* (instead of presenting a general view) *providing visited pages and the order of visited pages*. Practically, this means that server sessions are pre-processed into sessions holding interesting combinations of web pages, before SAM calculates distances between each pair of sessions. This also means that SAM distance measures between pre-processed sessions are used as distance measure for clustering. The resulting clusters provide groups of pre-processed sessions holding interesting combinations of web pages. Moreover, clusters represent profiles of interesting, order-based visiting patterns.

The reason why pre-processing of server sessions, based on the identified interesting related web pages, precedes instead of proceeds the calculation of SAM distance measures is because this approach deals with noise (i.e. uninteresting patterns) in an early stage of the analysis. Data sets within Web Usage Mining studies generally contain lots of patterns that are 'known' or 'obvious' due to the structure of direct hyperlinks between web pages that is offered as a 'navigating road' to web visitors. Dealing with uninteresting patterns in an early stage of the analysis provides an opportunity for SAM to handle large data sets. For example, in section 5.7, 75,855 server sessions are created from log files registering visiting behaviour on the web site <http://machines.hyperreal.org>. The data set is reduced from 75,855 to 7,266 server sessions after pre-processing the server sessions into server sessions holding interesting related, frequently visited web pages. This means that 68,589 server sessions do not hold interesting related, frequently visited web pages. If the original data set of 75,855 server sessions were first used to calculate SAM distance measures, we would end up with an explosion of SAM distance measures (i.e. $[75,855 \times 75,854] / 2 = 2,876,950,000$ SAM distance measures). Moreover, we would also face the problem of distance-based clustering (re. chapter seven) before we could 'post-process' the server sessions into server sessions holding interesting combinations of web pages. This way we unnecessarily burden the analysis with data, which is in fact noise.

The reason why server sessions, holding interesting related web pages, are clustered is to provide large groups of different interesting visiting patterns. This provides an overview of interesting patterns actually occurring on the web site. It also shows small difference in interesting patterns within the same cluster and large differences between interesting patterns across different clusters. If we omit the clustering procedure of server sessions holding

interesting related web pages, it would be difficult to provide an overview of several different large groups of interesting visiting patterns, to examine small differences within the groups and major differences across the groups.

In order to distinguish SAM from SAM based on interestingness, we will use SAM^I to refer to SAM distance measures between server sessions that are pre-processed into sessions holding interesting combinations of web pages, based on our approach of interestingness, which is explained in the following sections. Section 5.4 describes the framework of support logic, which is developed, illustrated and evaluated for Web Usage Mining studies in Cooley et al (1999b). We will also show how sets of beliefs are generated for Web Usage Mining within the support logic framework. In section 5.5, evidence is incorporated into the support logic framework for filtering interesting frequent item sets. Details and examples of this approach are given in Cooley et al (1999b). In section 5.6, we will show how SAM is integrated with the interesting frequent item sets (SAM^I), provided by the results of the support logic framework and evidence combination. After explaining our approach, the method is illustrated on real data sets of visiting behaviour on the web site <http://machines.hyperreal.org> in the following three sections.

For the following sections of this chapter, we define some frequently used concepts. (Interesting) *frequent item sets*, (interesting) *frequently visited pages* or (interesting) *(beliefs of) related (web) pages* all have the same meaning and do not provide order based information. Yet, (interesting) *navigations* or *visiting patterns* provide information about the order of visited pages. *Interesting (web) pages* are pages occurring within interesting frequent item sets or interesting navigations. Furthermore, (interesting) *profiles* refer to clusters grouping server sessions and provide information about the order of visited web pages. Finally, *usage behaviour* is a general term for the behaviour of users on the web and may be based on (interesting) frequent item sets and/or (interesting) navigations.

5.4 Support logic framework

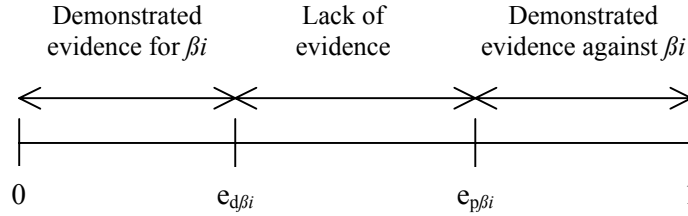
The support logic framework, used within this application, starts with the principles of Baldwin's support logic. We first explain the method and its possibilities. Then we construct the framework with beliefs for Web Usage Mining and we describe how we will apply the method within our research project.

5.4.1 Baldwin's support logic

Baldwin's support logic (Baldwin, 1987) values each piece of information, also called *belief*, by the *evidence for* and *evidence against*. For each such type of evidence, two kinds of evidence definitions exist. *Demonstrated evidence* is evidence that is proven or shown by the data and known by the researcher. *Possible evidence* is evidence that is not proven by the data. The researcher may have an idea about the existence of such evidence but it is not known for sure.

Figure 5.1 illustrates the conceptual frame of evidence (Cooley et al, 1999b). For each belief β_i , demonstrated evidence $e_{d\beta_i}$ and possible evidence $e_{p\beta_i}$ is represented by the evidence pair $[e_{d\beta_i}, e_{p\beta_i}]$. Furthermore, possible evidence against β_i , demonstrated evidence against β_i and lack of evidence with regard to β_i are represented in the framework by respectively $(1 - e_{d\beta_i})$, $(1 - e_{p\beta_i})$ and $(e_{p\beta_i} - e_{d\beta_i})$. Demonstrated as well as possible evidence must be nonnegative. Finally, summing demonstrated evidence supporting β_i with demonstrated evidence against β_i must not be greater than one.

For each belief β_i , evidence pairs are deducted from the framework presented above as follows. For example, assume that evidence is collected about a belief $\beta_1 = (X, Y)$, saying that web pages X and Y are related. If all of the evidence is for, or in support of β_1 , the evidence pair is $[1, 1]$. However, if all of the evidence is against β_1 , the evidence pair is $[0, 0]$. If, say, 20% of the cases in the data support β_1 and, say, 30% of the cases are against β_1 , the evidence pair is $[0.2, 0.7]$, indicating a proven evidence of 0.2 and a possible evidence of 0.7 supporting β_1 . This shows that there is a lack of evidence of 0.5. Finally, if there is no evidence at all to support β_1 , the evidence pair will be $[0, 1]$ indicating a complete lack of evidence of 100%.



where

βi = belief i with $i = 1, 2, \dots, B$;

B = total number of beliefs;

$e_{d\beta i}$ = demonstrated evidence for, in support of, βi ;

$e_{p\beta i}$ = possible evidence for, in support of, βi ;

$(1 - e_{d\beta i})$ = possible evidence against βi ;

$(1 - e_{p\beta i})$ = demonstrated evidence against βi ;

$(e_{p\beta i} - e_{d\beta i})$ = lack of evidence for or against βi ;

$[e_{d\beta i}, e_{p\beta i}]$ = evidence pair of βi ;

$e_{d\beta i} \geq 0$; $e_{p\beta i} \geq 0$; $e_{d\beta i} + (1 - e_{p\beta i}) \leq 1$;

Figure 5.1: Conceptual frame of evidence.

Following Baldwin's support logic programming (Baldwin, 1987), for every belief βi , evidence pairs $[e_{d\beta i}^1, e_{p\beta i}^1]$ and $[e_{d\beta i}^2, e_{p\beta i}^2]$, coming from two different sources, are combined into one evidence pair $[e_{d\beta i}^c, e_{p\beta i}^c]$ as follows:

$$K = 1 - e_{d\beta i}^1 (1 - e_{p\beta i}^2) - e_{d\beta i}^2 (1 - e_{p\beta i}^1) \quad (5.1)$$

$$e_{d\beta i}^c = [e_{d\beta i}^1 e_{d\beta i}^2 + e_{d\beta i}^1 (e_{p\beta i}^2 - e_{d\beta i}^2) + e_{d\beta i}^2 (e_{p\beta i}^1 - e_{d\beta i}^1)] / K \quad (5.2)$$

$$e_{p\beta i}^c = -[[(1 - e_{p\beta i}^1) (1 - e_{p\beta i}^2) + (e_{p\beta i}^1 - e_{d\beta i}^1) (1 - e_{p\beta i}^2) + (e_{p\beta i}^2 - e_{d\beta i}^2) (1 - e_{p\beta i}^1)] / K] + 1 \quad (5.3)$$

where

K = the difference between 1 and the sum of the products between demonstrated evidence for βi , provided by the first/second source and demonstrated evidence against βi , provided by the second/first source;

K = scaling factor, in order to satisfy the following conditions regarding combined evidence pair $[e_{d\beta i}^c, e_{p\beta i}^c]$:

$$0 \leq e_{d\beta i}^c \leq 1, 0 \leq e_{p\beta i}^c \leq 1, e_{d\beta i}^c \leq e_{p\beta i}^c;$$

$K > 0$;

$e_{d\beta i}^c$ = indicated by demonstrated evidence for βi , lack of evidence regarding βi , provided by both sources and divided by K;
 $e_{p\beta i}^c$ = indicated by demonstrated evidence against βi , lack of evidence regarding βi , provided by both sources, divided by K and summed to 1 in order to satisfy $e_{p\beta i}^c \geq 0$;

In the support logic framework, three types of comparisons are made between evidence sources. Comparing one of the original evidence sources with combined evidence identifies beliefs with conflicting evidence along with evidence only represented in the other source as interesting. This is useful when evidence pairs of one source are ‘known’ and evidence pairs of the other source are ‘new’. By combining the new evidence sources and comparing the known evidence pairs to the combined evidence pairs, all of the previously unknown and conflicting results will be labelled as interesting. If the two evidence sources are directly combined, all beliefs that have evidence from one of the sources are declared interesting in addition to any conflicting beliefs. This is useful when both sources of evidence are considered to be ‘new’ (Cooley et al, 1999b).

Examples of types of comparisons between different sources of evidence that would make sense within Web usage Mining Studies are given in table 5.1. Suppose you start analysing visiting behaviour on your web site one year after the site was developed. In order to find interesting visiting behaviour, source one, representing structure data of direct hyperlinks between web pages, is compared with source two, representing usage data stored in log files. Also, in order to be able to reason about evidence, coming from multiple sources, about a given belief, source one and two may be combined into 1_2. This way, instead of looking for beliefs with conflicting evidence from different sources, beliefs are identified with relatively strong/weak evidence from source two and relatively weak/strong evidence from source one (1 vs 1_2 and 2 vs 1_2). More information about comparisons with regard to combining evidence from structure and usage data is given in section 5.5.4. After interpretation of the results, the structure of the web site is adjusted conform to the behaviour of visitors, represented by evidence source 3. Two years after the site was initially developed, source three may be compared with source four in order to find interesting behaviour that may be useful for re-adjusting the layout of the web site conform to interesting visiting behaviour. Also, in order to search for interesting patterns across several years, source one and three are combined into source 1_3 while source two and four are combined into source 2_4. Both combinations are compared. Suppose that, after the second year no major changes are applied in the structure of the web site, source four and five may be combined and compared with source 3. More information about data sources

that are used to define interesting visiting behaviour throughout this thesis is given in the next sub-section.

Evidence source	Combined evidence	Type of comparison
1 = structure data year 1 2 = usage data year 1	1 and 2 = 1_2	1 vs 2 1 vs 1_2 2 vs 1_2
3 = structure data year 2 4 = usage data year 2	1 and 3 = 1_3 2 and 4 = 2_4	3 vs 4 1_3 vs 2_4
5 = usage data year 3	4 and 5 = 4_5	3 vs 4_5
...

Table 5.1: Examples of types of comparisons between different sources of evidence within Web Usage Mining studies.

Interesting results are defined as either a belief with a combined evidence pair that is significantly different from (conflicting with) one of the original evidence pairs, or original evidence pairs that are significantly different from (conflicting among) each other. *Significantly different* is determined by setting a threshold value τ for the differences between the evidence pairs. For a high value of τ , which is at or above 0.5 (re. Cooley et al, 1999b), relatively strong differences between evidence pairs provide high interesting results. For a low value of τ , which is below 0.5 (re. Cooley et al, 1999b), relatively weak differences between evidence pairs provide low interesting results. Ultimately, a belief β_i is interesting if:

$$\tau \leq IM_{\beta_i} \quad (5.4)$$

where

$$IM_{\beta_i} = \sqrt{(e_{d\beta_i})^2 + (e_{p\beta_i})^2} = \text{interestingness measure for } \beta_i;$$

$$e_{d\beta_i} = |e_{d\beta_i}^1 - e_{d\beta_i}^2|;$$

$$e_{p\beta_i} = |e_{p\beta_i}^1 - e_{p\beta_i}^2|;$$

$e_{d\beta_i}^1$ = demonstrated evidence for, in support of, β_i , provided by source 1;

$e_{d\beta_i}^2$ = demonstrated evidence for, in support of, β_i , provided by source 2;

$e_{p\beta_i}^1$ = possible evidence for, in support of, β_i , provided by source 1;

$e_{p\beta_i}^2$ = possible evidence for, in support of, β_i , provided by source 2;

$(e_{d\beta i})^2$ = squared difference between demonstrated evidences for βi , provided by source 1 and source 2;
 $(e_{p\beta i})^2$ = squared difference between possible evidences for βi , provided by source 1 and source 2;

In a plane, $IM_{\beta i}$ represents the straight-line distance between two points $(e_{d\beta i}^1, e_{p\beta i}^1)$ and $(e_{d\beta i}^2, e_{p\beta i}^2)$ and is called the Euclidean distance (Kaufman and Rousseeuw, 1990). In figure 5.2, evidence pairs coming from different sources are depicted in a plane. For each belief βi , demonstrated evidence is given on the horizontal axis while possible evidence is given on the vertical axis. Note that evidence pairs always fall within the area that is marked by the small spots in grey colour, at or above the diagonal. The area under the diagonal is marked by a rectangular pattern and specifies a range where evidence pairs cannot occur.

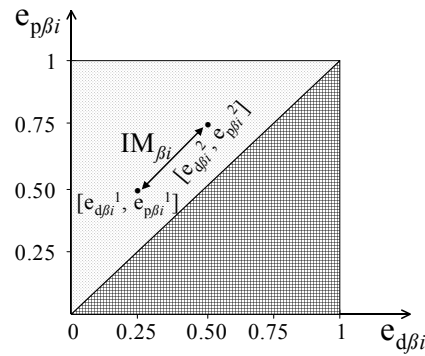


Figure 5.2: Frame of evidence presented in a plane.

Finally, for each belief βi , an interestingness measure $IM_{\beta i}$ is defined in the support logic framework by real-values. Practically this means that an ordering among interesting beliefs may be presented. Likewise, $IM_{\beta i}$'s with positive differences between demonstrated evidences for βi ($e_{d\beta i}^1 > e_{d\beta i}^2$) can be distinguished from $IM_{\beta i}$'s with negative differences between demonstrated evidences for βi ($e_{d\beta i}^1 < e_{d\beta i}^2$). As such, two groups of $IM_{\beta i}$'s may be discovered describing different situations.

5.4.2 Beliefs for Web Usage Mining

Within Web Usage Mining studies, beliefs along with their evidence pairs are automatically generated from two different sources. *Structure data* provide information about links between pages, which is incorporated into beliefs of

related pages. The stronger the topological connection between pages, the higher the value of $e_{d\beta i}^1$. *Usage data* provide information of visited pages on a web site, logged in a file and processed into server sessions. Likewise, this information is used to automatically construct beliefs and evidence pairs for pages being related by means of frequent item sets. The stronger the frequent item set, the higher the value of $e_{d\beta i}^2$. More information about structure, usage data and server sessions is given in chapter two. Examples of how evidence pairs are calculated from structure and usage data are given in the following section.

5.5 Filtering knowledge based on interestingness in Web Usage Mining

In order to define which beliefs are interesting and which are not, we will use the two different sources of structure data and usage data, providing for each belief βi , *structure evidence* [$e_{d\beta i}^s, e_{p\beta i}^s$] and *usage evidence* [$e_{d\beta i}^u, e_{p\beta i}^u$] of pages being related on a web site. A belief βi is interesting if the difference between its structure and usage evidence pairs $\geq \tau$ or if the difference between its structure (usage) and combined evidence pairs $\geq \tau$. We may also say that a belief βi is interesting if $IM_{\beta i} \geq \tau$, following equation (5.4).

5.5.1 Calculating structure evidence

In Cooley et al (1999b), a method for automatically calculating structure evidence pairs for beliefs of related web pages is given. Two factors define $e_{d\beta i}^s$. The *link factor* (lfactor) is a normalized measure for the number of links present among the pages of an item set. The *connectivity factor* (cfactor) is a measure for the strength of the topological connection among the pages in an item set. Structure evidence for a belief βi is defined as follows:

$$e_{d\beta i}^s = \text{lfactor} \times \text{cfactor} \quad (5.5)$$

where

lfactor = $L / [P (P-1)]$;

P = total number of pages in the item set;

L = number of direct hyperlinks between the pages in the item set;

cfactor = 1 if the graphical presentation for the pages in the item set is connected, which means that minimum one direct hyperlink must exist between every pair of pages in the item set;

cfactor = 0 otherwise;

$e_{p\beta_i^s}$ may be set anywhere between $e_{d\beta_i^s}$ and 1, depending on the degree of lack of evidence (5.6)

To illustrate how the cfactor is defined, six examples are given in figure 5.3. In the first column, at least one direct hyperlink exists between every pair of pages in item sets (X, Y), (X, Y, Z) and (W, X, Y, Z) and a cfactor equal to 1 is given to the item sets. In the second column, the graphical presentation for the pages in item sets (X, Y), (X, Y, Z) and (W, X, Y, Z) are not connected, which means that a cfactor of zero is given to the item sets.

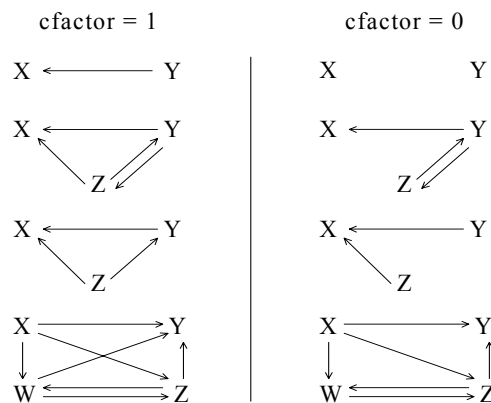


Figure 5.3: Illustration of direct hyperlinks between web pages defining the connectivity factor.

We remark that, within this research project, cfactors are either equal to one or to zero. This implies the following situations. A cfactor equal to zero for a four-item set indicates that the topological connection among these pages is not strong enough to support the belief that all of these four pages are structurally related. Yet, the cfactor for a three item set, composed with pages of the four item set, equal to one indicates that the topological connection among these pages is strong enough to support the belief that all of these three pages are structurally related. Instead of using two values for the cfactor, indicating whether the graphical presentation for the items in the item set is connected or not, it might be a good idea to use a range of values for the cfactor (i.e. $0 \leq \text{cfactor} \leq 1$), indicating how strong the graphical presentation for the items in the item set is connected. This topic is suggested in future research. The reason why we do not yet consider ranges of values for the cfactor is because first we would like to examine whether applying the basic definition of the cfactor,

developed and illustrated for Web Usage Mining studies by Cooley et al (1999b), provides good results. If so, we would be able to compare good results based on a relatively simple cfactor with ‘better’ results based on a relatively complicated cfactor. This makes it possible to test how sensitive the results are with regard to different scales for the cfactor. We may also provide a trade off between the results that are obtained from the analysis and the amount of work that was needed to perform the analysis.

Another remark is that, following equation (5.5), if cfactor $\neq 0$ then also $e_{d\beta i}^s \neq 0$, indicating that structure evidence will surely exist. Furthermore, following equation (5.5), if cfactor $\neq 0$ and $P = 2$ then $e_{d\beta i}^s = 1$ ($L = 2$) or 0.5 ($L = 1$). If cfactor $\neq 0$ and $P \geq 3$ then $0.5 \leq e_{d\beta i}^s \leq 1$.

To illustrate the calculation of structure evidence in Web Usage Mining, consider item set (W, X, Y, Z) presented in the first column of figure 6.3. Because at least one direct hyperlink between every pair of pages in the item set exists, cfactor = 1. Lfactor is defined by $L = 7$ and $P = 4$. Finally, structure evidence for belief $\beta(W, X, Y, Z)$ is defined as follows:

$$e_{d\beta(W, X, Y, Z)}^s = 7/12 \times 1 = 0.58 \text{ and } 0.58 \leq e_{p\beta(W, X, Y, Z)}^s \leq 1.$$

5.5.2 Calculating usage evidence

In Cooley et al (1999b), mined results from server session analyses, in the form of *frequent item sets*, representing frequently visited pages, are used to provide usage evidence for pages being related. Two measures are calculated for frequent item sets. *Support* (s) calculates the fraction of transactions that contain all of the items in the item set while *coverage* (c) measures the fraction of transactions that contain at least one of the items in the item set.

$$s = \text{count} (i_1 \wedge i_2 \dots \wedge i_p) / N \quad (5.7)$$

$$c = \text{count} (i_1 \vee i_2 \dots \vee i_p) / N \quad (5.8)$$

where

count (predicate) is the number of transactions containing the predicate;

i is a web page in the item set;

P is the total number of pages in the item set;

N is the total number of transactions or server sessions;

Note that support and coverage are both highly dependent on the total number of transactions. By taking the ratio of *support-to-coverage* (SCR), this dependency is eliminated. Besides, SCR gives a single measure of the strength

of a frequent item set independent of the total number of transactions in the data set. Finally, $e_{d\beta_i}^u$ is calculated as follows:

$$e_{d\beta_i}^u = SCR \quad (5.9)$$

where

$$SCR = s / c;$$

$e_{p\beta_i}^u$ may be set anywhere between $e_{d\beta_i}^u$ and 1, depending on the degree of lack of evidence (5.10)

To illustrate the calculation of usage evidence in Web Usage Mining, consider pages W, X, Y and Z. Suppose that 3 out of a total of 1000 server sessions hold page W as well as pages X, Y and Z, irrelevant in which order. Suppose that 300 out of a total of 1000 server sessions hold page W or pages X, Y or Z, irrelevant in which order. Usage evidence for belief $\beta(W, X, Y, Z)$ is defined as follows:

$$e_{d\beta(W, X, Y, Z)}^u = 3/300 = 0.01 \text{ and } 0.01 \leq e_{p\beta(W, X, Y, Z)}^u \leq 1.$$

5.5.3 Lack of evidence in Web Usage Mining

The conceptual frame of evidence is designed for a variety of applications in different research areas. This means that demonstrated, possible as well as lack of evidence (i.e. possible evidence – demonstrated evidence re. figure 5.1) may have different meanings in different applications. For example, in marketing research, loyalty analysis studies may define demonstrated and possible evidence of customer c to product p as follows. It might be useful to calculate demonstrated evidence from sequential sales (i.e. repeating sales of p to c within one year) and possible evidence from the results of annual household studies regarding distribution and spending of income. If an average household in Belgium spends yearly 0.5% of its income to product p, possible evidence may be equalized to that amount. This means that ‘lack of evidence’ may be used as a scale to indicate how loyal customer c actually is to product p. A relatively high degree of lack of evidence means that customer c is not loyal to product p or that demonstrated and possible evidence are scaled far away from each other in the conceptual frame of evidence. In other words, demonstrated evidence is plotted at the lower end of the scale, near zero, while possible evidence is plotted at the upper end of the scale, near one. A relatively low degree of lack of evidence means that customer c is loyal to product p or that demonstrated and possible evidence are scaled near each other in the conceptual frame of evidence. In other words, demonstrated as well as possible

evidence are both plotted at the lower end, both somewhere in the middle or both at the upper end of the scale. Yet, some standardization procedure is necessary in order to assure that possible evidence will always be equal to or larger than demonstrated evidence.

Considering ‘lack of evidence’ in Web Usage Mining studies, a meaningful application must be given to demonstrated and possible evidence. Looking for proof for beliefs of related web pages, server sessions in log files as well as direct hyperlinks between web pages in the structure of the web site provide information to define demonstrated usage and structure evidence. Yet, a meaningful way for defining possible usage and structure evidence for beliefs of related web pages in Web Usage Mining studies is hard to find. For this reason, with regard to beliefs of related web pages, possible evidence is equalized with demonstrated evidence in Web Usage Mining studies (Cooley et al, 1999b). Practically this means that, for the remaining of this chapter, lack of evidence will always be zero indicating that demonstrated and possible evidence are plotted at the same point in the scale of the conceptual frame of evidence.

5.5.4 *Combining structure and usage evidence*

Cooley et al (1999b) noticed the problem of scaling when combining structure and usage evidence into the support logic framework. Since the two sets of evidence are derived in different manners from different data sets, the scales do not necessarily match. For example, for the usage data, a factor that has not been considered in the generation of usage evidence is that the average mean path length of a server session equals about three pages (Pitkow, 1998). Furthermore, the distribution of the path length of server sessions fits the inverse Gaussian distribution and is heavily tailed to the right (Pitkow, 1998). This means that, if the number of related pages in a belief increases, the less likely it is that a corresponding frequent item set will be discovered. To deal with this, Cooley et al (1999b) scales usage evidence based on the number of pages in the item set:

$$e_{d\beta_i}^u = SCR \times sfactor \quad (5.11)$$

where

sfactor = number of pages in the item set;

In our illustration for belief $\beta(W, X, Y, Z)$, given in the previous examples, usage evidence is scaled with sfactor = 4 and becomes:

$$e_{d\beta(W, X, Y, Z)}^u = 0.01 \times 4 = 0.04 \text{ and } 0.04 \leq e_{p\beta(W, X, Y, Z)}^u \leq 1.$$

Suppose no lack of evidence is tolerated, then $e_{dB(W, X, Y, Z)}^s = e_{pB(W, X, Y, Z)}^s$ and $e_{dB(W, X, Y, Z)}^u = e_{pB(W, X, Y, Z)}^u$. Using equations (5.1) to (5.3), the combined evidence pair for $\beta(W, X, Y, Z)$, following from structure evidence [0.58; 0.58] and usage evidence [0.04; 0.04] with lack of evidence = 0, equals:
 $e_{dB(W, X, Y, Z)}^c = e_{pB(W, X, Y, Z)}^c = 0.05$.

5.5.5 Interesting frequent item sets

The algorithm presented in figure 5.4 is used for automatically discovering *interesting* frequent item sets. Usage evidence and structure evidence are combined using Baldwin's rules described in equations (5.1) to (5.3). Then, for each belief, three types of comparisons are made in order to filter interesting beliefs, represented by interesting item sets. First, usage evidence is compared with structure evidence. Second, usage evidence is compared with combined evidence. Third, structure evidence is compared with combined evidence. As stated in section 5.4.1, for each type of comparison, $IM_{\beta i}$'s with positive differences between demonstrated evidences for βi are distinguished from $IM_{\beta i}$'s with negative differences between demonstrated evidences for βi in order to investigate whether these groups describe different situations. Note that evidence pairs for usage and structure evidence are used, which do not take any degree for lack of evidence into account.

We illustrate the algorithm for automatically discovering interesting frequent item sets by means of our example given in the previous sections. Regarding belief $\beta(W, X, Y, Z)$, usage, structure and combined evidence are given in the second, third and fourth column of table 5.2. A value of 0.75 is given to τ for filtering interesting item sets. This means that beliefs of the highest interest levels are identified since studies showed that τ equal to 0.5 provides acceptable levels of interesting frequent item sets (Cooley et al, 1999b). Interestingness measures of at least 0.75, resulting from three different types of comparisons, along with indications for positive (+) and negative (-) differences, are given in the last three columns. This means that, in our example, we may consider belief $\beta(W, X, Y, Z)$ as an interesting frequent item set by comparing usage with structure evidence and structure with combined evidence. The difference between usage and structure evidence is negative; the difference between structure and combined evidence is positive. Following the results of this example we may say that pages W, X, Y and Z are used together less than would be expected from the structure of the web site, since there is a strong topological connection between web pages W, X, Y and Z with relatively less visiting behaviour.

Belief	Evidence			$IM_{\beta(W, X, Y, Z)} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
$\beta(W, X, Y, Z)$	[0.04 ; 0.04]	[0.58 ; 0.58]	[0.05 ; 0.05]	0.76 (-)	-	0.75 (+)

Table 5.2: Discovering interesting frequent item sets with regard to belief $\beta(W, X, Y, Z)$.

for each discovered frequent item set, representing belief βi , do

```

begin
 $e_{d\beta i}^u = e_{p\beta i}^u = SCR(\beta i) \times sfactor(\beta i);$  //calculate usage evidence pair for  $\beta i$ //
 $e_{d\beta i}^s = e_{p\beta i}^s = lfactor(\beta i) \times cfactor(\beta i);$  //calculate structure evidence pair for  $\beta i$ //
 $[e_{d\beta i}^c, e_{p\beta i}^c] = \text{Baldwin's combined evidence pair};$  //calculate combined evidence pair for  $\beta i$ //

 $e_{d\beta i} = |e_{d\beta i}^u - e_{d\beta i}^s|$  and  $e_{p\beta i} = |e_{p\beta i}^u - e_{p\beta i}^s|;$  //compare usage with structure evidence//
if  $T \leq \sqrt{[(e_{d\beta i})^2 + (e_{p\beta i})^2]}$  then
  begin
    add  $\beta i$  to interesting frequent item set by means of comparing usage with structure evidence;
    if  $e_{d\beta i}^u > e_{d\beta i}^s$  then  $positive\_difference[e_{d\beta i}^u, e_{d\beta i}^s] = true$ 
    else  $positive\_difference[e_{d\beta i}^u, e_{d\beta i}^s] = false;$ 
  end;

   $e_{d\beta i} = |e_{d\beta i}^u - e_{d\beta i}^c|$  and  $e_{p\beta i} = |e_{p\beta i}^u - e_{p\beta i}^c|;$  //compare usage with combined evidence//
  if  $T \leq \sqrt{[(e_{d\beta i})^2 + (e_{p\beta i})^2]}$  then
    begin
      add  $\beta i$  to interesting frequent item set by means of comparing usage with combined evidence;
      if  $e_{d\beta i}^u > e_{d\beta i}^c$  then  $positive\_difference[e_{d\beta i}^u, e_{d\beta i}^c] = true$ 
      else  $positive\_difference[e_{d\beta i}^u, e_{d\beta i}^c] = false;$ 
    end;

     $e_{d\beta i} = |e_{d\beta i}^s - e_{d\beta i}^c|$  and  $e_{p\beta i} = |e_{p\beta i}^s - e_{p\beta i}^c|;$  //compare structure with combined evidence//
    if  $T \leq \sqrt{[(e_{d\beta i})^2 + (e_{p\beta i})^2]}$  then
      begin
        add  $\beta i$  to interesting frequent item set by means of comparing structure with combined
        evidence;
        if  $e_{d\beta i}^s > e_{d\beta i}^c$  then  $positive\_difference[e_{d\beta i}^s, e_{d\beta i}^c] = true$ 
        else  $positive\_difference[e_{d\beta i}^s, e_{d\beta i}^c] = false;$ 
      end;
    end;
  end;

```

Figure 5.4: Algorithm for discovering interesting frequent item sets.

We remark that, the interestingness measure, based on the support logic framework and further developed for discovering interesting patterns within

Web Usage Mining (Cooley et al, 1999b), is not suitable for defining interesting frequent item sets consisting of one page. The reasons are the following. First, demonstrated structure evidence will always be zero because there are no direct hyperlinks with other pages to consider. Second, demonstrated usage evidence will always be one due to equal support and coverage. This results in frequent item sets that are always interesting, no matter where the page is located in the web site structure, no matter how often these pages are visited. Further research will discuss an interestingness measure for frequent item sets of one page. Nevertheless, we integrate the interestingness measure developed by Cooley et al (1999b) into SAM in order to investigate the order of visited pages within interesting visiting patterns. Our goal is to investigate whether the structure of direct hyperlinks between web pages may be improved and therefore we need interesting frequent item sets of minimum two pages.

Finally, table 5.3a and 5.3b present, by means of a decision table, the outcome if $IM \geq \tau$ for each type of comparison. Distinctions are made with regard to structure or usage evidence (not) equal to 0, 0.5 or 1. Also, no lack of evidence is tolerated, which means that the evidence shown from one of the sources defines demonstrated as well as possible evidence for a given belief. In other words, within Web Usage Mining studies (Cooley et al, 1999b) and in the remaining sections of this chapter, $e_{d\beta i}^u = e_{p\beta i}^u$ and $e_{d\beta i}^s = e_{p\beta i}^s$. Future research discusses how lack of evidence might be used and interpreted for Web Usage Mining studies. In general, the decision table provides the following information:

- If usage as well as structure evidence are different from 0, 0.5 and 1, comparing usage evidence with structure evidence identifies *interesting beliefs with conflicting evidence*. The outcome of the comparison does not indicate the strength of usage and structure evidence.
- If usage as well as structure evidence are different from 0, 0.5 and 1, comparing usage evidence with combined evidence identifies *interesting beliefs with strong usage and weak structure evidence*. The outcome of the comparison provides an indication of the strength of usage and structure evidence.
- If usage as well as structure evidence are different from 0, 0.5 and 1, comparing structure evidence with combined evidence identifies *interesting beliefs with weak usage and strong structure evidence*. The outcome of the comparison provides an indication of the strength of usage and structure evidence.

- If usage and/or structure evidence are equal to 0, 0.5 or 1, different types of comparisons identify *interesting beliefs* with ‘no’ (evidence equal to 0), ‘some’ (evidence equal to 0.5) or ‘strong’ (evidence > 0.5) usage or structure evidence. For some cases, an interestingness measure equal to 0 (IM = 0) is shown, which means that no interesting beliefs are identified.

In Web Usage Mining studies, decision table 5.3a and 5.3b may be used to predict the outcome of different comparison types when no lack of evidence is tolerated. For example, consider belief β (W, X, Y, Z) presented in table 5.2, with usage and structure evidence different from 0, 0.5 and 1. If we directly compare the original evidences i.e. usage with structure evidence providing an $IM \geq \tau$, belief β (W, X, Y, Z) will be declared interesting with conflicting evidence. If we compare combined with usage evidence, $IM < \tau$, which means that belief β (W, X, Y, Z) is not declared interesting for this type of comparison. Yet, if we compare structure with combined evidence, $IM \geq \tau$, which means that belief β (W, X, Y, Z) is identified as interesting with strong structure and weak usage evidence. Suppose that, instead of $e_{d\beta i}^u = e_{p\beta i}^u = 0.04$ and $e_{d\beta i}^s = e_{p\beta i}^s = 0.58$, evidence measures of $e_{d\beta i}^u = e_{p\beta i}^u = 0.34$ and $e_{d\beta i}^s = e_{p\beta i}^s = 1$ are used. Then, table 5.3b may be used to predict the outcome of different types of comparisons. Comparing usage with structure evidence will provide an interesting belief with conflicting and strong structure evidence since $IM \geq \tau$. Comparing usage with combined evidence will provide, since $IM \geq \tau$, an interesting belief with strong structure evidence. Comparing structure with combined evidence will not provide an interesting belief since $IM = 0$.

Sources of evidence / type of comparison	<i>Structure evidence $\neq 0$ and $\neq 0.5$ and $\neq 1$</i>			<i>Structure evidence = 0</i>		
	Usage vs structure	Usage vs combined	Structure vs combined	Usage vs structure	Usage vs combined	Structure vs combined
<i>Usage evidence $\neq 0$ and $\neq 0.5$ and $\neq 1$</i>	Interesting beliefs with conflicting evidence	Interesting beliefs with strong usage and weak structure evidence	Interesting beliefs with strong structure and weak usage evidence	Interesting beliefs with no structure evidence	Interesting beliefs with no structure evidence	(IM = 0)
<i>Usage evidence = 0</i>	Interesting beliefs with no usage evidence	(IM = 0)	Interesting beliefs with no usage evidence	(IM = 0)	(IM = 0)	(IM = 0)
<i>Usage evidence = 0.5</i>	Interesting beliefs with some usage evidence	Interesting beliefs with some usage evidence	(IM = 0)	Interesting beliefs with some usage and no structure evidence	Interesting beliefs with some usage and no structure evidence	(IM = 0)
<i>Usage evidence = 1</i>	Interesting beliefs with strong usage evidence	(IM = 0)	Interesting beliefs with strong usage evidence	Interesting beliefs with strong usage and no structure evidence	Interesting beliefs with strong usage and no structure evidence	(IM = 0)

Table 5.3a: Decision table presenting the outcome for $IM \geq \tau$ specified for different types of comparisons and evidence.

Sources of evidence / type of comparison	<i>Structure evidence = 0.5</i>			<i>Structure evidence = 1</i>		
	Usage vs structure	Usage vs combined	Structure vs combined	Usage vs structure	Usage vs combined	Structure vs combined
<i>Usage evidence ≠ 0 and ≠ 0.5 and ≠ 1</i>	Interesting beliefs with conflicting and some structure evidence	(IM = 0)	Interesting beliefs with some structure evidence	Interesting beliefs with conflicting and strong structure evidence	Interesting beliefs with strong structure evidence	(IM = 0)
<i>Usage evidence = 0</i>	Interesting beliefs with some structure and no usage evidence	(IM = 0)	Interesting beliefs with some structure and no usage evidence	Interesting beliefs with strong structure and no usage evidence	(IM = 0)	Interesting beliefs with strong structure and no usage evidence
<i>Usage evidence = 0.5</i>	(IM = 0)	(IM = 0)	(IM = 0)	Interesting beliefs with strong structure and some usage evidence	Interesting beliefs with strong structure and some usage evidence	(IM = 0)
<i>Usage evidence = 1</i>	Interesting beliefs with strong usage and some structure evidence	(IM = 0)	Interesting beliefs with strong usage and some structure evidence	(IM = 0)	(IM = 0)	(IM = 0)

Table 5.3b: Decision table presenting the outcome for $IM \geq \tau$ specified for different types of comparisons and evidence.

5.6 Integrating SAM with support logic framework and evidence combination

The interesting frequent item sets, representing interesting related web pages, provided by the algorithm of support logic framework and evidence combination, are integrated into the SAM algorithm. Hence, interesting visiting patterns, including order-based information, are automatically discovered. The advantage of integrating SAM with support logic framework and evidence combination is that, instead of presenting a general view of both interesting and uninteresting visiting patterns, now only the patterns that are interesting are given, including order-based information. From now on we will call SAM^I the SAM algorithm for discovering interesting visiting patterns. Practically this means that, when measuring distance between sequences by means of SAM^I, only the elements that represent interesting related pages are considered, instead of considering every element. SAM^I automatically filters out interesting pages (or combinations of pages) during the equalization process for sequence comparison. At the end of the process of sequence comparison, SAM^I identifies the interesting combination of pages within each sequence along with the distance measures. Sequences without any interesting combination of pages are not processed by SAM^I.

In particular, SAM^I distance between two sequences $S_1 = s_{11}, s_{12}, \dots, s_{1m}$ and $S_2 = s_{21}, s_{22}, \dots, s_{2n}$ is calculated using equation (5.12), which is deduced from equation (3.1) in chapter three. We remark that operations are performed on interesting pages only. We also note that substitution in equation (3.1) is replaced by reordering in equation (5.12) since *reordering* represents one deletion and one insertion of the same element affecting the same sequence (Joh et al, 2001).

$$d_{\text{SAM}^I}(S_1, S_2) = \min [(w_d D^I + w_i I^I) + \eta R^I] \quad (5.12)$$

where

d_{SAM^I} is the similarity or distance for interesting pages between two sequences S_1 and S_2 , based on SAM;

w_d is the weight value for the deletion operations, a positive constant not equal to 0, determined by the researcher ($w_d > 0$);

w_i is the weight value for the insertion operations, a positive constant not equal to 0, determined by the researcher ($w_i > 0$);

D^I is the number of deletion operations for interesting pages;

I^I is the number of insertion operations for interesting pages;

R^I is the number of reordering operations for interesting pages;

η is the reordering weight, a positive constant not equal to 0, determined by the researcher ($\eta > 0$);

and

m is the length of the first sequence (source);

n is the length of the second sequence (target);

s_{ij} is an element, representing a particular character, of a sequence;

i identifies the sequence number, $i = 1, \dots, N$;

N is the total number of sequences in the analysis;

j identifies the position in a sequence, $j = 1, \dots, m$ or $j = 1, \dots, n$;

Equation (5.12) indicates that the score, represented by SAM^I between two sequences, consists of the minimum costs for deleting and inserting unique interesting elements and the minimum costs for reordering common interesting elements.

In figure 5.5, the procedure is given that SAM^I uses for transforming original server sessions into sessions holding only interesting combinations of pages with respect of the order of pages. The source code of the program is given in appendix five. The algorithm reads the input file 'server_sessions_original' and checks, for every original server session, whether it holds interesting frequent item sets of two, three or four pages. If yes, the frequent item sets are written in an array. Then, for each original server session, starting with the first page up to the last page, this array is used to check whether the page is an element of an interesting frequent item set that was previously found in the original server session. If the page can be read from the array it is written in the output file 'server_sessions_transformed'.

```
begin
read original server session;
for  $i:=1$  to  $n$  do           //n = total number of interesting frequent item sets//
  begin
    read IFIS;                 //IFIS = interesting frequent item set//
    if IFIS  $\in$  original server session then write IFIS in array;
    readln;
  end
  ...
  for  $j:=1$  to  $m$  do         //m = length of server session or total number of elements=
    begin                       pages in server session//
      if  $j \in$  array then write  $j$  in server_sessions_transformed;
    end;
  end;
```

Figure 5.5: Procedure, used by SAM^I , for transforming server sessions into sessions with interesting combinations of pages, respecting the order of pages.

To give a clear understanding of how SAM^l works, the algorithm given in figure 5.5 is illustrated with an example in table 5.4. The interesting frequent item sets, discovered by the support logic framework and evidence combination, described in sections 5.4 and 5.5, are given in the first column. Note that alphanumerical characters are changed in integer values so as to apply the algorithm presented in figure 5.5. We remind that the order in which elements occur in frequent item sets is irrelevant. The second column presents two sequences s_1 and s_2 representing server sessions holding interesting and uninteresting combinations of pages. In the third column SAM^l between s_1 and s_2 is presented. Finally, in the last column, the original source and target sequences s_1 and s_2 are changed into sequences holding only interesting combinations of pages, respecting the order in which pages occur. Combinations of pages that are not interesting are filtered out of the sequences.

Interesting frequent item sets / interesting related pages	Source sequence: $s_1 = Y, T, U, X$ $= 2, 6, 7, 1$	$w_i = 1$ $w_d = 1$ $\eta = 2$	Source sequence (interesting comb. of pages): $s_1 = Y, X$ $= 2, 1$
$(X, Y) = (1, 2)$	Target sequence: $s_2 = X, Z, X, W, Y$ $= 1, 3, 1, 5, 2$	$d_{\text{SAM}^l}(s_1, s_2) = 5$	Target sequence (interesting comb. of pages): $s_2 = X, Z, X, W, Y$ $= 1, 3, 1, 5, 2$
$(V, W) = (4, 5)$			
$(Y, W) = (2, 5)$			
$(X, Y, Z) = (1, 2, 3)$			

Table 5.4. Sequence comparison based on SAM^l.

5.7 Application

For this application, log files registering visiting behaviour from 01/02/1999 till 28/02/1999 on the web site <http://machines.hyperreal.org> are analysed. After pre-processing the data using the method described in chapter four, section 4.3.1 Step 1: Pre-processing, a total number of 75,855 server sessions, navigating through web pages with 1,159 different logged URL addresses, are identified. Each URL address refers to a web page. For convenience of presentation, a unique page identification number is given to each distinct URL address. For example, page 349 is given to URL address <http://machines.hyperreal.org/manufacturers>. We remark that in this chapter the same web site is analysed as in chapter four (data set 2). Yet, different registration periods of the logged data are used. Compared to data set 2 of chapter four, this chapter analyses more data, providing more server sessions, which are analysed by SAM¹. If we use only the 3,131 server sessions of data set 2 in chapter four, we would end up with barely 180 server sessions holding interesting frequently visited web pages through SAM¹ analysis. This is not a realistic experimental analysis for testing SAM and Interestingness within Web Mining and Data Mining research. For this reason, we used more data and started the analysis of SAM¹ with more server sessions.

In the first section, statistics about the data used in this application are given. This is followed by the first step of the analysis, where interesting frequently visited pages are defined. Then, SAM¹ similarity measures for interesting frequently visited pages are calculated between the server sessions. Based on these similarity measures, the server sessions holding interesting related pages are clustered. Finally, in section 5.7.4, the clusters are examined.

5.7.1 Describing the data

Table 5.5 describes the data that is used for discovering interesting visiting patterns, providing order based information, on the web site <http://machines.hyperreal.org>. The shortest server session consists of one page whereas the longest server session is twenty pages long. The average length of the server sessions is 2.6. The total number of page views in the file is 196,550 with a total number of distinct URL addresses equal to 1,159.

Statistics	Server sessions representing visiting behaviour from 01/02/1999 – 28/02/1999 on the web site http://machines.hyperreal.org .
Total	75,855
Shortest	1
Longest	20
Average length	2.59
Total number of requests	196,550
Distinct pages	1,159

Table 5.5: Describing server sessions used within the SAM^I application.

Figure 5.6 provides an overview of the distribution of the server sessions' length. On the horizontal axis, the length of the server sessions ranging from one to twenty web pages is given; on the vertical axis the relative frequency (number of occurrences of the corresponding sessions' length divided by the total number of sessions in the analysis, multiplied by 100) is given. The three highest relative frequency values from the graph show that 54.45% of the server sessions in the analysis consist of one page, 16.40% consist of two pages and 9.33% consist of three pages. Finally, 16.45% of the server sessions in the analysis are between four and ten pages long and 3.37% of the server sessions in the analysis are between eleven and twenty pages long.

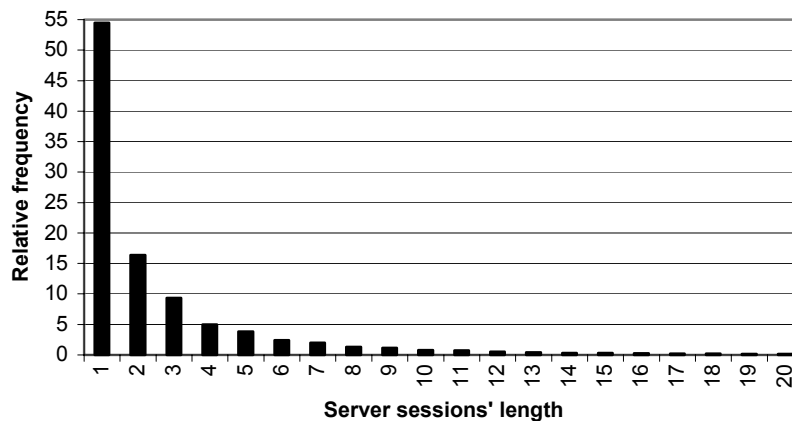


Figure 5.6: Distribution of the length of server sessions.

Figure 5.7 gives an overview of the distribution of distinct web pages. On the horizontal axis, 1,159 distinct web pages are presented by means of 50 groups. Each group reflects 23 web pages, except for the last group. For example, group 1 reflects page 1 to page 23, group 2 reflects page 24 to 46, group 3 reflects page 47 to page 69 etc. Finally, group 50 reflects page 1,128 to page 1,159. On the vertical axis, the frequency values (number of requests of the page_ids within the corresponding group divided by the total number of requests (i.e. 196,550) in the file, multiplied by 100) are given. The graph shows that 25% of the visited pages are pages within group 29, reflecting web pages 645 to 667 (including 645 and 667). The following two highest relative frequency values are 6.75% for group 50, reflecting web pages 1128 to 1159 (including 1128 and 1159) and 6.42% for group 45, reflecting web pages 1013 to 1035 (including 1013 and 1035).

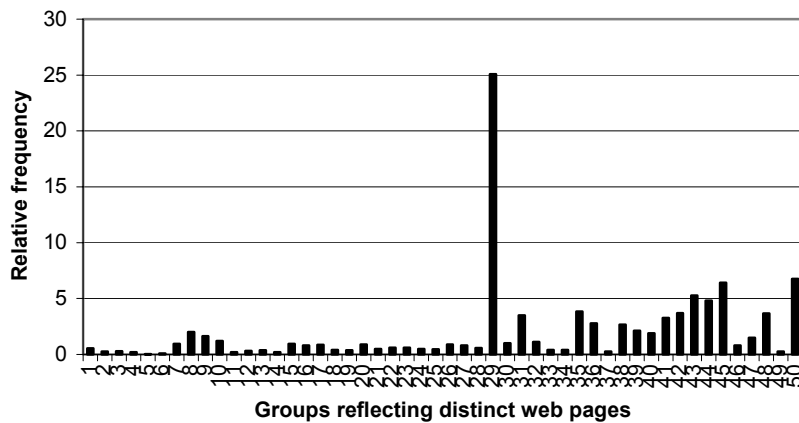


Figure 5.7: Distribution of 1159 distinct page_ids, represented in 50 groups.

Comparing figures 5.6 and 5.7, given above, with figures 4.4 (data set 2) and 4.6 in chapter four, presenting web usage behaviour on the same web site, the data sets are very similar with regard to distribution of length of server sessions and (groups of) distinct web pages. This means that the logged data that is used in chapter four (data set 2), storing visiting behaviour from 01/02/1999 to 03/02/1999, which is used to provide (un)interesting visiting patterns by means of SAM, is very similar to the logged data that is used in this chapter, storing visiting behaviour from 01/02/1999 to 28/02/1999, which is used for defining interesting visiting patterns by means of SAM¹.

5.7.2 *Interesting frequently visited pages*

Frequent item sets with minimum support of 0.1% (Cooley et al, 1999b) are calculated on the server sessions. Every frequent item set represents a belief of related pages and usage evidence pairs for pages being related are calculated from the support and coverage values of the frequent item sets using equations (5.10) and (5.11). Then, structure and combined evidence pairs are defined using equations (5.1) to (5.3), (5.5) and (5.6). Note that, as previously mentioned, no lack of evidence is tolerated in the analysis. In order to filter out interesting frequently visited pages, equation (5.4) is used along with the algorithm presented in figure 5.4.

An illustration is given in table 5.6. A total number of 539 beliefs, consisting of minimum two and maximum four related pages, are identified. They are given in the first column. For each belief, usage, structure and combined evidence pairs are presented in the following columns. An interestingness threshold value of $\tau = 0.75$ in equation (5.4) is used to filter out interesting beliefs of related pages. By setting the value of τ very high, related pages of the highest interest are discovered. Usually, a τ -value of 0.5 is satisfactory to filter out interesting from uninteresting beliefs (Cooley et al, 1999b). Three lists of interesting related pages are identified by comparing usage evidence with structure evidence (column five), usage evidence with combined evidence (column six), and structure evidence with combined evidence (column seven). Beliefs of related pages that are considered interesting are written in bold. Along with interesting beliefs, positive and negative differences between evidence pairs are written between brackets, next to the value of IM_{β_i} . Out of 539 beliefs of related pages, 91 are considered interesting. They are given in appendix five, ordered by level of interestingness and comparison manner.

Beliefs of related pages	Evidence			$IM_{\beta_i} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
$\beta_1(657, 984)$	[0.1700; 0.1700]	[0.5000; 0.5000]	[0.1700; 0.1700]	-	-	-
$\beta_2(815, 657)$	[0.1338; 0.1338]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2249 (-)	1.2249 (-)	-
$\beta_3(657, 947)$	[0.1100; 0.1100]	[0.5000; 0.5000]	[0.1100; 0.1100]	-	-	-
...
$\beta_{100}(1026, 1025)$	[0.1598; 0.1598]	[0.0000; 0.0000]	[0.0000; 0.0000]	-	-	-
...
$\beta_{200}(1129, 996)$	[0.0816; 0.0816]	[0.5000; 0.5000]	[0.0816; 0.0816]	-	-	-
...
$\beta_{300}(984, 163)$	[0.0356; 0.0356]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3638 (-)	1.3638 (-)	-
...
$\beta_{310}(62, 171)$	[0.2372; 0.2372]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0787 (-)	1.0787 (-)	-
...
$\beta_{450}(657, 1026, 713)$	[0.0420; 0.0420]	[0.0000; 0.0000]	[0.0000; 0.0000]	-	-	-
...
$\beta_{500}(815, 657, 810)$	[0.0123; 0.0123]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3968 (-)	1.3968 (-)	-
...
$\beta_{520}(657, 1026, 1025)$	[0.0192; 0.0192]	[0.0000; 0.0000]	[0.0000; 0.0000]	-	-	-
...
$\beta_{530}(657, 984, 993)$	[0.0174; 0.0174]	[0.5000; 0.5000]	[0.0174; 0.0174]	-	-	-
...
$\beta_{535}(657, 1026, 713, 868)$	[0.0084; 0.0084]	[0.0000; 0.0000]	[0.0000; 0.0000]	-	-	-
...
$\beta_{539}(815, 794, 657, 786)$	[0.0116; 0.0116]	[0.5834; 0.5834]	[0.0162; 0.0162]	0.8086 (-)	-	0.8021 (+)

Table 5.6: Interesting and uninteresting beliefs of related pages on <http://machines.hyperreal.org>.

Table 5.7 provides some statistics of beliefs of related pages that are considered interesting. A total number of 91 interesting frequent item sets, consisting of two to four pages, are found. The average length is 2.5. The total number of pages within 91 interesting frequent item sets is 229. Finally, the total number of distinct pages within 91 interesting frequent item sets is 61.

Statistics	Interesting frequent item sets
Total	91
Shortest	2
Longest	4
Average length	2.5
Total number of requests	230
Distinct pages	61

Table 5.7: Describing interesting frequent item sets used within the SAM^I application.

Figure 5.8 provides an overview of the distribution of the interesting frequent item sets' length. On the horizontal axis, the length ranging between two to four pages is given; on the vertical axis the relative frequency (number of occurrences of the frequent item sets' length divided by the total number of frequent item sets in the analysis, multiplied by 100) is given. The graph shows that 50.55% of the interesting frequent item sets consist of two pages, 47.25% consist of three pages and 2.20% consist of four pages.

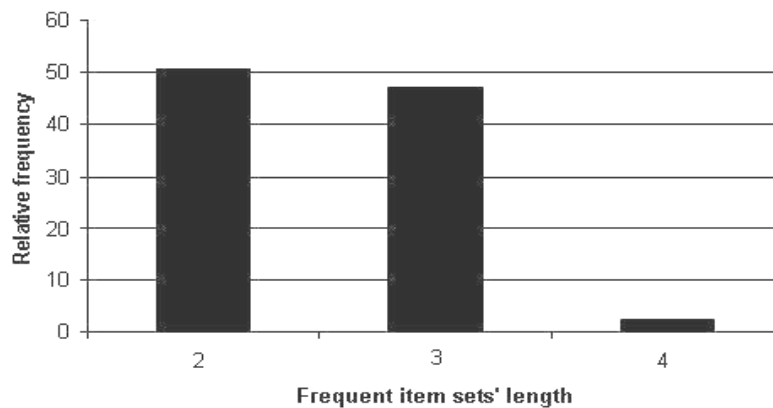


Figure 5.8: Distribution of interesting frequent item sets' length.

Figure 5.9 provides an overview of the distribution of distinct web pages within the 91 interesting frequent item sets. On the horizontal axis, 61 distinct web pages, starting with page_id 62 and ending with page_id 1,134, are presented. On the vertical axis, the frequency value for each distinct web page divided by the total number of requests (pages) in the file (i.e. 229), multiplied by 100, is given. The graph shows that 21.83% of the pages in the interesting frequent item sets are page 657. Likewise, pages 984 and 815 are highly represented in the interesting frequent items sets with respectively 19.65% and 9.17%.

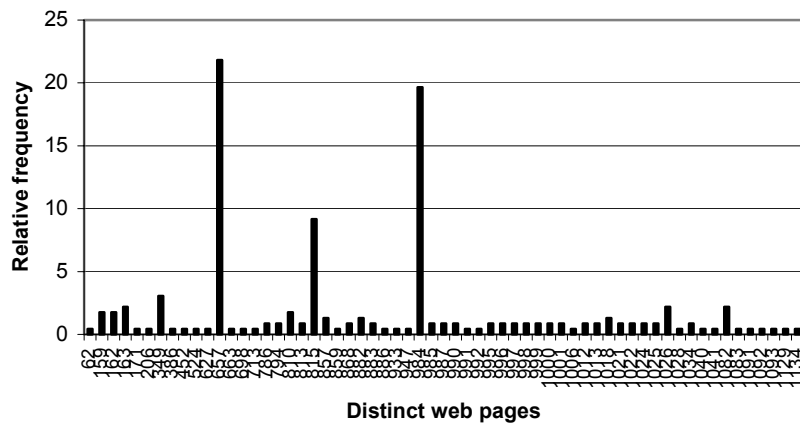


Figure 5.9: Distribution of distinct web pages in interesting frequent item sets.

Table 5.9 presents information about values of the cfactor, which is used in equation (5.5) for defining structure evidence, and of the sfactor, which is used in equation (5.11) for defining usage evidence on the data set presented in table 5.5. The table shows that, when comparing all of the beliefs with the interesting beliefs, distributions of values of the cfactor and sfactor are quite different. For example, out of the 539 beliefs of related pages that were found by frequent items sets with minimum support of 0.1%, 192 (35.62%) showed a cfactor equal to zero while 347 (64.38%) showed a cfactor equal to one. Out of the 91 beliefs of related pages that were declared interesting, all of them (100%) showed a cfactor equal to one. The distribution of the values of the sfactor among the 539 beliefs of related pages is as follows. 72.73%, 25.42% and 1.85% showed an sfactor of respectively two, three and four. Beliefs of related pages consisting of more than four pages, given minimum support of 0.1% for frequent item sets, were not found. The distribution of the values of the sfactor

among the 91 interesting beliefs of related pages is as follows. 49.45%, 48.35% and 2.20% showed an sfactor of respectively two, three and four. Some explanations for these differences are given below:

- If cfactor = 0 then structure evidence = 0 and, given $\tau = 0.75$, usage evidence must be ≥ 0.53034 in order to find an interesting belief. Beliefs with structure evidence = 0 and usage evidence ≥ 0.53034 were not found. For this reason, interesting beliefs of related pages are not found when cfactor = 0.
- Structure evidence for beliefs consisting of two pages can have only three values, which are 0, 0.5 or 1. Because of the relatively low usage evidence values along with a relatively high τ value (i.e. 0.75), all of the interesting beliefs are identified when structure evidence is 1 for two-item sets.
- Structure evidence for beliefs consisting of three pages can have five values, which are 0; 0.5; 0.67; 0.83 and 1. Because of the relatively low usage evidence values along with a relatively high τ value (i.e. 0.75), all of the interesting beliefs are identified when structure evidence ≥ 0.67 for three-item sets.
- Structure evidence for beliefs consisting of four pages can have the following values: 0; 0.33; 0.42; 0.5; 0.58; 0.67; 0.75; 0.83; 0.92 and 1. Because of the relatively low usage evidence values along with a relatively high τ value (i.e. 0.75), all of the interesting beliefs are identified when structure evidence ≥ 0.58 for four-item sets

Factor	539 beliefs of related pages		91 interesting beliefs of related pages	
	Number of beliefs	% of beliefs	Number of interesting beliefs	% of interesting beliefs
cfactor = 0	192	35.62	0	0.00
cfactor = 1	347	64.38	91	100.00
sfactor = 2	392	72.73	45	49.45
sfactor = 3	137	25.42	44	48.35
sfactor = 4	10	1.85	2	2.20

Table 5.9: Comparing cfactor and sfactor for total number of beliefs with interesting beliefs.

Finally, with regard to interesting beliefs of related pages presented by triple item sets, only one triple item set is also interesting for each combination of its dual items sets. Of the 43 remaining interesting beliefs of related pages

presented by triple item sets, none of them provides interesting dual item sets for each combination. For example, (815, 657, 810) is defined interesting along with (815, 657), (815, 810) and (657, 810). Yet, (815, 657, 813) is defined interesting along with (815, 657) and (657, 813). However, dual item set (815, 813) is not defined interesting. Considering the two interesting related pages presented by four-item sets, (815, 657, 1026, 984) shows interesting related pages for each combination of triple item sets while for (815, 794, 657, 786), triple item sets (794, 657, 786) and (815, 794, 786) are not defined interesting.

Figure 5.10 depicts how (815, 657, 810) and (815, 657, 813) are structured on <http://machines.hyperreal.org>, along with information of URL addresses for the page_ids in the item sets. On the left hand side of the figure, each combination of dual item sets is interesting because visiting behaviour between these pages occurs less frequent than expected, given the direct hyperlinks between the web pages. On the right hand side of the figure, (815, 813) is not interesting because, given only one direct hyperlink from 813 to 815, visiting behaviour between 815 and 813 occurs less frequent, which is expected. Table 5.10 provides information about usage-, structure evidence and interestingness measure.

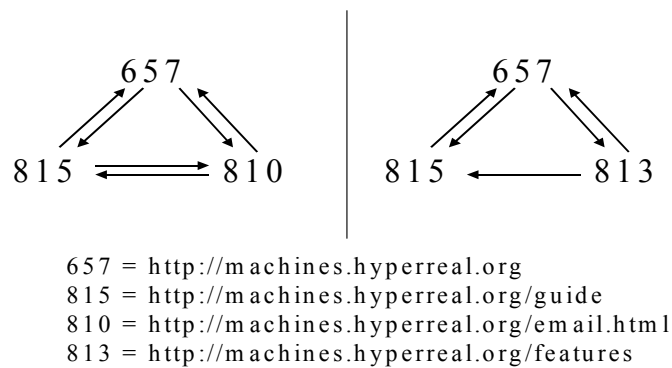


Figure 5.10: Example of interesting frequent item sets (815, 657, 810) and (815, 657, 813) along with direct hyperlinks on <http://machines.hyperreal.org>.

$r = 0.75$			
Frequent item set	Evidence		IM
	Usage	Structure	Usage – Structure
(815, 657, 810)	[0.0123; 0.0123]	[1.0000; 1.0000]	1.3968
(815, 657)	[0.1338; 0.1338]	[1.0000; 1.0000]	1.2249
(815, 810)	[0.0951; 0.0951]	[1.0000; 1.0000]	1.2797
(657, 810)	[0.0210; 0.0210]	[1.0000; 1.0000]	1.3845
(815, 657, 813)	[0.0189; 0.0189]	[0.8334; 0.8334]	1.1518
(815, 657)	[0.1338; 0.1338]	[1.0000; 1.0000]	1.2249
(657, 813)	[0.0345; 0.0345]	[1.0000; 1.0000]	1.3654
(815, 813)	[0.2466; 0.2466]	[0.5000; 0.5000]	0.3583

Table 5.10: Usage-, structure evidence and interestingness measure (IM) for interesting triple frequent item sets shown in figure 5.10 and combinations of dual item sets.

5.7.3 SAM^I distance measures

The interesting beliefs of related pages, which are also called interesting frequently visited pages or interesting frequent item sets, presented in appendix 5, are used by SAM^I to measure distances between server sessions. In this step, the algorithm described in equation (5.12) and in figure 5.5 is used. Due to the fact that SAM^I selectively aligns sequences based on interesting frequently visited pages, the original number of server sessions is reduced from 75,855 to 7,266. This means that 68,589 server sessions do not hold interesting frequently visited pages and therefore, are not considered for further analysis.

From now on, the same approach is used as shown in chapter four, steps 2 and 3: Processing and Post-processing. This means that the SAM^I distance measures for interesting frequently visited pages are used as distance measures for clustering. Figure 5.11 depicts the criteria for defining the number of clusters. The first two criteria, pseudo F statistic along with T-squared statistic, designate four clusters as a good cluster solution. Although R-squared reaches 57% at this level, the homogeneity of the data in four clusters is relatively high, indicated by a small value for the root mean squared standard deviation. We may also choose for six or eleven clusters. However, the result will end up in some small clusters holding small percentages of the data, which may be indicated by lower values for the pseudo F statistic. Since we are interested in large clusters of more or less equal sizes, we prefer four clusters.

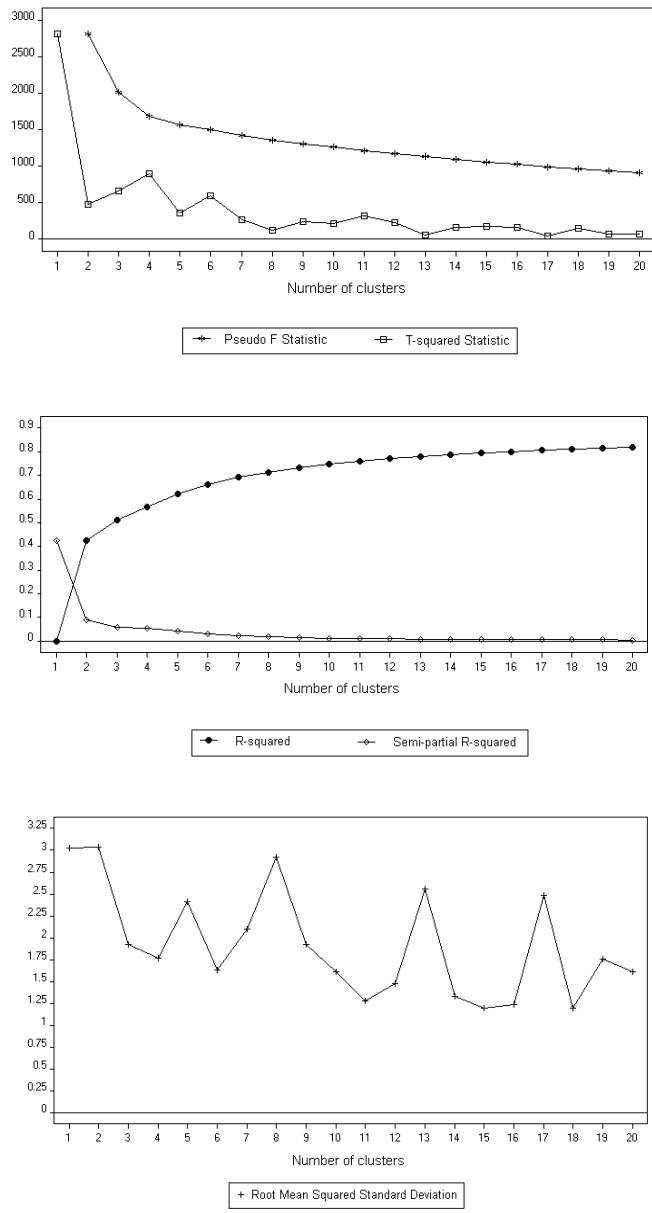


Figure 5.11: Information criteria for defining the number of clusters, using SAM¹ distance measures between server sessions at <http://machines.hyperreal.org>.

5.7.4 Cluster examination

Ward hierarchical clustering, based on SAM¹ distance matrix, results in four clusters. In this section, every cluster is examined by means of *interesting pages*, excluding order-based information, and *interesting navigations*, including order-based information of visited pages. First, some statistics are given with regard to the file holding all of the server sessions and the clusters.

5.7.4.1 Describing the data

Statistics for each cluster, as well as for the file holding the server sessions based on interesting related pages, are provided in table 5.11. A total number of 7,266 server sessions, showing 35,566 page requests, are clustered in four groups. Comparing the original server sessions given in table 5.5 with table 5.11, the total number of server sessions and total number of requests are reduced from 75,855 and 196,550 to 7,266 and 35,566 respectively. The shortest server session in the file is, instead of one, two pages long. The average length of server sessions holding interesting related web pages equals 4.9 instead of 2.59. Obviously, the total number of distinct pages in the file holding all of the server sessions based on interesting related pages is the same as in table 5.7.

Statistics	Server sessions based on interesting related pages				
	File	Cluster			
		1	2	3	4
Total	7,266	1,337	2,352	2,584	993
Shortest	2	2	2	2	2
Longest	20	13	19	20	19
Average length	4.9	3.2	5.3	5.5	4.7
Number of requests	35,566	4,280	12,458	14,199	4,629
Distinct pages	61	29	55	56	50

Table 5.11: Describing server sessions and clusters based on interesting related pages resulted from the SAM¹ application.

Figure 5.12 provides a 3D overview of the length of server sessions based on interesting related pages. Although at this stage of the analysis server sessions consisting of one page no longer exist (they are removed from the analysis because of the reasons given in section 5.5.4), for comparison reasons, the horizontal axis of the graph is the same as in figure 5.6, showing the length of the server sessions ranging from one to twenty pages. On the vertical axis, the relative frequency for each server sessions' length within each file/cluster is

presented. The relative frequency equals to the number of occurrences of the corresponding server sessions' length divided by the total number of sessions in the file or cluster, multiplied by 100. The figure shows that, with regard to the file holding all of the server sessions based on interesting related pages, 22.49% of the sessions are three pages long, 21% are two pages long and 13.71% are four pages long. The length of server sessions in cluster two, three and four show approximately the same division. Yet, cluster one provides another division with regard to sessions' length. Here, 46.97% of the server sessions are two pages long, 23.56% are three pages long and 13.31 are four pages long.

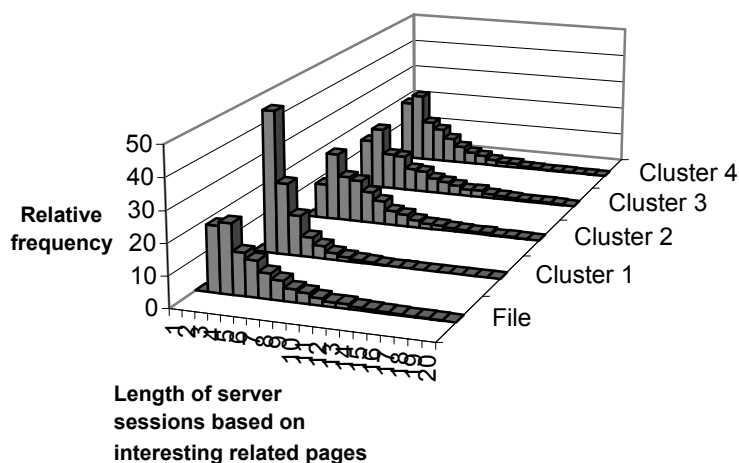


Figure 5.12: Distribution of the length of server sessions based on interesting related pages.

5.7.4.2 Cluster examination by means of interesting web pages

In figure 5.13, clusters are graphically presented with regard to *interesting web pages*. In a 3D view, the relative frequencies of interesting web pages, occurring in interesting frequent item sets (re. table 5.7), are given for the file holding all of the server sessions based on interesting related pages and for the clusters. The relative frequency equals to the number of requests (hits) of the corresponding web page divided by the total number of requests (hits) in the file or cluster, multiplied by 100. For example, the file holding all of the server sessions based on interesting related pages contains 1,334 requests of page 349.

The relative frequency of page 349 is $[1,334 / 35,566] * 100$ or 3.75%. Since page 349 shows 1,304 hits in cluster one, the relative frequency of page 349 in cluster one is $[1,304 / 4280] * 100$ or 30.46%. The following pages show relatively high frequency values in the following file or cluster:

- File holding all of the server sessions based on interesting related pages: 657 (32.28%), 984 (14.11%) and 815 (7.46%)
- Cluster 1: 349 (30.47%), 657 (12.62%), 163 (10.56%) and 159 (10.42%)
- Cluster 2: 984 (33.77%), 657 (23.68%)
- Cluster 3: 657 (49.67%), 815 (16.52%)
- Cluster 4: 657 (20.29%), 1082 (17.37%), 882 (10.09%),

The two smaller graphs below provide a rotated and elevated view of the graph above. Compared with figure 5.9, the file holding all of the server sessions based on interesting related pages shows approximately the same distribution of interesting web pages. Note that, in figure 5.13, if all of the interesting web pages were written on the horizontal axis like in figure 5.9, the graphical presentation becomes unclear. Therefore, without changing the scale of the horizontal axis, 21 instead of 61 interesting web pages, or 1 out of 3, are written down.

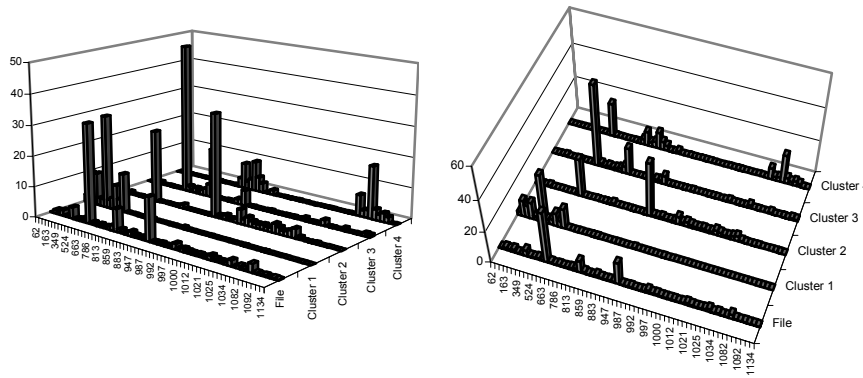
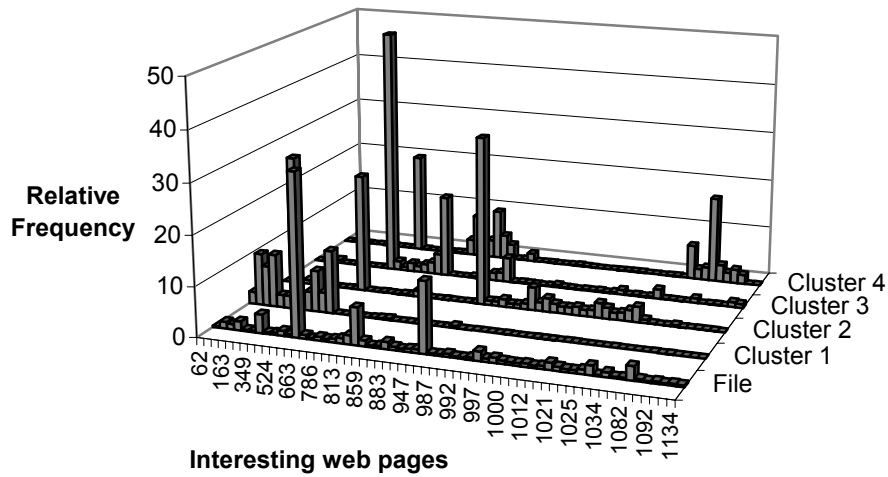


Figure 5.13: Distribution of interesting web pages.

In order to analyse the exclusivity of the clusters with regard to interesting pages, figure 5.14 provides exclusivity measures for each page within each cluster. *Exclusivity* for page x within cluster y is defined as the number of requests (hits) for page x within cluster y divided by the total number of requests (hits) for page x within the analysis, multiplied by 100. Exclusivity lies between zero and one. Exclusivity of a particular page equal to zero means that the corresponding cluster does not represent that page at all. Exclusivity of one indicates that the corresponding cluster is the only cluster that represents the page. Figure 5.14 shows that each cluster represents high exclusivities for

interesting web pages. The clusters given below represent exclusivities of at least 0.80 for 90% of the pages (or for 55 out of 61 pages):

- Cluster 1: 62, 159, 162, 171, 206, 349, 386, 452, 524, 627
- Cluster 2: 984, 985, 990, 991, 992, 995, 996, 997, 998, 999, 1000, 1001, 1006, 1012, 1021, 1022, 1024, 1025, 1028
- Cluster 3: 663, 698, 713, 786, 794, 810, 813, 815, 933, 947, 1129, 1134
- Cluster 4: 857, 859, 868, 882, 883, 886, 1034, 1040, 1041, 1082, 1083, 1091, 1092, 1093

Page 163 is mainly represented in cluster one with exclusivity of 0.69; pages 987, 1013 and 1018 are represented in cluster two with exclusivities of respectively 0.78, 0.76 and 0.72. Finally, two pages are not exclusively assigned to a cluster: page 657 (4.70% in cluster one, 25.69% in cluster two, 61.43% in cluster three, 8.18% in cluster four) and page 1026 (0.3% in cluster one, 58.32% in cluster two, 40.56% in cluster three, 1.03% in cluster four).

Within the web usage mining process of analysing visiting behaviour on <http://machines.hyperreal.org>, chapter four presented the results of exclusivities for (un)interesting web pages by means of clustering based on SAM. When comparing the results of chapter four (re. figure 4.20) with the results of exclusivities for interesting web pages by means of clustering based on SAM¹, high exclusivity measures based on SAM¹ are more evenly distributed among different clusters. Instead of one cluster providing most of the high exclusivities, now each cluster provides high exclusivities.

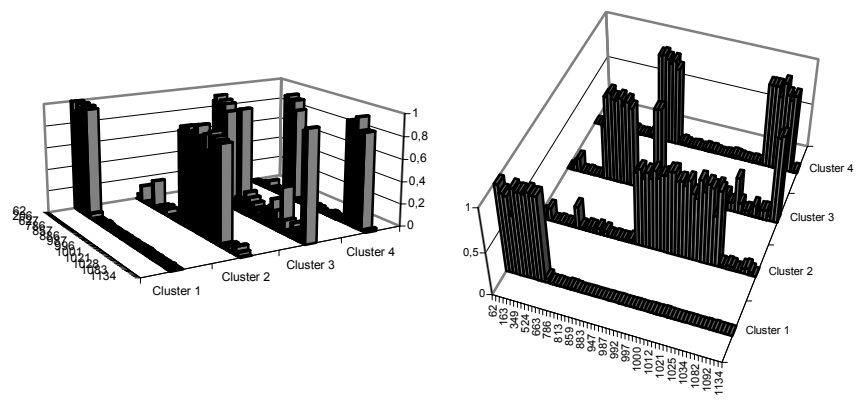
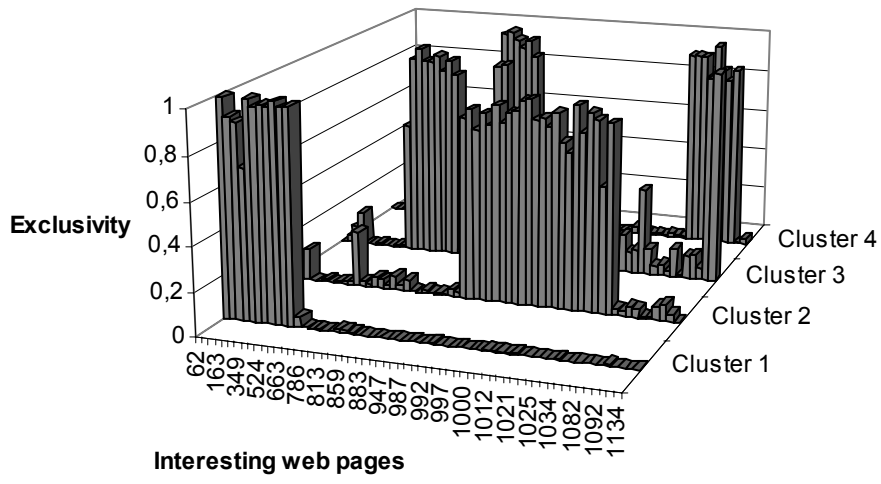


Figure 5.14: Exclusivity for interesting web pages within four clusters.

5.7.4.3 Cluster examination by means of interesting navigations

Cluster examination with regard to the order of interesting pages is done with open sequences. Open sequences are defined and illustrated in chapter four, section 4.8.2 (Step 3: Post-processing). For each cluster, all possible open sequences with minimum length of two elements and minimum support or confidence of 1% are calculated. Furthermore, for a thorough analysis and a clear view of the results, first open sequences having the *five highest support values* are selected for every cluster. The results are given in table 5.12. Cluster one generally represents interesting navigations to and from page 349. Cluster two and three represent navigations to and from page 984 and page 657 respectively. Finally, cluster four concerns visiting patterns to several web pages: 657, 857, 868, 882, 883 and 1082. Note that, with regard to open sequences with high support values, the same open sequences are not shown in different clusters. This indicates that the clusters are well separated.

To provide more order-based information for every cluster, also open sequences based on the *five highest confidence values* are given for each cluster. If more than five open sequences were found, all of them showing the same high confidence values, two additional selection criteria were applied, based on the longest open sequences and on the highest support values. For example, in cluster three, 11 open sequences provided the same highest confidence values of 100%:

- 1 (815, 1018, 657, 813, 657) Support = 1.04; Confidence = 100.00
- 2 (815, 794, 657, 813, 657) Support = 1.01; Confidence = 100.00
- 3 (815, 794, 657, 815, 657) Support = 1.16; Confidence = 100.00
- 4 (815, 984, 657, 813, 657) Support = 1.20; Confidence = 100.00
- 5 (810, 984, 657, 815, 657) Support = 1.01; Confidence = 100.00
- 6 (657, 1018, 657, 813, 657) Support = 1.12; Confidence = 100.00
- 7 (657, 794, 657, 813, 657) Support = 1.20; Confidence = 100.00
- 8 (815, 657, 794, 657, 815, 657) Support = 1.04; Confidence = 100.00
- 9 (815, 657, 984, 657, 813, 657) Support = 1.08; Confidence = 100.00
- 10 (657, 815, 810, 657, 813, 657) Support = 1.04; Confidence = 100.00
- 11 (657, 815, 984, 657, 813, 657) Support = 1.04; Confidence = 100.00

Instead of presenting all of these open sequences to show order-based information within cluster three, we first selected the longest ones. To show more than four open sequences, the fifth one is selected based on highest support. If, after the additional selection criteria, more than one sequence is found with highest support, all of them are given. Table 5.13 provides order-based information of interesting navigations within four clusters, based on the

five highest confidence values and, if necessary, additional selection criteria. For example, the following four different interesting navigations are extracted from the data:

- Cluster 1: The chance that page 349 is visited after the following pattern (respecting the order of pages in the pattern) is 66.04%:
657, 163
- Cluster 2: The chance that page 1013 is visited after the following pattern (respecting the order of pages in the pattern) is 91.67%:
984, 996, 999, 657
- Cluster 3: The chance that page 657 is visited after the following pattern (respecting the order of pages in the pattern) is 100.00%:
815, 657, 984, 657, 813
- Cluster 4: The chance that page 657 is visited after the following pattern (respecting the order of pages in the pattern) is 85.71%:
657, 1082, 1091, 1082

For evaluation purposes, support and confidence values of the open sequences selected for describing order-based information within each cluster, are also given for the other clusters in table 5.14. The support and confidence values of the open sequences used to describe clusters in table 6.12 and 6.13 are written in bold and represent the cluster that is printed at the head of the columns. For example, page 163 followed by page 349 (first row) represents cluster one. Page 657 followed by page 984 (eleventh row) represents cluster two. In general, we may state that the more zero values at the non-diagonal places (or the more zero values not printed in bold) in table 6.14, the better the model fits the data, i.e. the better open sequences printed in bold represent clusters.

One remark is that, at first sight, it might seem that open sequences (657, 984) and (984, 657) do not strongly represent cluster two, since the support values for these open sequences are more than 17% and 12% for cluster three. However, cluster two and three represent different interesting navigations related to pages 657 and 984. Cluster two represents navigations regarding pages 657 and 984 along with navigations to pages 996 and 998. Cluster three represents navigations to pages 657 and 984 along with navigations to pages 815 and 813. This means that, server sessions in cluster two holding pages 657 and 984 also hold pages 996 and 998. Likewise, server sessions in cluster three holding pages 657 and 984 also hold pages 815 and 813. Although not all of

the numbers outside the diagonal have zero values, we may say that most of them do (or are lower than 1%) and that the model fits the data well.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(163, 349)	22.29	67.27
	(349, 524)	14.88	22.90
	(349, 657)	12.79	19.68
	(159, 349)	11.29	41.37
	(163, 159)	10.92	32.96
2	(657, 984)	62.71	86.82
	(984, 657)	40.01	40.03
	(657, 984, 657)	26.66	42.51
	(984, 996)	17.43	17.44
3	(984, 998)	11.73	11.74
	(657, 815)	65.63	65.69
	(815, 657)	57.97	80.15
	(657, 815, 657)	50.35	76.71
	(657, 813)	18.30	18.32
4	(657, 984)	17.65	17.66
	(657, 1082)	17.42	36.50
	(882, 883)	17.02	68.42
	(1082, 657)	15.01	37.06
	(657, 882)	13.29	27.85
	(857, 868)	12.79	62.56

Table 5.12: Open sequences having five highest support values within each cluster.

Cluster	Open sequences	Support (%)	Confidence (%)
1	(657, 163, 349)	2.62	66.04
	(349, 627, 349)	3.07	39.81
	(163, 159, 349)	3.44	31.55
	(657, 159, 349)	1.57	30.88
	(349, 452, 349)	2.24	30.61
2	(984, 996, 999, 657, 1013)	1.40	91.67
	(984, 996, 998, 657, 1013)	1.19	73.68
	(657, 984, 1006, 984, 657)	2.17	68.00
	(657, 984, 1001, 984, 657)	1.74	67.21
	(984, 657, 984, 1006, 657)	1.02	64.86
3	(815, 657, 984, 657, 813, 657)	1.08	100.00
	(815, 657, 794, 657, 815, 657)	1.04	100.00
	(657, 815, 810, 657, 813, 657)	1.04	100.00
	(657, 815, 984, 657, 813, 657)	1.04	100.00
	(657, 794, 657, 813, 657)	1.20	100.00
	(815, 984, 657, 813, 657)	1.20	100.00
4	(657, 1082, 1091, 1082, 657)	1.21	85.71
	(657, 1091, 1082, 657)	1.21	85.71
	(882, 883, 882, 883, 882)	1.11	84.62
	(984, 657, 1082, 657)	1.01	83.33
	(657, 1082, 1091, 1082)	1.41	82.35
	(657, 1082, 1091, 657)	1.41	82.35

Table 5.13: Open sequences having five highest confidence values within each cluster.

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(163, 349)	22.29	67.27	0.17	4.26	0.19	5.56	0.00	0.00
(349, 524)	14.88	22.90	0.13	20.00	0.00	0.00	0.00	0.00
(349, 657)	12.79	19.68	0.43	66.67	0.31	72.73	0.10	100.00
(159, 349)	11.29	41.37	0.13	14.29	0.00	0.00	0.00	0.00
(163, 159)	10.92	32.96	0.30	7.45	0.15	4.44	0.00	0.00
(657, 163, 349)	2.62	66.04	0.00	0.00	0.00	0.00	0.00	0.00
(349, 627, 349)	3.07	39.81	0.00	0.00	0.00	0.00	0.00	0.00
(163, 159, 349)	3.44	31.51	0.00	0.00	0.00	0.00	0.00	0.00
(657, 159, 349)	1.57	30.88	0.00	0.00	0.00	0.00	0.00	0.00
(349, 452, 349)	2.24	30.61	0.00	0.00	0.00	0.00	0.00	0.00
(657, 984)	0.45	1.46	62.71	86.82	17.65	17.66	3.63	7.59
(984, 657)	0.75	58.82	40.01	40.03	12.73	67.01	3.42	70.83
(657, 984, 657)	0.22	50.00	26.66	42.51	10.99	62.28	2.72	75.00
(984, 996)	0.15	11.76	17.43	17.44	2.13	11.20	1.01	20.83
(984, 998)	0.00	0.00	11.73	11.74	0.58	3.05	0.20	3.27
(984, 996, 999, 657, 1013)	0.00	0.00	1.40	91.67	0.00	0.00	0.00	0.00
(984, 996, 998, 657, 1013)	0.00	0.00	1.19	73.68	0.00	0.00	0.00	0.00
(657, 984, 1006, 984, 657)	0.00	0.00	2.17	68.00	0.35	81.82	0.00	0.00
(657, 984, 1001, 984, 657)	0.00	0.00	1.74	67.21	0.23	85.71	0.10	100.00
(984, 657, 984, 1006, 657)	0.00	0.00	1.02	64.86	0.00	0.00	0.00	0.00
(657, 815)	0.75	2.43	4.85	6.71	65.63	65.69	11.58	24.26
(815, 657)	0.67	75.00	4.04	72.52	57.97	80.15	11.78	84.78
(657, 815, 657)	0.45	60.00	3.27	67.54	50.35	76.71	9.16	79.13
(657, 813)	0.52	1.70	0.47	0.65	18.30	18.32	0.81	1.69
(657, 984)	0.45	1.46	62.71	86.82	17.65	17.66	3.63	7.59
(815, 657, 984, 657, 813, 657)	0.00	0.00	0.00	0.00	1.08	100	0.00	0.00
(815, 657, 794, 657, 815, 657)	0.00	0.00	0.00	0.00	1.04	100	0.00	0.00
(657, 815, 810, 657, 813, 657)	0.00	0.00	0.00	0.00	1.04	100	0.00	0.00
(657, 815, 984, 657, 813, 657)	0.00	0.00	0.00	0.00	1.04	100	0.00	0.00
(815, 984, 657, 813, 657)	0.00	0.00	0.00	0.00	1.20	100	0.00	0.00
(657, 794, 657, 813, 657)	0.00	0.00	0.00	0.00	1.20	100	0.00	0.00
(882, 883)	0.00	0.00	0.13	100.00	0.23	28.57	17.02	68.42
(657, 1082)	0.00	0.00	0.94	1.29	2.36	2.36	17.42	36.50
(1082, 657)	0.00	0.00	0.55	54.17	2.55	88.00	15.01	37.06
(657, 882)	0.00	0.00	0.13	0.18	0.70	0.70	13.29	27.85
(857, 868)	0.00	0.00	0.13	100.00	0.00	0.00	12.79	62.56
(657, 1082, 1091, 1082, 657)	0.00	0.00	0.00	0.00	0.15	100.00	1.21	85.71
(657, 1091, 1082, 657)	0.00	0.00	0.00	0.00	0.15	100.00	1.21	85.71
(882, 883, 882, 883, 882)	0.00	0.00	0.00	0.00	0.00	0.00	1.11	84.62
(984, 657, 1082, 657)	0.00	0.00	0.30	43.75	0.23	85.71	1.01	83.33
(657, 1082, 1091, 1082)	0.00	0.00	0.00	0.00	0.15	50.00	1.41	82.35
(657, 1082, 1091, 657)	0.00	0.00	0.00	0.00	0.23	75.00	1.41	82.35

Table 5.14: Evaluating open sequences in other clusters.

Finally, a random sample was drawn of 12 open sequences that were not selected by high support values for cluster description. Table 5.15 provides support and confidence values of open sequences with low or average support values for each cluster.

Open sequences	1		2		3		4	
	S	C	S	C	S	C	S	C
(163, 657)	5.46	16.48	1.28	31.91	2.55	73.33	0.00	0.00
(159, 657)	7.26	26.58	0.00	0.00	0.00	0.00	0.00	0.00
(159, 162)	3.89	14.25	0.00	0.00	0.00	0.00	0.00	0.00
(997, 984)	0.00	0.00	4.00	50.00	0.00	0.00	0.00	0.00
(1021, 984)	0.00	0.00	7.14	80.00	0.00	0.00	0.00	0.00
(1026, 1025)	0.00	0.00	1.96	14.42	0.00	0.00	0.00	0.00
(1129, 657)	0.00	0.00	0.00	0.00	3.48	90.00	0.00	0.00
(813, 657)	0.00	0.00	0.00	0.00	14.74	73.41	0.00	0.00
(1092, 1082)	0.00	0.00	0.00	0.00	0.00	0.00	7.65	68.47
(1082, 1083)	0.00	0.00	0.00	0.00	0.00	0.00	11.18	27.61
(1082, 815)	0.00	0.00	0.00	0.00	1.47	50.67	6.24	15.42
(882, 886)	0.00	0.00	0.00	0.00	0.00	0.00	9.97	40.08

Table 5.15: Evaluating open sequences with low or average support values.

5.8 Results

Table 5.16 presents the number of interesting frequent item sets that are found in the experiment per type of comparison. Furthermore, for each type of comparison, positive differences between demonstrated evidences for βi are distinguished from negative differences between demonstrated evidences for βi . If differences between evidences are zero, the corresponding frequent item set can never be interesting. Because evidence pairs are used without taking any degree for lack of evidence into account (i.e. $e_{d\beta i}^u = e_{p\beta i}^u$, $e_{d\beta i}^s = e_{p\beta i}^s$, $e_{d\beta i}^c = e_{p\beta i}^c$), differences between demonstrated evidences for βi will always be the same as differences between possible evidences against βi . By comparing usage with structure evidence, all of the interesting frequent item sets in the analysis are identified showing negative differences. Moreover, by comparing usage with combined evidence and structure with combined evidence, all of the interesting frequent item sets in the analysis are identified showing respectively negative and positive differences.

Type of comparison					
Usage - Structure		Usage - Combined		Structure - Combined	
Positive difference ($e_{d\beta i}^u - e_{d\beta i}^s$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^s$) < 0	Positive difference ($e_{d\beta i}^u - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^c$) < 0	Positive difference ($e_{d\beta i}^s - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^s - e_{d\beta i}^c$) < 0
0	91	0	46	45	0

Table 5.16: Number of interesting frequent item sets per type of comparison and for positive and negative differences between demonstrated evidences for βi .

In table 5.17a and 5.17b, given the results of our experiment of applying SAM¹ to <http://machines.hyperreal.org>, a meaning is given to each category of outcome of interesting frequent items. In a decision table, the outcome is predicted per level of evidence and per type of comparison, taking into account positive and negative differences between sources of evidence. Also, distinctions are made whether usage and/or structure evidence are $\neq 0$, $\neq 0.5$ and $\neq 1$. Suggestions for improving the structure of the web site may be as follows. Frequent item sets that are found interesting in category (1) suggest inserting links between web pages. Frequent item sets that are found interesting in category (2) suggest deleting links between web pages or moving pages elsewhere in the structure of the web site. ‘-’ indicates that comparing evidence can never be larger or smaller than zero. Detailed information about

suggestions for improving the structure of the web site <http://machines.hyperreal.org> is given in section 5.9 Deploying the results.

Sources of evidence / type of comparison	<i>Structure evidence $\neq 0$ and $\neq 0.5$ and $\neq 1$</i>						<i>Structure evidence = 0</i>					
	Usage vs structure		Usage vs combined		Structure vs combined		Usage vs structure		Usage vs combined		Structure vs combined	
	Positive difference ($e_{d\beta i}^u - e_{d\beta i}^s$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^s$) < 0	Positive difference ($e_{d\beta i}^u - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^c$) < 0	Positive difference ($e_{d\beta i}^s - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^s - e_{d\beta i}^c$) < 0	Positive difference ($e_{d\beta i}^u - e_{d\beta i}^s$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^c$) < 0	Positive difference ($e_{d\beta i}^u - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^u - e_{d\beta i}^c$) < 0	Positive difference ($e_{d\beta i}^s - e_{d\beta i}^c$) > 0	Negative difference ($e_{d\beta i}^s - e_{d\beta i}^c$) < 0
<i>Usage evidence $\neq 0$ and $\neq 0.5$ and $\neq 1$</i>	(1)	(2)	(1)	(2)	(1)	(2)	(1)	-	(1)	-	-	-
<i>Usage evidence = 0</i>	-	(2)	-	-	(2)	-	-	-	-	-	-	-
<i>Usage evidence = 0.5</i>	(1)	(2)	(1)	(2)	-	-	(1)	-	(1)	-	-	-
<i>Usage evidence = 1</i>	(1)	-	-	-	-	(1)	(1)	-	(1)	-	-	-
(1) = Identification of interesting frequent item sets that are used together <i>more</i> than would be expected from the structure of the web site (2) = Identification of interesting frequent item sets that are used together <i>less</i> than would be expected from the structure of the web site												

Table 5.17a: Decision table providing a meaning for interesting frequent item sets if $IM \geq \tau$, specified for different levels of evidence and positive/negative differences between types of comparisons.

Sources of evidence / type of comparison	<i>Structure evidence = 0.5</i>						<i>Structure evidence = 1</i>					
	Usage vs structure		Usage vs combined		Structure vs combined		Usage vs structure		Usage vs combined		Structure vs combined	
	Positive difference ($e_{d\beta i^u} - e_{d\beta i^s}$) > 0	Negative difference ($e_{d\beta i^u} - e_{d\beta i^s}$) < 0	Positive difference ($e_{d\beta i^u} - e_{d\beta i^c}$) > 0	Negative difference ($e_{d\beta i^u} - e_{d\beta i^c}$) < 0	Positive difference ($e_{d\beta i^s} - e_{d\beta i^c}$) > 0	Negative difference ($e_{d\beta i^s} - e_{d\beta i^c}$) < 0	Positive difference ($e_{d\beta i^u} - e_{d\beta i^s}$) > 0	Negative difference ($e_{d\beta i^u} - e_{d\beta i^s}$) < 0	Positive difference ($e_{d\beta i^u} - e_{d\beta i^c}$) > 0	Negative difference ($e_{d\beta i^u} - e_{d\beta i^c}$) < 0	Positive difference ($e_{d\beta i^s} - e_{d\beta i^c}$) > 0	Negative difference ($e_{d\beta i^s} - e_{d\beta i^c}$) < 0
<i>Usage evidence ≠ 0 and ≠ 0.5 and ≠ 1</i>	(1)	(2)	-	-	(2)	(1)	-	(2)	-	(2)	-	-
<i>Usage evidence = 0</i>	-	(2)	-	-	(2)	-	-	(2)	-	-	(2)	-
<i>Usage evidence = 0.5</i>	-	-	-	-	-	-	-	-	-	-	-	-
<i>Usage evidence = 1</i>	(1)	-	-	-	-	(1)	-	-	-	-	-	-
(1) = Identification of interesting frequent item sets that are used together <i>more</i> than would be expected from the structure of the web site (2) = Identification of interesting frequent item sets that are used together <i>less</i> than would be expected from the structure of the web site												

Table 5.17b: Decision table providing a meaning for interesting frequent item sets if $IM \geq \tau$, specified for different levels of evidence and positive/negative differences between types of comparisons.

In the sections that follow, first a graphical overview is given of interesting navigations on the web site <http://machines.hyperreal.org>, providing the structure of the web site along with URL addresses and the order in which pages are visited. Second, interesting information is provided about non-existing navigations given a provided structure.

5.8.1 Interesting navigations on <http://machines.hyperreal.org> presented in a graph.

5.8.1.1 Composition of the graph

In figures 5.15 and 5.16, interesting navigations, along with direct hyperlinks between pages and parts of the structure of the web site <http://machines.hyperreal.org>, are graphically depicted in each cluster. For each interesting page, the page_id is given along with (a part of) the URL address of this particular page, which is written under the page_id inside the rectangle. The complete URL address of each page can be read taking into account the level in the web site structure and the links. For example, page 657 constitutes the main page with URL address <http://machines.hyperreal.org>. Going one level downwards, three different web pages appear. The complete URL address of page 349 is <http://machines.hyperreal.org/manufacturers>. Proceeding towards, for example, page 868, the URL address <http://machines.hyperreal.org/manufacturers/ARP/Odyssey> is given. Other examples of how to read URL addresses are <http://machines.hyperreal.org/manufacturers/Roland/Juno> for page 996 and <http://machines.hyperreal.org/manufacturers/Roland/JX> for page 998.

The dashed rectangles in figures 5.15 and 5.16 originated from different logged URL addresses in the files. However, the content of the web page appears to be exactly the same as the one given by the solid rectangles. Further analysis revealed that the log files also stored information of people who used the URL address www.hyperreal.org and navigations from this main page on. For example, page 159 appears to be exactly the same as page 815. The only difference is that page 159 is navigated through www.hyperreal.org/guide and page 815 is navigated through <http://machines.hyperreal.org/guide>. We would like to keep this distinction in our analysis because these web pages appear within interesting related pages.

As already mentioned in chapter four, links between pages are drawn by *thin black solid arrows*, while interesting navigations, including order-based information, are given by the *bigger dashed arrows*. For example, from page 657, people can go to pages 349, 815, 810 and 813 and from each of these pages a link points back to the home page. Also, from page 349 other pages

may be visited like 857, 984, 882, 1082 as well as 657, 815 and 810. In figure 5.15, interesting navigations are represented by *open sequences with the five highest support values* for each cluster. The same navigations were previously given in table 5.12. In figure 5.16, for each cluster, interesting navigations are represented by *the five highest support values of combinations of the order of pages within interesting related pages*. Support (s) and confidence (c) values are written next to or above the arrows. For example, in cluster one of figure 5.15 and 5.16, 22.29% of the cases visited page 163 before page 349. The confidence value indicates that, if people visit page 163, the chance that they will visit page 349 thereafter is 67.27%. In cluster four of figure 5.15, 17.42% of the cases visit page 657 before page 1082 and 15.01% of the cases visit page 1082 before page 657. Nevertheless, we could also use open sequences or interesting related web pages selected by high confidence values. Yet, the open sequences with high support values are more efficient for graphical presentations since they summarize, for each cluster, the *most occurring* (i.e. most frequent) navigations.

For evaluation purposes, distribution of server sessions is given in the upper left corner of every cluster. For example, in figure 5.15, 18.40% of the server sessions in the input file are grouped in cluster one. Practically, this means that 1,337 out of the 7,266 server sessions are grouped in cluster one.

In order to avoid complex drawings of arrows making the figures unclear, some modifications are made in figures 5.15 and 5.16. First, with regard to the links between pages, some arrows point towards a particular `page_id`. For example, from pages 857, 984, 882, 1082 one may proceed to pages 657, 815 and 810. Likewise, from pages 868, 996, 998, 883 one may proceed to pages 349, 657, 815 and 810. Second, the dashed parts of the links indicate that there is no intersection with other links. If there were no dashed parts, the links could be misinterpreted, saying, for example, that from page 984 a link points to page 882. Third, with regard to the presentation of interesting navigations, lines showing arrows in the middle of navigations, instead of at the beginning or at the end, may appear. For example, in cluster two of figure 6.14, when navigating from page 657 to page 984 and from page 984 to page 657, somewhere in the middle of both navigations, an arrow is drawn. These arrows are used for interpreting open sequences or frequently visited pages having more than two elements. Support and confidence values are given next to or above the arrow of the last navigation. In cluster two, an interesting navigation appears in the following order: 657, 984, 657 with support and confidence values of 26.66% and 42.51%. Fourth, with regard to the magnitude of the structured web site with interesting related pages, for each cluster, only part of the site is given that is relevant for describing the interesting navigations.

5.8.1.2 Interesting navigations presented by open sequences with high support values

In figure 5.15, *cluster one* mainly represents navigations to and from the ‘*manufacturers*’ page. In general, visitors also use the URL address www.hyperreal.org and the navigations represent usage behaviour that is very interesting. In appendix 5 relatively high interestingness measures of 1.3968, 1.3087, 1.2415, 1.1183 and 1.0794 are defined for the following interesting beliefs of related pages: (657, 349), (163, 159), (163, 349), (349, 159) and (349, 524). The added value of the information extracted by cluster one, compared with the information provided the table of appendix 5, is that cluster one shows not only information about pages but also the order in which those pages are visited. Interesting to know is that people go from ‘*manufacturers*’ to the home page www.machines.hyperreal.org, instead of the other way around. Also, if people use the URL address www.hyperreal.org only one-way of traffic is interesting, going from www.hyperreal.org towards ‘*manufacturers*’. Referring to tables 5.16 and 5.17, all of the interesting navigations in cluster one fall within the same categories. This means that cluster one identifies profiles of interesting navigations between web pages that are used together *less* than would be expected from the structure of the web site.

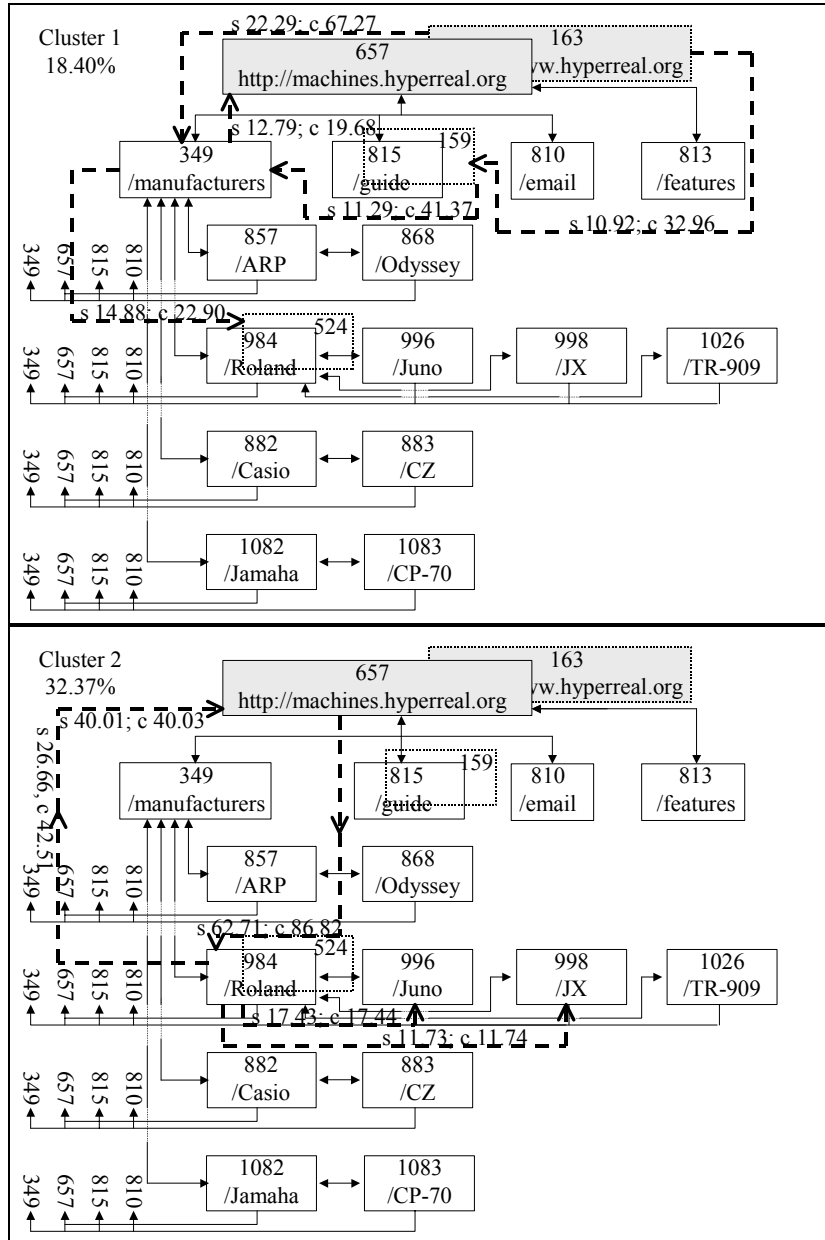
Cluster two mainly extracts interesting navigations that are concentrated around the ‘*Roland*’ page. For example, visitors use the link from ‘*Roland*’ to ‘*JX*’ and from ‘*Roland*’ to ‘*Juno*’ less than expected from the structure of the web site. Note that (657, 984) or (984, 657) is not written as an interesting belief of two related pages in appendix 5. We remind that open sequences are used to describe the order of pages within server sessions grouped into clusters. They are primarily used to give a clear view of the reality and most occurring order based visiting behaviour. Open sequences are not used to distinguish between interesting and non-interesting navigations.

In *cluster three*, interesting navigations with regard to the ‘*home*’ page are extracted. Links from ‘<http://machines.hyperreal.org>’ to ‘*guide*’ and the other way around, from ‘*guide*’ to ‘<http://machines.hyperreal.org>’ are used less than expected from the structure of the web site. Yet, from ‘<http://machines.hyperreal.org>’ to ‘*features*’ is interesting in only one direction. This one-way, order based information indicated by open sequences is not provided in appendix 5.

Finally, in *cluster four*, interesting navigations from ‘*Casio*’ to ‘*CZ*’ and from ‘*ARP*’ to ‘*Odyssey*’ occur less than expected from the structure of the web site. This also means that, given a T-value of 0.75, navigations the other way around i.e. from ‘*CZ*’ to ‘*Casio*’ and from ‘*Odyssey*’ to ‘*ARP*’ are not

considered interesting since they occur as frequent as expected from the structure of the web site.

In the analysis above, interesting navigations are discovered based on the second (2) and fourth (4) category of comparison. Yet, in figure 5.15, some of the navigations present visiting patterns between web pages that are used together less than would be expected from the structure of the web site (interesting navigations), others present visiting patterns between web pages that are used together as frequent as expected from the structure of the web site (uninteresting navigations). The reason why uninteresting navigations are also given in figure 5.15 is because they have high support values in the data. In order to provide only interesting navigations, the clusters are examined by combinations of the order of pages within interesting beliefs of related pages. The results are given in the following section.



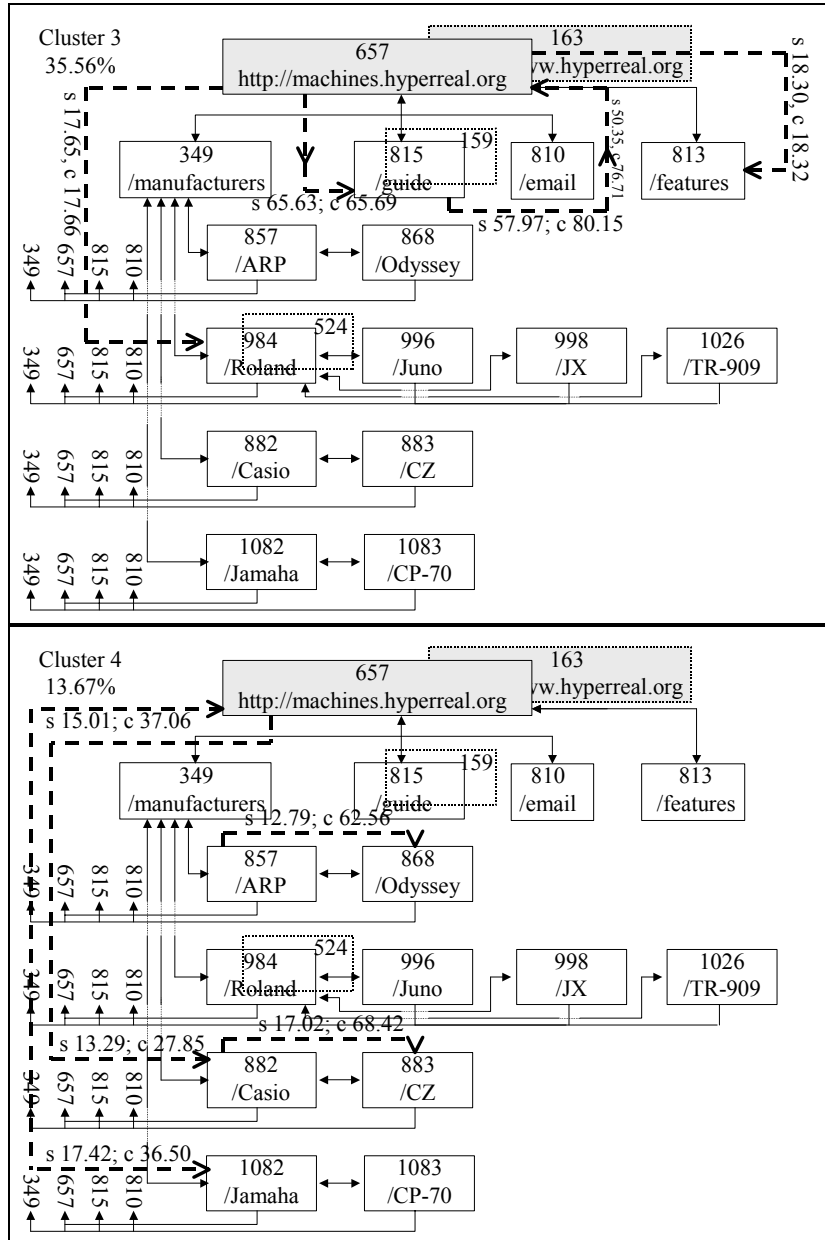


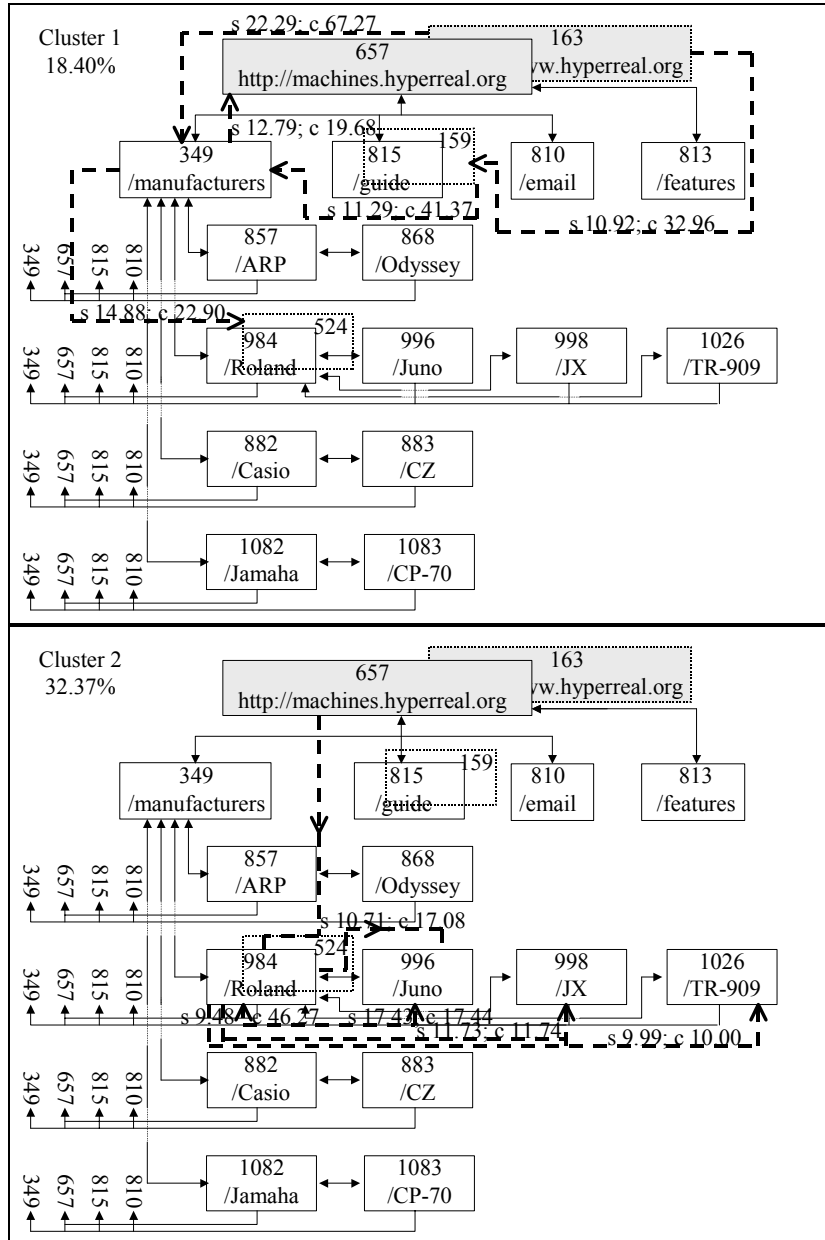
Figure 5.15: Interesting navigations on <http://machines.hyperreal.org>, presented by open sequences with high support values.

5.8.1.3 Interesting navigations presented by high support values of combinations of the order of pages within interesting beliefs of related pages

In this section, for each cluster, support and confidence are calculated for every combination of the order of pages within interesting beliefs of related pages. Out of a total number of 91 frequent item sets, given appendix 5, 278 different combinations of the order of pages are defined. For each cluster, the five highest support values are used for presenting interesting navigations in figure 5.16.

Compared with figure 5.15, cluster one represents the same interesting navigations. In general, eight out of twenty navigations in figure 5.16 are not given in figure 5.15. For example, in cluster two of figure 5.16, proceeding from page 657 to page 984, followed by page 996 is not presented in figure 5.15 because the support measure is not high enough.

All of the navigations in figure 5.16 are declared interesting and fall within the second (2) and fourth (4) category, providing visiting patterns between web pages that are used together less than would be expected from the structure of the web site. One exception occurs for navigation (657, 984, 996). The frequent item set of this navigation is declared interesting in table 5.16 for category (2) and (5). However, navigation (657, 984, 996) provides the same interesting information i.e. usage behaviour from page 657 to page 984 and from page 984 to page 996 that occurs less frequent than expected from the structure of direct hyperlinks between the pages 657, 984 and 996.



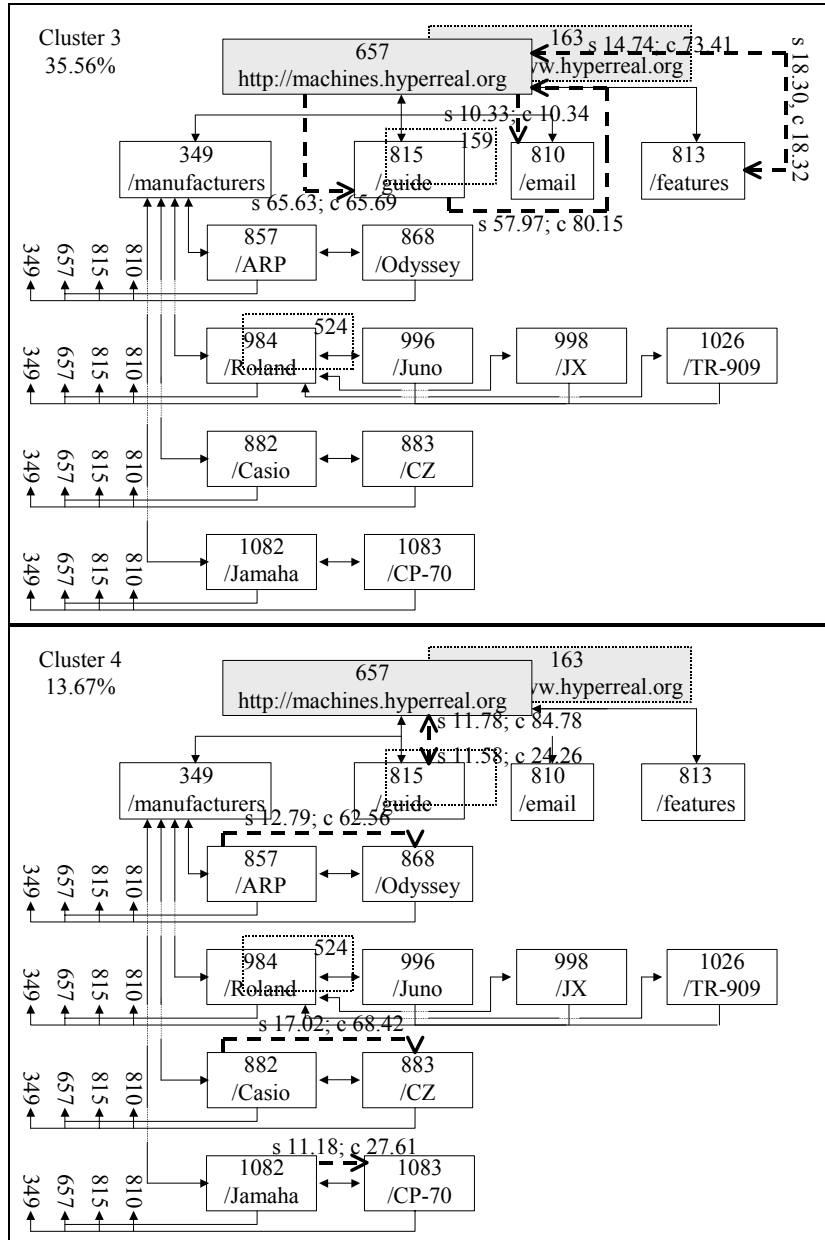


Figure 5.16: Interesting navigations on <http://machines.hyperreal.org>, presented by high support values of combinations of the order of pages within interesting beliefs of related pages.

5.8.2 *Non-existing navigations given a provided structure*

Another group in which results are categorized provides interesting information about links between pages that are nearly not (support of frequent item sets is less than 0.1%) used by the visitors. This kind of information may be used for deleting links between pages so as to prevent the web site for being too complicated or inefficient for the users. From the previous analysis, sets of pages without a frequent item set (i.e. support is less than 0.1%) automatically fall into this category. We may call these sets of pages beliefs of not related pages. Interesting beliefs of pages not being related, or interesting not frequently visited pages, are identified by three types of comparisons. Table 5.18 gives the results, ordered by level of interestingness. Because the interesting combinations of not related pages do not represent any real frequent visiting behaviour, the order is irrelevant.

All of the interesting beliefs of not related pages, following our analysis on <http://machines.hyperreal.org>, fall within category (2). This means that interesting frequent item sets of web pages are identified that are, given a provided structure of direct hyperlinks, barely used by the visitor.

Interesting beliefs of not related pages	Evidence			$IM_{bi} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
(349, 406)	[0.0019; 0.0019]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.4115 (-)	1.4115 (-)	-
(349, 937)	[0.0022; 0.0022]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.4111 (-)	1.4111 (-)	-
(349, 959)	[0.0027; 0.0027]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.4103 (-)	1.4103 (-)	-
(349, 407)	[0.0042; 0.0042]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.4082 (-)	1.4082 (-)	-
(820, 827)	[0.0337; 0.0337]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3665 (-)	1.3665 (-)	-
(857, 862)	[0.2100; 0.2100]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.1172 (-)	1.1172 (-)	-
(852, 853)	[0.2647; 0.2947]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0398 (-)	1.0398 (-)	-

Table 5.18: Interesting beliefs of not related pages on <http://machines.hyperreal.org>.

5.9 Deploying the results

Both groups of interesting existing navigations providing order-based information of visited pages as well as non-existing navigations given a provided structure, described in sections 5.8.1 and 5.8.2, may be used for link optimisation studies. In order to develop a web site structure conform to visiting behaviour of users, links between pages that are not optimally used may be deleted or the pages may be moved elsewhere in the structure of the web site. Likewise, interesting navigations between web pages, without any direct links, may suggest link insertion for the convenience of the user. Finally, considering interesting non-existing navigations given a provided structure suggests link deletion.

5.9.1 *Suggestions for reorganizing pages or deleting direct links*

For this project we may give some suggestions with regard to link optimisation at the web site <http://machines.hyperreal.org>. First, links between the following pages are not used as frequently as expected from the current structure of the web site. Therefore we may suggest reorganizing the pages or deleting the direct links between pages with page_id and corresponding URL address given in table 5.19. In the last column the interestingness measure is given. This measure may give an indication of the ‘urgency’ of reacting to the behaviour of web users. The higher the interestingness measure, the more urgent it is to respond to visiting behaviour by optimising the structure of the web site. Note that in table 6.19 suggestions are given for the most frequent occurring and interesting patterns presented by the open sequences in figure 5.15. Suggestions for reconsidering the structure between pages, inserting or deleting links may also be studied for other interesting order-based navigation patterns, for example those not having high support values. Likewise, figure 5.16 may be used for suggestions with regard to link optimisation.

From	To	IM
349 http://machines.hyperreal.org/manufacturers	657 http://machines.hyperreal.org	1.3905
657 http://machines.hyperreal.org	813 http://machines.hyperreal.org/features	1.3654
163 http://www.hyperreal.org	159 http://www.hyperreal.org/guide	1.3084
159 http://www.hyperreal.org/guide	349 http://machines.hyperreal.org/manufacturers	1.3084
163 http://www.hyperreal.org	349 http://machines.hyperreal.org/manufacturers	1.2415
984 http://machines.hyperreal.org/manufacturers/Roland	998 http://machines.hyperreal.org/manufacturers/Roland/JX	1.2385
657 http://machines.hyperreal.org	815 http://machines.hyperreal.org/guide	1.2249
815 http://machines.hyperreal.org/guide	657 http://machines.hyperreal.org	1.2249
857 http://machines.hyperreal.org/manufacturers/ARP	868 http://machines.hyperreal.org/manufacturers/ARP/Odyssey	1.1815
984 http://machines.hyperreal.org/manufacturers/Roland	996 http://machines.hyperreal.org/manufacturers/Roland/Juno	1.1518
349 http://machines.hyperreal.org/manufacturers	524 http://www.hyperreal.org/manufacturers/Roland	1.0794
882 http://machines.hyperreal.org/Casio	883 http://machines.hyperreal.org/Casio/CZ	0.9350

Table 5.19: Suggestions for reorganizing pages or deleting direct links.

Second, with regard to direct links between pages, which are barely used by the visitor (re. non-existing navigations given a provided structure, presented in table 5.18), we suggest link deletion. Suggestions for link deletion are given in table 5.20.

From	To	IM
349 http://machines.hyperreal.org/manufacturers	406 http://machines.hyperreal.org/manufacturers/email	1.4115
406 http://machines.hyperreal.org/manufacturers/email	349 http://machines.hyperreal.org/manufacturers	1.4115
349 http://machines.hyperreal.org/manufacturers	937 http://machines.hyperreal.org/manufacturers/links	1.4111
937 http://machines.hyperreal.org/manufacturers/links	349 http://machines.hyperreal.org/manufacturers	1.4111
349 http://machines.hyperreal.org/manufacturers	959 http://machines.hyperreal.org/manufacturers/Opus	1.4103
959 http://machines.hyperreal.org/manufacturers/Opus	349 http://machines.hyperreal.org/manufacturers	1.4103
349 http://machines.hyperreal.org/manufacturers	407 http://machines.hyperreal.org/manufacturers/EML	1.4082
407 http://machines.hyperreal.org/manufacturers/EML	349 http://machines.hyperreal.org/manufacturers	1.4082
820 http://machines.hyperreal.org/incoming	827 http://machines.hyperreal.org/incoming/info	1.3665
827 http://machines.hyperreal.org/incoming/info	820 http://machines.hyperreal.org/incoming	1.3665
857 http://machines.hyperreal.org/manufacturers/ARP	862 http://machines.hyperreal.org/manufacturers/ARP/Axxe	1.1172
862 http://machines.hyperreal.org/manufacturers/ARP/Axxe	857 http://machines.hyperreal.org/manufacturers/ARP	1.1172
852 http://machines.hyperreal.org/manufacturers/Alesis	853 http://machines.hyperreal.org/manufacturers/Alesis/MMT-8	1.0398
853 http://machines.hyperreal.org/manufacturers/Alesis/MMT-8	852 http://machines.hyperreal.org/manufacturers/Alesis	1.0398

Table 5.20: Suggestions for deleting direct links.

5.9.2 Suggestions for inserting direct links

Finally, links between web pages may be inserted due to the fact that, although no direct links appear between these pages, the visiting pattern is higher than expected from the structure of the site. A suggestion for link insertion with regard to the web site <http://machines.hyperreal.org> is, unfortunately, not found, given the log files and given $\tau = 0.75$. This means that inserting direct hyperlinks on <http://machines.hyperreal.org> is not very urgent.

5.10 Calculating a less severe structure evidence

In section 5.5.1 structure evidence is calculated by defining a cfactor equal to 1 if at least one direct hyperlink exists between every pair of pages in an item set (re. figure 5.3). Otherwise, cfactor equals 0. The results of applying such a severe definition of structure evidence are given in sections 5.8 and 5.9. Generally, interesting navigations are discovered providing usage behaviour that occurs *less* frequent than expected from the structure of the web site.

We may question ourselves what would have happened if less severe structure evidence had been used. Therefore, the analysis is repeated on the same data set, using the same value for τ and applying a less severe cfactor for calculations of structure evidence. In figure 5.17 the same examples of figure 5.3 are given, yet another cfactor is defined. Here, if at least one path occurs between web pages, a cfactor equal to 1 is used. Otherwise, cfactor equals 0.

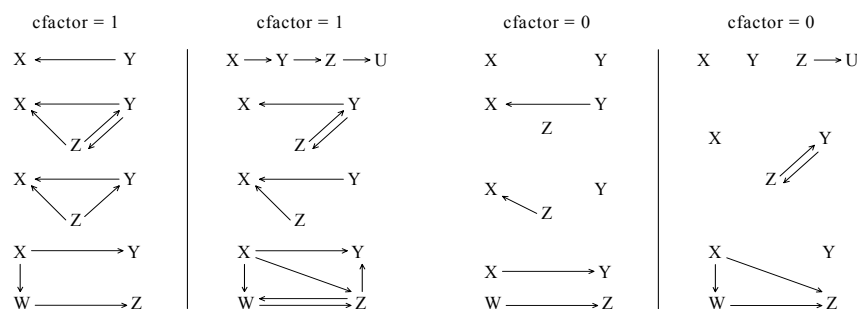


Figure 5.17: Illustration of direct hyperlinks between web pages defining a less severe connectivity factor.

The results show that 116 frequent item sets are defined interesting, which is more compared to the 91 frequent item sets that are declared interesting previously. Also, all of the previously discovered 91 interesting frequent item sets re-occur within the new set. Moreover, all of the 116 interesting frequent item sets present beliefs of related web pages that are used together *less* than would be expected from the structure of the web site. This may be explained by the following reasons. Usage evidence pairs are the same for both types of analyses and appear to be relatively low. Also, structure evidence pairs that were previously different from zero (i.e. equal to one) remain the same. Yet, some of the structure evidence pairs that were previously equal to zero rise to a higher level in the analysis if a less severely defined cfactor is used. For example, in the second column of figure 5.17, structure evidence for frequent item sets (X, Y, Z, U) and (X, Y, W, Z) equals respectively 0.25 and 0.5. Yet, structure evidence for the same frequent item sets equals 0 using the cfactor of

the previous analysis because direct hyperlinks between (X, Z), (X, U), (Y, U) and (Y, W) do not exist.

5.11 Usage behaviour that occurs more frequent than expected from the web structure

We may also question ourselves how we would have been able to discover interesting navigations providing usage behaviour that occurs *more* frequent than expected from the structure of the web site. This means that usage evidence must be larger than structure evidence with $\tau \leq IM_{\beta_i}$ and $IM_{\beta_i} =$

$= \sqrt{(|e_{d\beta_i}^1 - e_{d\beta_i}^2|)^2 + (|e_{p\beta_i}^1 - e_{p\beta_i}^2|)^2}$. Unfortunately, from the data used in the experimental tests with $\tau = 0.75$, this situation was not found due to the relatively low usage evidence for beliefs of related pages (or frequent item sets).

However, we may provide a theoretical case where beliefs of related pages are defined interesting with usage evidence larger than structure evidence. Consider a preliminary example of a web site structure given in figure 5.18. The homepage is written as 'H'. One level deeper page A, B and C are structured. At the deepest level pages D, E, F and G appear. Suppose the following five server sessions are logged in a file. Note that the backspace key is used in S_3 and S_5 .

$S_1 = H B F H B C G$

$S_2 = B C G$

$S_3 = H A E B C G$

$S_4 = H A D A B$

$S_5 = H A E B C$

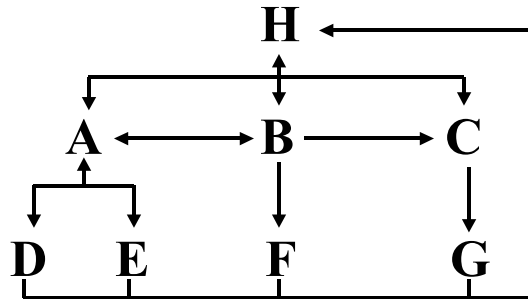


Figure 5.18: Theoretical simplified example of a web site structure.

If we consider frequent item set (B, G), structure evidence = 0 and usage evidence = 0.6 using equations (5.5) and (5.9), which provides an interestingness factor of 0.85. Using $\tau = 0.75$, frequent item set (B, G) is defined as an interesting frequent item set representing web usage behaviour that occurs more frequent than expected from the structure of the web site.

5.12 Conclusion and Future Research

In this chapter, SAM is integrated with an interestingness measure in order to discover navigations or visiting patterns that are interesting on a web site. Navigations are interesting if they are unexpected, surprising or contradicting with the structure of the web site or direct hyperlinks between web pages. Navigations are uninteresting if they are expected, known, obvious or resulting from the structure of the web site or direct hyperlinks between web pages. Interesting navigations provide information that may be used for optimising the layout of the web site through structuring of direct hyperlinks between web pages and for web personalization studies.

The principles of Baldwin's support logic create a support logic framework with a conceptual frame of evidence and beliefs for Web Usage Mining. Beliefs along with evidence pairs are automatically generated from two different sources. Structure data provide information about links between pages, which is incorporated into beliefs of related pages. Likewise, usage data provide information about visited pages on a web site, which is incorporated into beliefs of related pages. For each belief of related pages, evidence pairs are

defined from structure data and usage data, which are called structure evidence and usage evidence respectively.

Within Web Usage Mining studies, two evidence pairs, coming from different sources and referring to the same belief, may be combined into one evidence pair, which is called *combined usage evidence*, *combined structure evidence* or *combined evidence*, in order to be able to reason about evidence coming from different sources. Combined usage evidence is combining usage evidence sources of different log files registering usage behaviour of different periods. Likewise, combined structure evidence is combining sources of modified web structures of different periods. Finally, combined evidence means that usage and structure evidence are combined.

If usage evidence as well as structure evidence $\neq 0$, $\neq 0.5$ and $\neq 1$, comparing usage with structure evidence will identify interesting beliefs with conflicting evidence. Comparing usage with combined evidence will identify interesting beliefs with strong usage and weak structure evidence. Yet, comparing structure evidence with combined evidence will identify interesting beliefs with strong structure and weak usage evidence.

Interesting beliefs of related pages are defined using a threshold value τ for the differences between evidence pairs. For a high value of τ , which is at or above 0.5 (re. Cooley et al, 1999b), relatively strong differences between evidence pairs provide high interesting results. For a low value of τ , which is below 0.5 (re. Cooley et al, 1999b), relatively weak differences between evidence pairs provide low interesting results.

In search for interesting navigation patterns on <http://machines.hyperreal.org>, the value of τ is set at 0.75 in order to find patterns of the highest interest level. First, 75,855 server sessions are created out of log files registering web usage behaviour from 01/02/1999 till 28/02/1999. Frequent items sets with minimum support of 0.1% represent beliefs of related pages. From the log files, a total number of 539 beliefs, consisting of minimum two and maximum four related pages, are identified, of which 91 are declared interesting.

For each belief, usage and structure evidence are calculated. Also combined evidence is calculated in order to identify interesting beliefs with relatively strong/weak structure evidence and relatively weak/strong usage evidence. The results of our experiment provided 91 interesting beliefs of related pages when comparing usage with structure evidence. 46 interesting beliefs showed structure evidence equal to one and weak usage evidence; 44 interesting beliefs showed structure evidence equal to 0.667 and weak usage evidence; 1 interesting belief showed structure evidence equal to 0.8334 and weak usage evidence. In general, all of the 91 interesting beliefs identified related web pages that are used together less than would be expected from the structure of

the web site. No interesting beliefs are identified with regard to related pages that are used together more than would be expected from the structure of the web site, because of two reasons. First, usage evidence of frequent item sets is relatively low. Second, by setting τ at 0.75, related pages of the highest interest only are identified.

For each interesting belief of related pages, positive and negative differences between evidence pairs are examined per type of comparison. In a decision table, the outcome is predicted for usage/structure evidence \neq or $= 0$, \neq or $= 0.5$ and \neq or $= 1$. Two different meanings may be given to interesting beliefs of related pages, given our experiment on <http://machines.hyperreal.org>. First, interesting beliefs of related pages identify interesting related pages that are used together *more* than would be expected from the structure of the web site (1). This means that, despite a relatively weak topological connection, the related pages are frequently visited. Second, interesting beliefs of related pages identify interesting related pages that are used together *less* than would be expected from the structure of the web site (2). This means that, despite a relatively strong topological connection, the related pages are not visited frequent enough. The results of our experiment that are considered interesting all fall into the second meaning (2), showing negative differences when comparing usage with structure and usage with combined evidence and showing positive differences when comparing structure with combined evidence.

A typical characteristic of interesting frequent item sets, representing interesting beliefs of related pages within Web Usage Mining studies, is that underlying combinations of item sets usually are not interesting, given the results of our experiment.

During the process of identifying interesting frequent item sets (or interesting related (web) pages) on <http://machines.hyperreal.org>, no lack of evidence is tolerated, which means that the evidence shown from one of the sources defines demonstrated as well as possible evidence for a given belief. This approach is also followed by Cooley et al (1999b) and practically, this means that $e_{\alpha\beta i}^u = e_{p\beta i}^u$ and $e_{\alpha\beta i}^s = e_{p\beta i}^s$. Another remark is that, the interestingness measure, based on the support logic framework and further developed for discovering interesting patterns within Web Usage Mining (Cooley et al, 1999b), is not suitable for defining interesting frequent item sets consisting of one page. Nevertheless, the interestingness measure is useful for our studies because we investigate the order of visited pages within interesting visiting patterns. Our goal is to investigate whether the structure of direct hyperlinks between web pages may be improved and therefore we need interesting frequent item sets of minimum two pages.

SAM is integrated with the results of interesting beliefs of related pages (SAM^l). This means that SAM^l is now able to recognize interesting frequent item sets within server sessions. This also means that SAM^l ignores uninteresting frequent item sets when measuring distances between server sessions. In other words, SAM^l filters the interesting frequent item sets out of the server sessions. Following our approach of Web Usage Mining (re. chapter four, figure 4.3), SAM^l distance measures are used for clustering server sessions consisting of interesting related web pages.

The reason why server sessions are filtered (also known as ‘pre-processed’), based on the identified interesting related web pages, before instead of after the calculation of SAM distance measures is because this approach deals with noise (i.e. uninteresting patterns) in an early stage of the analysis. Data sets within Web Usage Mining studies generally contain lots of patterns that are ‘known’ or ‘obvious’ due to the structure of direct hyperlinks between web pages that is offered as a ‘navigating road’ to web visitors. Dealing with uninteresting patterns in an early stage of the analysis provides an opportunity for SAM to handle large data sets. For example, in our experiment of analysing web usage behaviour on <http://machines.hyperreal.org>, the data set is reduced from 75,855 to 7,266 server sessions after pre-processing the server sessions into server sessions holding interesting related, frequently visited web pages. This means that 68,589 server sessions do not hold interesting related, frequently visited web pages. If the original data set of 75,855 server sessions were first used to calculate SAM distance measures, we would end up with an explosion of SAM distance measures (i.e. $[75,855 \times 75,854] / 2 = 2,876,950,000$ SAM distance measures). Moreover, we would also face the problem of distance-based clustering (re. chapter seven) before we could ‘post-process’ the server sessions into server sessions holding interesting combinations of web pages. This way we unnecessarily burden the analysis with data, which is in fact noise.

The reason why server sessions, holding interesting related web pages, are clustered is to provide large groups of different interesting visiting patterns. This provides an overview of interesting patterns actually occurring on the web site. It also shows small difference in interesting patterns within the same cluster and large differences between interesting patterns across different clusters. If we omit the clustering procedure of server sessions holding interesting related web pages, it would be difficult to provide an overview of several different large groups of interesting visiting patterns, to examine small differences within the groups and major differences across the groups.

Clustering server sessions based on SAM^l distance measures identifies profiles representing interesting navigations on a web site. The difference between frequent item sets and navigations is that navigations provide

information about the order of visited pages. Frequent item sets do not provide order-based information. The profiles, providing interesting order-based information of navigations on <http://machines.hyperreal.org> are the following. First, interesting navigations to and from <http://machines.hyperreal.org/manufacturers> are given. Second, interesting navigations to and from <http://machines.hyperreal.org/manufacturers/Roland> are shown. Third, interesting navigations with regard to the home page are given. Fourth, interesting navigations are concentrated around <http://machines.hyperreal.org/manufacturers/Casio> and <http://machines.hyperreal.org/manufacturers/Jamaha>.

The added value of clustering based on SAM¹ (compared to interesting frequent item sets) is shown through large groups presenting profiles of interesting information about the order in which pages are visited. Also, small differences within profiles and large differences across profiles are presented. This means that clustering server sessions based on SAM¹ provides more information about what is actually going on at the web site. For example, interesting belief of related pages (163, 349) and (159, 349) do not provide order-based information and show an interestingness measure of respectively 1.2415 and 1.1183. Clustering server sessions based on SAM¹ additionally informs us that people actually and mostly navigate through the web site from page 163 followed by page 349 and from page 159 followed by page 349, instead of the other way around. Moreover, interesting navigations with regard to pages 163, 349 and 159 are quite similar to interesting navigations with regard to page 524 (indicated by profile or cluster 1), while they are quite dissimilar to interesting navigations with regard to pages 657, 984, 996, 998 (indicated by profile or cluster 2). The information provided by the clusters may be used for re-structuring web pages or deleting/inserting direct hyperlinks between web pages. In the example above, since people navigate from page 163 to page 349 less than expected from the structure, the direct hyperlink from page 163 to page 349 may be deleted. Likewise, since people actually navigate less from page 349 to 163, the direct hyperlink from page 349 to 163 may be deleted as well. The web developer may also consider moving pages 163 and 349 elsewhere in the structure of the web site, conform to visiting behaviour of users.

We may conclude that the model of clustering based on SAM¹ fits the data well because of the following reasons. First, all of the four clusters show high exclusivities for interesting web pages. In particular, 90% of the pages have exclusivity above 0.80. Second, order-based information is well represented by the clusters. Open sequences show relatively high support and confidence values for only one of the four clusters.

Re-structuring a web site, by means of moving pages elsewhere in the structure, deleting links between particular pages and inserting links between other pages, conform with the behaviour of visitors, shows that the discovered information is easily integrated with existing processes. The ease of integration is one of the criteria for a successful data mining project, mentioned in chapter two, section 2.3. Note that our project falls into the category of Web Usage Mining as well as data mining, because Web Usage Mining is in fact a data mining project where web usage data is analysed.

Finally, since the data set of 75,855 server sessions used for clustering based on SAM^I is very similar to the data set of 3,131 server sessions used for clustering based on SAM (re. chapter four, data set 2), we may provide an indication how the results would look like if SAM had been applied to 75,855 server sessions. Comparing four clusters based on SAM^I with five clusters based on SAM, differences between clusters of both algorithms are found due to presence and absence of interesting frequent item sets within navigations. This means that, if SAM^I is used for analysing usage behaviour on <http://machines.hyperreal.org>, navigations are discovered with an interestingness measure at or above τ . If SAM is used, navigations are discovered which are generally smaller than τ and, as such, not declared interesting. For example, given $\tau = 0.75$, cluster one based on SAM^I provides the following interesting navigations: 163 followed by 349 (IM = 1.2415), 163 followed by 159 (IM = 1.3084), 159 followed by 349 (IM = 1.1183) and 349 followed by 524 (IM = 1.0794). Yet, cluster one based on SAM provides the following actual but, unfortunately, uninteresting navigations: 657 followed by 1082 (IM = 0.6289), 657 followed by 947 (IM = 0.5840), 338 followed by 1153 (IM = 0.5050), 804 followed by 190 (IM = 0.5587), 657 followed by 933 and vice versa (IM = 0.5523).

Topics for future research include extending MDSAM with an interestingness measure in order to distinguish interesting two-dimensional navigations from those that are uninteresting. Also, the interestingness measure should be sensitive to the 'depth' of pages in the web site structure. Studies must verify whether the a-priori probability of finding related pages, which are situated 'deep' in the web site structure is smaller than the probability of finding related pages, which are situated at the 'top' in the web site structure. Another topic for future research is developing the cfactor for a range of values instead of using two values for the cfactor (zero and one). Also, instead of working with no lack of evidence, the effect of different categories of lack of evidence on the results must be further investigated. For example, with regard to usage evidence, demonstrated evidence is provided by real web usage behaviour registered in log files that are used within the analysis while possible evidence may be provided by probabilities. Calculating probabilities for

possible evidence may be done through examination of web usage behaviour registered in log files during a relatively long period of, say, three or more years. Finally, SAM¹ may be extended with an interestingness measure for discovering interesting frequent item sets of one page. Usage as well as structure evidence must be developed in such a way that surprising effects measure real situations.

CHAPTER 6

SENSITIVITY ANALYSIS

In the preceding chapters, web usage data is analysed using SAM. This means that SAM, 2-DIM SAM or SAM¹ distance measures are calculated between server sessions. The distance measures are inserted into a distance matrix, which is used by a hierarchical clustering algorithm for defining clusters of server sessions. Finally, the clusters of server sessions represent profiles of visiting behaviour showing order-based information of (interesting) visited pages and/or visiting times with regard to web usage behaviour.

The parameters of the SAM application are reflected by operation weights (also called costs) i.e. weights for deletion, insertion and substitution operations. In order to obtain more insight into the influence of SAM parameters on the calculations of and relations between SAM distance measures as well as on the final results (i.e. clusters of server sessions), we will analyse in this section the sensitivity of SAM. First, we provide an overview of properties with regard to SAM. In the following section, we evaluate, through experimental tests, whether these properties are true. We also examine to what extent SAM distance measures are changed if SAM parameters are changed. Then, we study how SAM distance measures change for sampled sets with differences in the total number of items and in average sequence length. Finally, before we conclude and define avenues for future research, we examine how changes in SAM parameters influence the final clustering results on a real data set.

Parts of this chapter regarding the experimental tests, which are performed in order to examine the sensitivity of SAM to differences in operational weights and differences in sampled sets, are also described in Van Baelen (2003). Within our research group, this project is executed by Walter Van Baelen and myself. A first draft of the results is printed for a Masters degree in Economic Engineering Computer Science (Van Baelen, 2003). In this chapter, the results are used for further investigation how SAM's parameter settings may influence the final clustering results.

6.1 SAM properties

SAM uses three different parameters within its analysis of measuring distances between sequences. Each parameter refers to an operation and bears an operational cost (also called weight) reflecting the amount of work for changing one sequence into the other. *Sensitivity analysis* measures how changes in parameters affect (relations between) SAM distance measures, which will affect the *final results* (i.e. clusters of server sessions). Before proceeding to the sensitivity analysis of SAM, we first illustrate some properties with regard to different parameter settings of SAM.

6.1.1 Operation weights providing equal SAM distance measures

Different weights for substitution operations will not always provide different SAM distance measures. Below, situations are given when different parameters provide equal SAM distance measures. Due to the minimum cost principle of SAM (re. Chapter three, equation (3.1)), substitution operations are replaced by deletion and insertion operations if substitution costs more than the sum of deletion and insertion. This means that, if the weights given to deletion and insertion are equal and substitution weights are equal to or higher than the sum of the weights for deletion and insertion, the SAM distance measures will be the same for different operation weights. Table 6.1 provides some examples of different weights assigned to insertion, deletion and substitution operations. For s_1 and s_2 , given in the first row, SAM distance measures are calculated using the parameters of the corresponding examples. Following the SAM algorithm, deletion and insertion operations must be applied to the source (first) sequence in order to change the source into the target sequence. The operations, necessary to equalize s_1 with s_2 (or to change s_1 into s_2), are one substitution, two deletions in s_1 (i.e. elements 4 and 7) and one insertion in s_1 (i.e. element 5). The examples show that, although different parameter settings are used, due to the minimum costs principles of SAM, the same SAM distance measures are calculated between s_1 and s_2 .

Source sequence: $s_1 = 1\ 2\ 3\ 4\ 7$ Target sequence: $s_2 = 2\ 1\ 3\ 5$				
Example	Parameters			$d_{SAM}(s_1, s_2)$
	w_d	w_i	w_s	
1	1	1	2	5
2	1	1	3	5
3	1	1	100	5

Table 6.1: Examples of operation weights providing equal SAM distance measures ($w_d = w_i$ and $w_s \geq w_d + w_i$).

6.1.2 Operation weights providing equally related SAM distance measures

Different weights for deletion, insertion and substitution operations will not always affect the relations between SAM distance measures. This may indicate that clustering by means of the SAM distance matrix, will provide equal results if relations between SAM distance measures are not changed. If, for different examples, the weights of operations are multiplied/divided by a constant factor, the SAM distance measures between sequences are also multiplied/divided by that constant factor. Examples are given in table 6.2. In the first (second), weights of operations in the second (first) divided (multiplied) by 50 are used. Likewise, SAM distance measures are multiplied (divided) by 50, indicating that, in each example, relations between SAM distance measures remain unchanged. If sequences are clustered based on SAM distance measures, the final clustering results of example one and two will be the same.

Source sequence: $s_1 = 1\ 2\ 3\ 4\ 7$ Target sequence: $s_2 = 2\ 1\ 3\ 5$				
Example	Parameters			$d_{SAM}(s_1, s_2)$
	w_d	w_i	w_s	
1	2	2	4	10
2	100	100	200	500

Table 6.2. Examples of operation weights providing equally related SAM distance measures.

6.1.3 Operation weights with some influence on relations between SAM distance measures

The relation between weights for deletion and insertion operations, irrelevant of the substitution weight, provides information about the influence of SAM's parameters on relations between SAM distance measures. If, within several examples, the deletion-to-insertion weights ratios are equal, different parameters will have some influence on relations between SAM distance measures. Likewise, if, within several examples, the weights for deletions are much larger (smaller) than the weights for insertions, different parameters will have some influence on relations between SAM distance measures. Examples are given in table 6.3. In examples one, two and three the deletion-to-insertion weights ratios are 0.5. In example four and five the deletion-to-insertion weights ratios are 1. The last three examples illustrate parameters where weights for deletions are much larger than weights for insertions. In the end, small changes in relations between SAM distance measures may produce small changes in the final results (i.e. clusters of server sessions). Practically this means that, although SAM distance between s_1 and s_2 equals 6 in the first example and 14 in the second, they may end up in the same clusters of example one and two because the relations between the SAM distance measures in the data set of example one and two has been changed only a little bit.

Source sequence: $s_1 = 1\ 2\ 3\ 4\ 7$ Target sequence: $s_2 = 2\ 1\ 3\ 5$				Examples using parameters with some influence on relations between SAM distance measures	
Example	Parameters				$d_{SAM}(s_1, s_2)$
	w_d	w_i	w_s		
1	1	2	2	6	1, 2, 3
2	2	4	6	14	
3	2	4	10	14	
4	2	2	2	8	4, 5
5	100	100	2	302	6, 7, 8
6	100	1	2	203	
7	100	4	104	308	
8	100	2	200	304	

Table 6.3. Examples of operation weights with some influence on relations between SAM distance measures.

6.1.4 Operation weights with more influence on relations between SAM distance measures

Finally, if for several examples, no uniform rule exists about the relation between operation weights for deletions and insertions or whether the weights for deletions are much larger (smaller) than the weights for insertions, different parameters will have more influence on relations between SAM distance measures. Examples are given in table 6.4. In example one and two, the weights given to deletions are larger than those given to insertions. However, in example one, deletion operations cost twice as much as insertion operations whereas in example two, deletion operations cost 100 times as much as insertion operations. In example three, the weight for insertion is larger than the weight for deletion. In the end, more changes in relations between SAM distance measures may produce more changes in the final results (i.e. clusters of server sessions). Practically this means that, the first, second and third example may end up with more differences in clustering results because the relations between SAM distance measures in the data set of example one, two and three has been changed more.

Source sequence: $s_1 = 1\ 2\ 3\ 4\ 7$ Target sequence: $s_2 = 2\ 1\ 3\ 5$				
Example	Parameters			$d_{SAM}(s_1, s_2)$
	w_d	w_i	w_s	
1	2	1	2	7
2	100	1	2	203
3	1	100	2	104

Table 6.4. Examples of operation weights with more influence on relations between SAM distance measures.

6.2 Parameterisation tests

Parameterisation tests apply different SAM parameters on a sampled set and examine the relations between SAM distance measures. In the sections that follow, we first list the goals of parameterisation tests. Then we provide an overview of the different operation weights that are invoked on a synthetic sampled set throughout the tests. The following section describes how the relation between SAM distance measures, resulting from different parameters, is examined. Finally, the results of the parameterisation tests are described, given the experiments and the sampled set.

6.2.1 Goals

The goals of parameterisation tests are the following:

- Examine whether the properties with regard to SAM, described in the previous section, are true.
- Examine whether our software utilities meet the requirements of SAM's properties.
- Examine the sensitivity of SAM or measuring the extent of changes in SAM parameters on relations between SAM distance measures.

6.2.2 Experiments and sampled set

The parameterisation tests are performed by different SAM experiments, specifying different operation weights for deletion (w_d), insertion (w_i) and substitution (w_s), presented in decision table 6.5. Based on the properties of SAM, which are illustrated in section 6.1, and on the levels of w_d , w_i and w_s , nine different categories of experimental tests are performed. Every category is identified by 'C' followed by an integer. Every experiment is identified by means of three underscored ('_') delimited integer values. The first integer refers to the weight for deletion (w_d), the second refers to the weight for insertion (w_i) and finally the third refers to the weight for substitution (w_s). For example, category C1 holds three experiments, which are designed to examine the first SAM property. The first, second and third experiment in C1 are identified by respectively 1_1_2, 1_1_3 and 1_1_100. Following SAM's properties, these operational weights provide equal SAM distance measures. We remark that this study examines the algorithm of SAM through studying

relational changes in SAM distance measures. Future research discusses how meaningful costs may be applied within Web Usage Mining studies by means of changes in SAM parameter settings.

The reason why categories and experiments of table 6.5 are chosen is given in the following paragraphs. C1 is designed to examine whether our software utilities meet the requirements of SAM's first property. Three instead of two experiments are identified in C1 in order to test the software's performance on the following characteristics:

- The ability to handle small and large numerical values given to operational weights
- The ability to handle two levels of parameter settings i.e. $w_s = w_d + w_i$ and $w_s > w_d + w_i$

The second level of parameter settings (i.e. $w_s < w_d + w_i$) is not relevant for the first SAM property since equal SAM distance measures do not occur for different operational weights if $w_s < w_d + w_i$. Also, equal SAM distance measures only occur if $w_d = w_i$ across different experiments.

C2 and C3 are designed to examine whether our software utilities meet the requirements of SAM's second property. Here, two levels of parameter settings (i.e. $w_s \geq w_d + w_i$ and $w_s < w_d + w_i$) are relevant and need to be examined. The only thing that matters for equally related SAM distance measures, using different operational weights, is to multiply or divide the weights of experiment a by x to obtain the weights of experiment b. This explains why in C2 equal weights are chosen (i.e. $w_{d1} = w_{i1}$ and $w_{d2} = w_{i2}$) and in C3 different weights are chosen (i.e. $w_{d1} \neq w_{i1}$ and $w_{d2} \neq w_{i2}$). Another explication for choosing this diversity in weights is to test the software's performance regarding equally related SAM distance measures for small and large integer values.

C4 to C9 are designed for two reasons. First, we examine whether the third and fourth properties of SAM are true. Second, we examine the influence or extent of changes in SAM parameters on relations between SAM distance measures. In other words, we specify categories of levels of influence and provide more detailed information based on changes in parameter settings. Both levels of parameter settings (i.e. $w_s \geq w_d + w_i$ and $w_s < w_d + w_i$) are relevant. C8 and C9 are distinguished from C4, C5, C6 and C7 based on different relations between weights for deletion and insertion operations. Furthermore, C4 and C5 specify three experiments, which are 2_1_3, 3_1_4, 100_1_101 in C4 and 2_1_2, 4_2_2, 100_1_2 in C5. The parameters in C4 are chosen in order to examine whether changing parameters 2_1_3 into 3_1_4 causes less relational changes compared to 100_1_101. For the same reason, parameters in C5 are chosen. We also expect that differences in SAM parameters will be quite small for experiments 2_1_2 and 4_2_2 in C5 because $w_{d1}/w_{i1} = w_{d2}/w_{i2}$, or $2/1 = 4/2$. Categories C6 and C7 present experiments

which are the inverse of those in C4 and C5. This means that the weights given to deletion and insertion operations are turned around to verify whether the same relational changes in SAM distance measures are observed under different circumstances: deletion > insertion and deletion < insertion. Also, experiment 1_2_3 is added in C8 in order to verify whether changing parameters 2_1_3 into 1_2_3 causes less relational changes compared to 1_100_101. For the same reason, experiment 1_2_2 is added in C9. Finally, weights are provided for small and large integer values.

Some of the weights were previously shown in the examples of tables 6.1 to 6.4 in section 6.1 to illustrate the SAM properties. For example, operational weights in C1 of decision table 6.5 are equal to the examples in table 6.1.

Levels of w_d, w_i, w_s	Sam Property				
	1	2	3		4
$w_d = w_i$		$w_{d1} =, >, < w_{i1}$ $w_{d2} =, >, < w_{i2}$ ($w_{d1} \neq w_{d2}$) ($w_{i1} \neq w_{i2}$)	$w_{d1} > w_{i1}$ $w_{d2} > w_{i2}$ ($w_{d1} \neq w_{d2}$)	$w_{d1} < w_{i1}$ $w_{d2} < w_{i2}$ ($w_{i1} \neq w_{i2}$)	$w_{d1} > w_{i1}$ and $w_{d2} < w_{i2}$
$w_s \geq w_d + w_i$	C1 1_1_2 1_1_3 1_1_100	C2 1_1_2 100_100_200	C4 2_1_3 3_1_4 100_1_101	C6 1_2_3 1_3_4 1_100_101	C8 2_1_3 1_2_3 1_100_101
$w_s < w_d + w_i$		C3 1_2_1 100_200_100	C5 2_1_2 4_2_2 100_1_2	C7 1_2_2 2_4_2 1_100_2	C9 2_1_2 1_2_2 1_100_2
Description of SAM properties 1 = Operation weights providing equal SAM distance measures 2 = Operation weights providing equally related SAM distance measures 3 = Operation weights with some influence on relations between SAM distance measures 4 = Operation weights with more influence on relations between SAM distance measures					

Table 6.5. Decision table presenting different categories of parameter settings in SAM. Each category holds different experiments.

The experiments presented in table 6.5, defining different SAM parameters, are used on a sampled, synthetic set with $N = 500$ sequences, representing server sessions. The number of items I , representing web pages, is 20 with $i =$

1, 2, ..., 20. For a given value of N and I, data is sampled in such a way that every integer i has more or less the same probability in the set. The average length (avg_seq_length), minimum length (min_seq_length) and maximum length (max_seq_length) of the sequences in the sampled set are respectively two, one and five. Figure 6.1 summarizes the algorithm that is used to sample the set. Details of the algorithm are given in appendix 6. To give an idea how the sampled set that is used throughout the parameterisation tests, looks like, table 6.6 shows the first ten and the last ten records.

```

begin
  define equal probability function for integer values i of or between 1 and 20
  for N:=1 to 50 do
    begin
      sample six sequences with length = 1
      sample one sequence with length = 2
      sample one sequence with length = 3
      sample one sequence with length = 4
      sample one sequence with length = 5
    end
  end;

```

Figure 6.1: Summarized algorithm for sampled set.

Number of record	Sequence (representing server session)
1	1
2	1
3	18
4	5
5	6
6	14
7	7 4
8	8 9 2
9	10 2 17 2
10	6 19 8 16 7
...	...
491	14
492	7
493	12
494	11
495	9
496	3
497	11 16

498	15 5 9
499	20 16 17 9
500	1 19 11 10 10

Table 6.6: First and last ten records, representing server sessions, of synthetic sampled set (created by algorithm in figure 6.1), used throughout parameterisation tests.

6.2.3 Correlation and dissimilarity

A good method for *measuring the relation between SAM distance measures*, resulting from different experiments, is *Pearson's correlation (c)* (Kaufman and Rousseeuw, 1990), which looks for a linear relation between two variables x and y as follows:

$$c(x, y) = \frac{\sum_{j=1}^n (x_j - \text{avg}_x) (y_j - \text{avg}_y)}{\sqrt{[\sum_{j=1}^n (x_j - \text{avg}_x)^2]} \sqrt{[\sum_{j=1}^n (y_j - \text{avg}_y)^2]}} \quad (6.1)$$

where

x represents all of the pair-wise SAM distance measures between sequences in the analysis, based on operation weights of experiment a or d_{SAM}^a ;

y represents all of the pair-wise SAM distance measures between sequences in the analysis, based on operation weights of experiment b or d_{SAM}^b ;

j identifies the pair-wise SAM distance measures;

n is the total number of pair-wise SAM distance measures between sequences in the analysis and equals $[N \times (N - 1)] / 2$;

avg_x is the average SAM distance measure of x ;

avg_y is the average SAM distance measure of y ;

and

a identifies the first argument in the correlation, $a = 1, 2, \dots, E$;

b identifies the second argument in the correlation, $b = a + 1$;

E is the total number of different experiments in the parameterisation tests;

N is the total number of sequences in the analysis;

Pearson's correlation lies between -1 and +1 and is useful for clustering purposes because the extent to which two variables are related is measured (Kaufman and Rousseeuw, 1990). On the one hand, a perfect positive correlation between two variables is indicated by $c = 1$ and means that a high/low value of the first variable occurs with a high/low value of the second variable. On the other hand, a perfect negative correlation between two variables is indicated by $c = -1$ and means that, a high/low value of the first variable occurs with a low/high value of the second variable. Interpretations for c are given in table 6.7 (Texas A&M University, 2003; SPSS Tutorial, 2003; Northwest Missouri State University, 2003; Biz/ed, 2003).

Strength of correlation	Meaning
$0.8 \leq c \leq 1$	Strong
$0.5 < c < 0.8$	Moderate
$ c \leq 0.5$	Weak

Table 6.7: Interpreting Pearson's correlation.

In Kaufman and Rousseeuw (1990), correlations are converted to *dissimilarities* as follows:

$$\text{dissimilarity}(x, y) = [1 - c(x, y)] / 2 \quad (6.2)$$

Dissimilarities lie between 0 and 1 and *measure the extent of changes in relations between SAM distance measures*, due to differences in SAM parameters. Using equation (6.2), variables (representing SAM distance measures resulting from different experiments) with a high positive correlation receive dissimilarity close to 0, whereas variables with a small positive correlation receive dissimilarity close to 0.5. With regard to negative correlations, dissimilarities lie between 0.5 and 1. The dissimilarity is particularly useful to indicate the effect of negatively correlated variables on the final clustering results. In the extreme case, a dissimilarity of 1 (correlation = -1) between SAM distance measures resulting from two experiments will end up with huge differences in clustering results since, for each sequence pair, small/large SAM distance measures in one experiment become large/small SAM distance measures in the other. Following, sequences with small distance measures in one experiment are grouped together during clustering based on the SAM distance matrix. Yet, the same sequences that were grouped together will end up in different clusters in the other experiment since their distance

measures in the SAM distance matrix are large. For this reason, negative correlations are always associated with high dissimilarities.

Along with interpretations of c , provided in table 6.7, corresponding interpretations of the dissimilarities are given in table 6.8. The last column provides information about the magnitude of changes in relations between SAM distance measures (no, minor, considerable or major), which may indicate the effect on the expected, final clustering results.

Dissim (x, y)	C (x, y)	Meaning	Magnitude of changes in relations between SAM distance measures
0	1	Sam distance measures are similar and perfectly correlated.	Zero
> 0 and ≤ 0.1	≥ 0.8 and < 1	SAM distance measures are nearly similar and strongly correlated.	Minor
> 0.1 and < 0.25	> 0.5 and < 0.8	SAM distance measures are moderately similar and moderately correlated.	Considerable
≥ 0.25	≤ 0.5	SAM distance measures are dissimilar and weakly correlated.	Major

Table 6.8: Influence of SAM parameters on relations between SAM distance measures.

6.2.4 Results

Using the sampled set in table 6.6 and the SAM parameters given in the experiments of table 6.5, correlations and dissimilarities are calculated between SAM distance measures of every pair of experiments within each category (C1 to C9), using equations (6.1) and (6.2). The results are given in table 6.9. In the first column, categories are given along with pairs of experiments in the second column. In the third and fourth column, the dissimilarity and correlation between SAM distance measures for each corresponding pair of experiments in the second column, is given. Finally, in the last column, sensitivity is shown, using the different categories of the dissimilarities provided in table 6.8. *Sensitivity* of changes in SAM parameter settings expresses the magnitude of changes in relations between SAM distance measures.

Category	Experiments	Dissimilarity	Correlation	Sensitivity
C1	1 1 2 and 1 1 3	0	1	Zero
C1	1 1 2 and 1 1 100	0	1	Zero
C1	1 1 3 and 1 1 100	0	1	Zero
C2	1 1 2 and 100 100 200	0	1	Zero
C3	1 2 1 and 100 200 100	0	1	Zero
C4	2 1 3 and 3 1 4	0.0055	0.989	Minor
C4	2 1 3 and 100 1 101	0.055	0.89	Minor
C4	3 1 4 and 100 1 101	0.0255	0.949	Minor
C5	2 1 2 and 4 2 2	0.005	0.99	Minor
C5	2 1 2 and 100 1 2	0.06	0.88	Minor
C5	4 2 2 and 100 1 2	0.055	0.89	Minor
C6	1 2 3 and 1 3 4	0.0055	0.989	Minor
C6	1 2 3 and 1 100 101	0.053	0.893	Minor
C6	1 3 4 and 1 100 101	0.025	0.95	Minor
C7	1 2 2 and 2 4 2	0.005	0.99	Minor
C7	1 2 2 and 1 100 2	0.055	0.89	Minor
C7	2 4 2 and 1 100 2	0.053	0.893	Minor
C8	2 1 3 and 1 2 3	0.1165	0.767	Considerable
C8	2 1 3 and 1 100 101	0.301	0.398	Major
C9	2 1 2 and 1 2 2	0.12	0.76	Considerable
C9	2 1 2 and 1 100 2	0.305	0.39	Major

Table 6.9: dissimilarities and correlations between SAM distance measures for sampled set presented in table 6.6.

First, we examine whether the properties with regard to the SAM algorithm, given in section 6.1, are true. We also verify whether the software utilities meet SAM's requirements and handle properly large numerical values given to the operational weights. Table 6.9 provides the following results. The superscripts of d_{SAM} indicate the experimental id (re. table 6.5).

- $\text{dissim}(d_{\text{SAM}}^{1-1-2}, d_{\text{SAM}}^{1-1-3}) = 0$; $\text{dissim}(d_{\text{SAM}}^{1-1-2}, d_{\text{SAM}}^{1-1-100}) = 0$; $\text{dissim}(d_{\text{SAM}}^{1-1-3}, d_{\text{SAM}}^{1-1-100}) = 0$; $\text{dissim}(d_{\text{SAM}}^{1-1-2}, d_{\text{SAM}}^{100-100-200}) = 0$; $\text{dissim}(d_{\text{SAM}}^{1-2-1}, d_{\text{SAM}}^{100-200-100}) = 0$, indicating that the first and second property in table 6.5 are true. Moreover, large (i.e. 100, 200) as well as small (i.e. 1, 2, 3) integers are handled properly by SAM's software.
- Dissimilarities between SAM distance measures for experiments in C4, C5, C6 and C7 are generally smaller than those in C8 and C9. For example, $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{3-1-4}) = 0.0055$ whereas $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{1-2-3}) = 0.1165$. Likewise, $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{100-1-101}) = 0.055$ whereas $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{1-100-101}) = 0.301$. This means that the third and fourth

property in table 6.5 are true. Moreover, large (i.e. 100, 101) as well as small (i.e. 1, 2, 3, 4) integers are handled properly by SAM's software.

Second, we examine whether small/large differences in parameter settings in C4 to C9 provide small/large differences in relations between SAM distance measures. In table 6.9, the following dissimilarities are found.

- C4: $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{3-1-4}) < \text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{100-1-101})$
- C5: $\text{dissim}(d_{\text{SAM}}^{2-1-2}, d_{\text{SAM}}^{4-2-2}) < \text{dissim}(d_{\text{SAM}}^{2-1-2}, d_{\text{SAM}}^{100-1-2})$
- C6: $\text{dissim}(d_{\text{SAM}}^{1-2-3}, d_{\text{SAM}}^{1-3-4}) < \text{dissim}(d_{\text{SAM}}^{1-2-3}, d_{\text{SAM}}^{1-100-101})$
- C7: $\text{dissim}(d_{\text{SAM}}^{1-2-2}, d_{\text{SAM}}^{2-4-2}) < \text{dissim}(d_{\text{SAM}}^{1-2-2}, d_{\text{SAM}}^{1-100-2})$
- C8: $\text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{1-2-3}) < \text{dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{1-100-101})$
- C9: $\text{dissim}(d_{\text{SAM}}^{2-1-2}, d_{\text{SAM}}^{1-2-2}) < \text{dissim}(d_{\text{SAM}}^{2-1-2}, d_{\text{SAM}}^{1-100-2})$

Third, we verify whether changing parameters in such a way that $w_{d1}/w_{i1} = w_{d2}/w_{i2}$ produces less changes in relations between SAM distance measures compared to circumstances where $w_{d1}/w_{i1} \neq w_{d2}/w_{i2}$. This is illustrated by the following experiments in table 6.9.

- C5: $\text{dissim}(d_{\text{SAM}}^{2-1-2}, d_{\text{SAM}}^{4-2-2}) < \text{C4: dissim}(d_{\text{SAM}}^{2-1-3}, d_{\text{SAM}}^{3-1-4})$
- C7: $\text{dissim}(d_{\text{SAM}}^{1-2-2}, d_{\text{SAM}}^{2-4-2}) < \text{C6: dissim}(d_{\text{SAM}}^{1-2-3}, d_{\text{SAM}}^{1-3-4})$

We remark that these results are dependent on the sampled set and may be influenced by randomness. In order to provide more general results about the effect of changes in SAM parameters on (relations between) SAM distance measures and on final clustering results, parameterisation tests are executed on a real data set in section 6.4.

Given the experiments specifying different SAM parameters (re. table 6.5) and given the sampled set (re. table 6.6) used for parameterisation tests, the following rules may be deduced. In order to know which rule will be applied to a group of parameter settings, changes in operational weights are examined starting with the first three rules. If operational weights do not match with rule 1, 2 or 3, rule 4 is analysed. If again, no match is found, the analysis proceeds to rule 5.

If, for two experiments of SAM parameter settings, w_{d1} , w_{i1} , w_{s1} are operational weights in experiment 1 and w_{d2} , w_{i2} , w_{s2} are operational weights in experiment 2:

1. $w_{d1} = w_{d2} = w_{i1} = w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$, $w_{s2} \geq w_{d2} + w_{i2}$
then experiment 1 and 2 will provide *equal* SAM distance measures and the influence of different parameters on relations between SAM distance measures will be *zero*.
2. $w_{d1} = w_{d2} \neq w_{i1} = w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$, $w_{s2} \geq w_{d2} + w_{i2}$
then experiment 1 and 2 will provide *equal* SAM distance measures and the influence of different parameters on relations between SAM distance measures will be *zero*.
3. $w_{d1} = x w_{d2}$ and $w_{i1} = x w_{i2}$ and $w_{s1} = x w_{s2}$ with $x = 2, 3, \dots, \infty$
then experiment 1 and 2 will provide *equal* relations between SAM distance measures and the influence of different parameters on relations between SAM distance measures will be *zero*..

Else if

4. $w_{d1} / w_{i1} = w_{d2} / w_{i2}$ and $w_{s1} < w_{d1} + w_{i1}$ and $w_{s2} < w_{d2} + w_{i2}$ or
 $w_{d1} > w_{i1}$ and $w_{d2} > w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$ and $w_{s2} \geq w_{d2} + w_{i2}$ or
 $w_{d1} > w_{i1}$ and $w_{d2} > w_{i2}$ and $w_{s1} < w_{d1} + w_{i1}$ and $w_{s2} < w_{d2} + w_{i2}$ or
 $w_{d1} < w_{i1}$ and $w_{d2} < w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$ and $w_{s2} \geq w_{d2} + w_{i2}$ or
 $w_{d1} < w_{i1}$ and $w_{d2} < w_{i2}$ and $w_{s1} < w_{d1} + w_{i1}$ and $w_{s2} < w_{d2} + w_{i2}$
then the influence of different parameters on relations between SAM distance measures will be *minor*

Else if

5. $w_{d1} / w_{i1} \neq w_{d2} / w_{i2}$ and $w_{d1} > w_{i1}$ and $w_{d2} < w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$ and
 $w_{s2} \geq w_{d2} + w_{i2}$ or
 $w_{d1} / w_{i1} \neq w_{d2} / w_{i2}$ and $w_{d1} > w_{i1}$ and $w_{d2} < w_{i2}$ and $w_{s1} < w_{d1} + w_{i1}$ and
 $w_{s2} < w_{d2} + w_{i2}$ or
 $w_{d1} / w_{i1} \neq w_{d2} / w_{i2}$ and $w_{d1} < w_{i1}$ and $w_{d2} > w_{i2}$ and $w_{s1} \geq w_{d1} + w_{i1}$ and
 $w_{s2} \geq w_{d2} + w_{i2}$ or
 $w_{d1} / w_{i1} \neq w_{d2} / w_{i2}$ and $w_{d1} < w_{i1}$ and $w_{d2} > w_{i2}$ and $w_{s1} < w_{d1} + w_{i1}$ and
 $w_{s2} < w_{d2} + w_{i2}$
then the influence of different parameters on relations between SAM distance measures will be *considerable* or *major*

For example, examining operational weights in experiments 2_1_3 and 3_1_4, rule 4 predicts that the influence on relations between SAM distance measures will be minor. Yet, considering the operational weights in experiments 2_1_2 and 1_2_2, rule 5 predicts that, if $w_{d1} = 2$, $w_{i1} = 1$, $w_{s1} = 2$ are changed into $w_{d2} = 1$, $w_{i2} = 2$, $w_{s2} = 2$, relations between SAM distance measures will at least be considerably influenced. In experiment 2_1_2, the weight for deletion is larger than insertion whereas in experiment 1_2_2, the weight for deletion is smaller than insertion.

Section 6.4 examines whether these rules might be generalized for a real web usage data set. Before proceeding to section 6.4, we first examine whether particular characteristics of sampled sets might cause differences in the magnitude of changes in relations between SAM distance measures.

6.3 Characteristics of sampled sets

Sampled sets may differ in total number of distinct items in the set and in average sequence length. Instead of applying different SAM parameters on one sampled set, one set of parameters is applied to different sampled sets and the relations between SAM distance measures resulting from different sampled sets are analysed. In the sections that follow, we first list the goals of the tests. Then we provide an overview of different sampled sets that are used throughout the tests. The following subsection describes how the relation between SAM distance measures, resulting from different sampled sets, is analysed. Finally, the results of the tests are described.

6.3.1 Goals

The goals of the tests are the following:

- Examine whether SAM is sensitive to the total number of items or average sequence length in the sampled set.

6.3.2 Experiments and sampled sets

The parameter settings that are used throughout the tests are given by experiment 1_1_2 (C1) of table 6.5. Also, in order to verify that the results did not occur by chance, two more experiments are used throughout the tests: experiment 100_1_2 and 2_4_10. Experiments 1_1_2, 100_1_2 and 2_4_10 are

chosen because of the variety in parameter settings. In experiment 1_1_2, $w_d = w_i$ and $w_s = w_d + w_i$. In experiment 100_1_2, $w_d > w_i$ and $w_s < w_d + w_i$. Finally, in experiment 2_4_10, $w_d < w_i$ and $w_s > w_d + w_i$.

Table 6.10 describes six sampled sets that are used within the tests. The sampled sets are synthetic with $N = 500$ sequences, representing server sessions. The number of items I , representing pages, is 20, with $i = 1, 2, \dots, 20$ for the first three sampled sets. The number of items I is 100, with $i = 1, 2, \dots, 100$ for the last three sampled sets. For a given value of N and I , data is sampled in such a way that every integer i has approximately the same probability in the sampled sets. The average length (*avg_seq_length*), minimum length (*min_seq_length*) and maximum length (*max_seq_length*) of the sequences in the sampled sets are given in the fourth, fifth and sixth column of table 6.10. This means that, besides I , sensitivity of SAM is also examined with regard to sequences of relatively short lengths (20_2_5_1 and 100_2_5_1), long lengths (20_12_15_10 and 100_12_15_10) and sequences of both short and long lengths (20_7_15_1 and 100_7_15_1). Note that each sampled set is identified by means of four underscored ('_') delimited integer values. The first integer refers to I . The following integers refer to average, maximum and minimum sequence length in the sampled set.

Id	N	I	avg_ seq_length	max_ seq_length	min_ seq_length
20_2_5_1	500	20	2	5	1
20_12_15_10	500	20	12	15	10
20_7_15_1	500	20	7	15	1
100_2_5_1	500	100	2	5	1
100_12_15_10	500	100	12	15	10
100_7_15_1	500	100	7	15	1

Table 6.10. Synthetic sampled sets.

6.3.3 Correlation and dissimilarity

Using equation (6.1), Pearson's correlation (c) is calculated between two variables v and w where

- v represents all of the pair-wise SAM distance measures between sequences in the analysis of data set p , based on operation weights of experiment a or d_{SAM}^a ;
- w represents all of the pair-wise SAM distance measures between sequences in the analysis of data set q , based on operation weights of experiment a or d_{SAM}^a ;

and

- a identifies the experiment and the weights used by SAM;
- p identifies the first data set in the correlation, $p = 1, 2, \dots, F$;
- q identifies the second data set in the correlation, $q = p + 1$;
- F is the total number of different data sets in the tests;

Characteristics and interpretations of Pearson's correlation are described in section 6.2.3. Likewise, correlations are converted to dissimilarities using equation (6.2). Interpretations for *dissimilarities* are as follows. Dissimilarities lie between 0 and 1 and *measure the influence of changes in sampled sets, particularly with regard to the number of items and differences in sequence length, on relations between SAM distance measures*. Also, table 6.8 is used throughout the tests. Yet, instead of measuring the influence of SAM parameters, the influence of the sampled set is measured on SAM distance measures.

6.3.4 Results

Using the sampled sets of table 6.10 and the SAM parameters of experiments 1_1_2, 100_1_2 and 2_4_10, for each experiment, the correlations and dissimilarities are calculated between the SAM distance measures of every pair of sampled sets, using equations (6.1) and (6.2). Sensitivity tables are calculated using table 6.8. The results for experiment 1_1_2 are given in tables 6.11, 6.12 and 6.13. A *dissimilarity table* provides the dissimilarity between SAM distance measures for each pair of sampled sets. In a *correlation table* the value of c between SAM distance measures for each pair of sampled sets is given. Because the tables are symmetric, the lower half of the cells in the tables are marked with upward diagonal lines, so as to avoid repeating information. Dissimilarity, correlations and sensitivities of experiments 100_1_2 and 2_4_10 are given in appendix 6.

Experiment 1_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	0	0.6615	0.6585	0.6865	0.189	0.739
	20 12 15 10		0	0.3625	0.4525	0.5275	0.5495
	20 7 15 1			0	0.235	0.7115	0.2695
	100 2 5 1				0	0.666	0.321
	100 12 15 10					0	0.6805
	100 7 15 1						0

Table 6.11: Dissimilarity table for experiment 1_1_2.

Experiment 1_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	1	-0.323	-0.317	-0.373	0.622	-0.478
	20 12 15 10		1	0.275	0.095	-0.055	-0.099
	20 7 15 1			1	0.53	-0.423	0.461
	100 2 5 1				1	-0.332	0.358
	100 12 15 10					1	-0.361
	100 7 15 1						1

Table 6.12: Correlation table for experiment 1_1_2.

Experiment 1_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	ZERO	MAJOR	MAJOR	MAJOR	CONS	MAJOR
	20 12 15 10		ZERO	MAJOR	MAJOR	MAJOR	MAJOR
	20 7 15 1			ZERO	CONS	MAJOR	MAJOR
	100 2 5 1				ZERO	MAJOR	MAJOR
	100 12 15 10					ZERO	MAJOR
	100 7 15 1						ZERO

Table 6.13: Sensitivity table for experiment 1_1_2.

Dissimilarities do not exclusively measure differences between sampled sets with regard to the number of items I and sequence length. However, due to approximate equal probabilities of the items i in the sampled sets, dissimilarities between sampled sets are mostly due to differences in the number of items I and differences in sequence length. Yet, care must be taken to interpret dissimilarities, correlations as well as sensitivities between pairs of sampled sets. Given the sampled sets of table 7.10 and experiments 1_1_2, 100_1_2 and 2_4_10, the following results are obtained from the dissimilarity, correlation and sensitivity tables:

- Differences BETWEEN sensitivity tables of different experiments are quite small, which argues for stable results. In table 6.14, the characters printed in bold show different values regarding sensitivities across different experiments. Comparing experiment 1_1_2 with 100_1_2, three cell values are different. Comparing experiment 1_1_2 with 2_4_10, one cell value is different. Yet, comparing experiment 100_1_2 with 2_4_10, four cell values are different. Different cell values across different sensitivity tables are due to different parameter settings. For example, the sensitivity of SAM distance measures from sampled sets 20_7_15_1 and 100_2_5_1 in experiments 1_1_2 and 100_1_2 is less than in experiment 2_4_10. The reason for this difference might be that less deletions and more insertions in both sampled sets 20_7_15_1 and 100_2_5_1 occur. In experiments 1_1_2 and 100_1_2 operation weights for insertions are equal to one. Yet, in experiment 2_4_10 operation weight for insertion equals four, which means that more fluctuations occur in relations between SAM distance measures.
- Likewise, differences WITHIN sensitivity tables of different experiments are quite small. Most of the cell values show major changes in relations between SAM distance measures. This means that both I and sequence length have an influence on changes in relations between SAM distance measures.

1 = Experiment 1_1_2 2 = Experiment 100_1_2 3 = Experiment 2_4_10 Z = zero C = considerable M = major		S A M P L E D S E T																	
		20_2_5_1			20_12_15_10			20_7_15_1			100_2_5_1			100_12_15_10			100_7_15_1		
SAMPLED SET	20_2_5_1	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
		Z	Z	Z	M	M	M	M	M	M	M	M	M	C	M	C	M	M	M
	20_12_15_10				Z	Z	Z	M	M	M	M	M	M	M	M	M	M	M	M
	20_7_15_1							Z	Z	Z	C	C	M	M	M	M	M	C	M
	100_2_5_1										Z	Z	Z	M	M	M	M	C	M
	100_12_15_10													Z	Z	Z	M	M	M
100_7_15_1																Z	Z	Z	

Table 6.14: Merging three sensitivity tables for experiments 1_1_2, 100_1_2 and 2_4_10.

Tables 6.15 and 6.16 provide an overview of the number of cases (i.e. number of cells above the diagonal within sensitivity table 6.14) with considerable (table 6.15) and major (table 6.16) differences in relations between SAM distance measures of sampled sets with equal or different I (columns) and equal or different sequence length (rows). For example, table 6.15 indicates that, out of six cases showing *considerable* differences in relations between SAM distance measures of two different sampled sets:

- Zero cases were found with equal I and equal sequence length (cell in first column and first row)
- One case is found with different I and equal sequence length (cell in second column and first row)
- One case is found with equal I and different sequence length (cell in first column and second row)
- Four cases were found with different I and different sequence length (cell in second column and second row)

Table 6.16 indicates that, out of 39 cases showing *major* differences in relations between SAM distance measures of two different sampled sets:

- Zero cases were found with equal I and equal sequence length (cell in first column and first row)
- Eight cases are found with different I and equal sequence length (cell in second column and first row)
- Seventeen cases are found with equal I and different sequence length (cell in first column and second row)
- Fourteen cases are found with different I and different sequence length (cell in second column and second row)

Characteristics of compared sampled sets with <i>considerable</i> influence on SAM distance measures	Equal I	Different I	TOTAL
Equal sequence length	0	1	1
Different sequence length	1	4	5
TOTAL	1	5	6

Table 6.15: Number of cases with considerable differences in relations between SAM distance measures distinguishing sampled sets with different/equal I and/or different/equal sequence length.

Characteristics of compared sampled sets with <i>major</i> influence on SAM distance measures	Equal I	Different I	TOTAL
Equal sequence length	0	8	8
Different sequence length	17	14	31
TOTAL	17	22	39

Table 6.16: Number of cases with major differences in relations between SAM distance measures distinguishing sampled sets with different/equal I or/or different/equal sequence length.

We remark that, occasionally, SAM may be more or less sensitive than intuitively expected. For example, in table 6.11, for sampled sets 20_2_5_1 and 100_12_15_10, showing differences in I as well as in sequence length, a relatively low dissimilarity measure of 0.189 is given. Yet, for sampled sets 20_2_5_1 and 100_2_5_1, showing differences in I only, a relatively high dissimilarity measure of 0.6865 is given. Likewise, for sampled sets 20_2_5_1 and 20_12_15_10, showing differences in sequence length only, a relatively high dissimilarity measure of 0.6615 is given. However, we would expect that, the more sampled sets differ, the more relations between SAM distance measures change resulting in a higher dissimilarity measure.

We provide an example how, by chance, relations between SAM distance measures do not change the way we expect. Consider the first five sequences of sampled sets 20_2_5_1, 100_2_5_1, 20_12_15_10 and 100_12_15_10, holding respectively sequences of relatively short lengths and long lengths, shown in table 6.17. The lower half of the table provides the first five SAM distance measures for each sampled set using the following operational weights: 1 for deletion, 1 for insertion and 2 for substitution. Table 6.18 provides Pearson's correlations as well as dissimilarities. The highest dissimilarity is shown for SAM distance measures of sampled sets 20_2_5_1 and 100_2_5_1, both

holding sequences of the same lengths but different I. Unexpected, the lowest dissimilarity (i.e. highest similarity!) is shown for SAM distance measures of sampled sets 20_2_5_1 and 100_12_15_10, holding sequences of different lengths and different I.

Sampled set			
20_2_5_1	100_2_5_1	20_12_15_10	100_12_15_10
$s_1 = 4$	$s_1 = 6 \ 9 \ 10$ 104	$s_1 = 2 \ 1 \ 20 \ 19 \ 3 \ 10 \ 9$ 8 5 12 18 5 5 6 3	$s_1 = 22 \ 72 \ 28 \ 21 \ 55 \ 54$ 71 99 100 22
$s_2 = 1$	$s_2 = 77$	$s_2 = 4 \ 3 \ 20 \ 19 \ 18 \ 17$ 5 2 1 6 9 6 7 3 2	$s_2 = 31 \ 61 \ 44 \ 72 \ 89 \ 80$ 39 29 74 21
$s_3 = 10$	$s_3 = 20$	$s_3 = 8 \ 8 \ 3 \ 2 \ 1 \ 11 \ 19$ 17 16 20 9 8 7 2 15	$s_3 = 42 \ 52 \ 76 \ 93 \ 60 \ 40$ 28 15 39 54
$s_4 = 18$	$s_4 = 1 \ 99 \ 54$	$s_4 = 11 \ 14 \ 16 \ 18 \ 17$ 17 2 4 1 6	$s_4 = 93 \ 82 \ 49 \ 29 \ 73 \ 48$ 91 1 20 8
$s_5 = 2 \ 8 \ 9$	$s_5 = 44$	$s_5 = 5 \ 17 \ 18 \ 11 \ 12 \ 4 \ 2$ 9 8 16 16 13 12	$s_5 = 91 \ 75 \ 88 \ 29 \ 1 \ 29$ 100 82 6 4 22 77 18 54 63
SAM distance measures			
$d(s_1, s_2) = 2$	$d(s_1, s_2) = 5$	$d(s_1, s_2) = 18$	$d(s_1, s_2) = 10$
$d(s_1, s_3) = 2$	$d(s_1, s_3) = 5$	$d(s_1, s_3) = 20$	$d(s_1, s_3) = 10$
$d(s_1, s_4) = 2$	$d(s_1, s_4) = 7$	$d(s_1, s_4) = 19$	$d(s_1, s_4) = 10$
$d(s_1, s_5) = 4$	$d(s_1, s_5) = 5$	$d(s_1, s_5) = 20$	$d(s_1, s_5) = 15$

Table 6.17: First five sequences and SAM distance measures in sampled sets 20_2_5_1, 100_2_5_1, 20_12_15_10 and 100_12_15_10.

First five SAM distance measures coming from sampled sets	Correlation	Dissimilarity
20_2_5_1 and 100_2_5_1	-0.333	0.6665
20_2_5_1 and 20_12_15_10	0.522	0.239
20_2_5_1 and 100_12_15_10	1	0

Table 6.18: Correlations and dissimilarities between first five SAM distance measures from different sampled sets.

6.4 Parameterisation tests on real data set

In this section we examine how changes in SAM parameters influence the final clustering results on a real data set. To this end we investigate the effect of *minor*, *considerable* or *major* changes in relations between SAM distance measures, resulting from different operational weights, on the final clustering results. The data set is presented in chapter four, providing visiting behaviour on the web site of a Belgian telecom provider. The operational weights are given in table 6.19. We remark that, in order to test whether the rules, stated in section 6.2.4 and deduced from parameterisation tests described by experiments in table 6.5, may be generalized to a real data set, the operational weights given in table 6.19 are different from those provided earlier in table 6.5. For example, rule 4 in section 6.2.4 predicts that the influence of different parameters on relations between SAM distance measures will be minor if the three previous rules do not hold and if $w_{d1} > w_{i1}$, $w_{d2} > w_{i2}$, $w_{s1} \geq w_{d1} + w_{i1}$ and $w_{s2} \geq w_{d2} + w_{i2}$. This is illustrated by experiments 5_3_8 and 7_6_13.

Experiment identification	w_d	w_i	w_s	Predicted influence on relations between SAM distance measures
5_3_8	5	3	8	Minor
7_6_13	7	6	13	
4_7_11	4	7	11	Considerable or Major
7_4_11	7	4	11	
2_6_3	2	6	3	Considerable or Major
150_4_3	10	6	3	

Table 6.19: Experiments specifying different SAM parameters and predicting influence on relations between SAM distance measures.

The operational weights, given by six different experiments of table 6.19, are used by SAM while measuring SAM distances between every pair of server sessions in the data set. In each experiment, a total number of $(773 \times 772) / 2 = 298,378$ SAM distance measures are calculated. Table 6.20 shows dissimilarities, correlations and sensitivities between SAM distance measures resulted from different experiments. For example, with regard to the data set of a Belgian Telecom provider, the correlation between SAM distance measures of experiments 5_3_8 and 7_6_13 equals 0.88, indicating a strong relation (re. table 6.7). Using equation (6.2), the dissimilarity equals 0.06, indicating that SAM distance measures are nearly similar (re. table 6.8). Furthermore, predictions are made about the influence of changing operational weights $w_d =$

5, $w_i = 3$, $w_s = 8$ into $w_d = 7$, $w_i = 6$, $w_s = 13$. Changes in relations between SAM distance measures, also called the sensitivity of SAM distance measures towards changes in operational weights, will be minor. Finally, the prediction rules about the influence of changes in operational weights on relations between SAM distance measures, stated in section 6.2.4, are true and may be generalized to a real data set of web usage behaviour on the web site of a Belgian Telecom Provider.

Data set: Belgian Telecom Provider		Dissimilarity	Correlation	Sensitivity
Experiments	$\frac{5}{7} \frac{3}{6} \frac{8}{13}$	0.06	0.88	Minor
	$\frac{4}{7} \frac{7}{4} \frac{11}{11}$	0.175	0.65	Considerable
	$\frac{2}{150} \frac{6}{4} \frac{3}{3}$	0.29	0.42	Major

Table 6.20: Dissimilarities, correlations and sensitivities between SAM distance measures using different operational weights on a data set representing web usage behaviour on the web site of a Belgian Telecom Provider.

In order to examine how the final clustering results will change due to changes in operational weights in the SAM algorithm, distance matrices are constructed for each experiment. A distance matrix is described in the second (i.e. processing) step of our approach of web usage mining process in chapter four. A total number of six distance matrices are constructed and, conform to our approach of web usage mining, Ward hierarchical clustering is invoked on each distance matrix. The number of clusters is based on information criteria. We remark that definitions of information criteria for defining the number of clusters are given in chapter four. Table 6.21 provides, for each experiment, the number of clusters. In table 6.22, the number of server sessions in each cluster, for each experiment, is given.

Experiment	Total number of clusters
5_3_8	4
7_6_13	5
4_7_11	5
7_4_11	5
2_6_3	5
150_4_3	5

Table 6.21: Total number of clusters in 6 different experiments.

Experiment	Cluster					TOT
	1	2	3	4	5	
5_3_8	242	75	178	278	-	773
7_6_13	220	98	170	264	21	773
4_7_11	244	198	160	48	123	773
7_4_11	238	225	56	112	142	773
2_6_3	246	188	170	113	56	773
150_4_3	238	244	156	78	57	773

Table 6.22: Clustering 773 server sessions in 6 different experiments.

The clustering results of six different experiments, using different operational weights on the same data set of a Belgian telecom provider, are compared by means of equality tables. Tables 6.23 to 6.25 present an *equality table*, showing equalities between server sessions grouped in each cluster for each pair of experiments, given in table 6.20. The objective of using equality tables is to examine to what extent changes in operational weights influence the final clustering results. In other words, how many server sessions are clustered differently when the influence of changes in operational weights on the relation between SAM distance measures is ‘minor’, ‘considerable’ or ‘major’ (re. prediction rules of section 6.2.4)?

Tables 6.23 to 6.25 are constructed as follows. The total number of server sessions in the analysis are written in the corner at the bottom right of each table. The last column shows the number of server sessions in each cluster of the experiment that is presented vertically, in the first column. The last row shows the number of server sessions in each cluster of the experiment that is presented horizontally, in the first row. For example, table 6.23 compares equal server sessions in clusters of experiments 5_3_8 and 7_6_13. Out of 242 server sessions in cluster one of experiment 5_3_8, 215 are grouped in cluster one of experiment 7_6_13. Of the remaining server sessions, 8 are grouped in cluster two, 7 in cluster four and 12 in cluster five of experiment 7_6_13. The cells printed in bold show the maximum number of server sessions (in each row) that are ‘*equally grouped together*’, which means that, for two different

experiments, the same server sessions are grouped together irrelevant of the fact whether they are assigned, as a group, to the same type of clusters. 'Equally clustered' means that, for two different experiments, the same server sessions are grouped together and are assigned to the same type of clusters. For example, comparing the final clustering results of experiments 4_7_11 and 7_4_11 (re. table 6.24), 645 out of 773 server sessions of experiment 4_7_11 are equally grouped together in experiment 7_4_11. Yet, 625 out of 773 server sessions of experiment 4_7_11 are equally clustered in experiment 7_4_11. Finally, comparing the final clustering results of experiments 5_3_8 and 7_6_13 (re. table 6.23), 4_7_11 and 7_4_11 (re. table 6.24), 2_6_3 and 150_4_3 (re. table 6.25), respectively 91.72%, 80.85% and 76.84% of the server sessions are equally clustered.

Experiment 5_3_8: 4 clusters	Experiment 7_6_13: 5 clusters					TOT
	1	2	3	4	5	
1	215	8	0	7	12	242
2	5	70	0	0	0	75
3	0	11	167	0	0	178
4	0	9	3	257	9	278
TOT	220	98	170	264	21	773

Table 6.23: Equality table comparing experiment 5_3_8 with 7_6_13.

Experiment 4_7_11: 5 clusters	Experiment 7_4_11: 5 clusters					TOT
	1	2	3	4	5	
1	238	0	6	0	0	244
2	0	62	0	0	136	198
3	0	159	1	0	0	160
4	0	0	14	34	0	48
5	0	4	35	78	6	123
TOT	238	225	56	112	142	773

Table 6.24: Equality table comparing experiment 4_7_11 with 7_4_11.

Experiment 2_6_3: 5 clusters	Experiment 150_4_3: 5 clusters					TOT
	1	2	3	4	5	
1	0	239	0	0	7	246
2	182	5	0	1	0	188
3	46	0	103	14	7	170
4	2	0	34	52	25	113
5	8	0	19	11	18	56
TOT	238	244	156	78	57	773

Table 6.25: Equality table comparing experiment 2_6_3 with 150_4_3.

6.5 Conclusions and Future Research

SAM uses three different parameters within its analysis of measuring distances between sequences. Each parameter refers to an operation (deletion, insertion and substitution) and bears an operational cost (also called weight) reflecting the amount of work for changing one sequence into the other. In the previous chapters, weights of 1 for deletion, 1 for insertion and 2 for substitution are used in experimental tests. In order to obtain more information about the influence of changes in operational weights on SAM distance measures and on the final clustering results, this chapter examines the sensitivity of SAM.

First, parameterisation tests, given by different experiments using different weights for deletion, insertion and substitution, are applied on a synthetic sampled set in order to examine whether specific changes in SAM parameters cause ‘no’, ‘some’ or ‘more’ changes in (relations between) SAM distance measures. Changes in relations between SAM distance measures may provide some information about changes in final clustering results. Furthermore, instead of predicting ‘no’, ‘some’ and ‘more’ influence on relations between SAM distance measures, categories of influence on relations between SAM distance measures are specified. For example, Pearson’s correlation (c) between SAM distance measures resulting from two different experiments applied to the same sampled set measures the extent to which two variables are related. From (c), dissimilarities (dissim) are deducted, providing information how dissimilar two variables are. Interpretations of dissimilarities and correlations provide four categories of influence: no, minor, considerable and major.

In order to derive the influence of changes in weights for deletion, insertion and substitution on changes in relations between SAM distance measures, five rules are deducted from the parameterisation tests. In order to know which rule will be applied to a group of parameter settings, hierarchical examination is

necessary starting with rule 1, 2 or 3. If operational weights do not match, rule 4 is analysed. If again, no match is found, the analysis proceeds to rule 5. Finally, the analysis ends by applying rule 5 if no match was found with the previous five rules.

Second, examination is done on whether particular characteristics of sampled sets, such as the total number of distinct items (I), average sequence length (avg_seq_length), maximum sequence length (max_seq_length) and minimum sequence length (min_seq_length) might cause differences in the magnitude of changes in relations between SAM distance measures. To this end, instead of applying different experiments on one sampled set, one experiment (1_1_2) is applied to different sampled sets. The sampled sets vary in I and in sequence length. The relations between SAM distance measures resulting from different sampled sets are analysed. In addition, in order to verify that the results did not occur by chance, two more experiments (100_1_2 and 2_4_10) are used throughout the tests.

With regard to the influence of I and sequence length on changes in relations between SAM distance measures, the following conclusions are made. In most of the cases relations between SAM distance measures provide major changes. This means that correlations between SAM distance measures are less or equal than 0.5, providing dissimilarity measures higher or equal than 0.25. Given the experimental results of SAM applied to sampled sets, we conclude that I and sequence length have major influence on relations between SAM distance measures.

Finally, we remark that the results of the tests for measuring the influence of I and sequence length on changes in SAM distance measures must be carefully interpret. Dissimilarities do not exclusively measure differences between sampled sets with regard to I and sequence length. However, due to approximate equal probabilities of the items i ($i = 1, 2, \dots, I$) in the sampled sets, dissimilarities between sampled sets are mostly due to differences in I and sequence length.

In order to examine how changes in SAM parameters influence the final clustering results on a real data set, the effect of *minor*, *considerable* or *major* predicted changes in relations between SAM distance measures, resulting from different operational weights, is investigated on the final clustering results. To this end, different experiments are applied to a real web usage data set, storing visiting behaviour towards the web site of a Belgian telecom provider. The results of the tests confirm the predictions. For example, rule four predicts that changing $w_d = 5$, $w_i = 3$, $w_s = 8$ into $w_d = 7$, $w_i = 6$, $w_s = 13$ will provide a minor change in the relations between SAM distance measures. After examining the test results, the relation between SAM distance measures has indeed been changed minor, falling within the range of predicted change. With

regard to changes in the final clustering results, the following conclusions are made. Given the data set, minor, considerable and major changes in relations between SAM distance measures produced respectively 91.72%, 80.85% and 76.84% of equally clustered server sessions. This indicates that, given the results of our experimental tests, the final clustering results of our approach of Web Usage Mining by means of SAM are relatively insensitive to changes in SAM's operational weights. This means that, given the results of our experimental tests, SAM is a relatively stable method for analysing differences between server sessions in Web Usage Mining studies.

Future research should examine the influence of changes in parameters on changes in relations between SAM distance measures and on the final clustering results, using more different experiments, more synthetic sampled sets and more real data sets, in order to generalize the results to a broad area of tests and sampled/data sets. Also, the sensitivity of changes in SAM's weights for particular web pages instead of operations should be studied. Through assigning low weights for operations on semantically related web pages, it might be possible to extract clusters holding server sessions including semantically related web pages.

An example of a meaningful way for introducing costs into Web Usage Mining studies would be as follows. Suppose that we would like to obtain different clusters of visiting behaviour towards a particular web page, called page x. The clusters should provide information about typical web pages preceding and following page x. Suppose that page x is the home page of a web site, it might be interesting to distinguish different behaviour where people are coming from before visiting the home page and where people are going to after visiting the home page. The analysis starts with omitting server sessions from the database if page x is not included. Then, in the SAM algorithm, we assign weights of 1 for insertion, 1 for deletion and 2 for substitution to page x and to operations of other pages if they precede or follow page x. Also, we assign weights of 0 for insertion, deletion and substitution to operations of web pages which do not precede or follow page x, since clusters should not be based upon these pages. Further research should examine how such typical behavioural patterns can be extracted from the database and presented by clusters.

CHAPTER 7

SAM HEURISTIC

When applying SAM to web usage data in large databases (i.e. data sets of one megabyte or more, or data sets storing at least 10,000 records), efficiency problems arise in terms of computational complexity, because the number of pair-wise comparisons, or the total number of SAM distance measures, is proportional to the total number of server sessions in the analysis squared. For example, if 10,000 server sessions are analysed by means of SAM, the total number of SAM distance measures equals $(10,000 \times 9,999) / 2 = 49,995,000$, ending up with a high computational complexity. Moreover, the time needed to perform of the analysis would be approximately eight days.

To overcome the problems of computational complexity when applying SAM to large databases, we introduce in this chapter a heuristic method based on SAM, also called SAM heuristic, starting with selecting a subset of server sessions to perform the analysis. Computational complexity of the SAM heuristic is proportional to the number of server sessions in the subset squared instead of the total number of server sessions in the analysis squared. This means that SAM heuristic reduces computational complexity considerably and the duration of the analysis is shortened to minutes, maybe hours, instead of days.

The SAM heuristic randomly selects a server session, called s_k , from the database. Then, SAM distance measures are calculated between all of the remaining server sessions in the database and s_k . The algorithm proceeds with ordering the server sessions in the database, based on their SAM distance measure towards s_k , from low to high. From the ordered database, a subset of server sessions is selected in such a way that the subset holds server sessions, which are (very) similar to s_k and (very) dissimilar to s_k . After selecting the subset, SAM distance measures are calculated between every pair of server sessions in the subset. The SAM distance measures are inserted into a distance matrix and used by Ward hierarchical clustering to perform cluster analysis on the data in the subset. The number of clusters is defined following a consensus among five criteria: pseudo F statistic, T-squared statistic, R-squared, semi-partial R-squared and root mean squared standard deviation. The criteria are

also used by SAM applications in previous chapters (re. four, five and six). After assigning the server sessions in the subset to the clusters, cluster centres are defined based on the minimum sum of SAM distance measures between pairs of server sessions within each cluster. The method proceeds with assigning the remaining server sessions, which were not selected in the subset, to the clusters based on the minimum SAM distance measure with the cluster centre.

To illustrate the functionality of the SAM heuristic on real web usage data, the method is applied to files of logged data on the web site <http://machines.hyperreal.org>. A total number of 151,712 server sessions are analysed by means of the SAM heuristic. In order to examine how sensitive the results are with regard to the first randomly selected server session, ten different runs are executed on the data. Each run starts with a different starting value. The magnitude of the subset equals 0.5% or 759 server sessions.

The final clustering results are validated by comparing equalities within the division of server sessions in clusters for each pair of runs. Given the results of our experiments we may conclude that the amount of server sessions that are equally clustered across different runs is at least 80%. Also, a huge advantage in processing time is accomplished reducing time of analysis from more than 25,000 days, which is horrifying long, to approximately an hour. Therefore, given the results of our experiments, we may suggest SAM heuristic for analysing large databases of web usage data.

7.1 Large databases

Companies have had decades to accumulate masses of data about their customers, products or services. For example, an annual survey conducted by The Winter Group in 1998 showed that the ten largest Unix data warehouses range in size from 150 to 700 gigabytes (Nautilus Systems inc., 2003). Moreover, the Palomar-STScI Digital Sky Survey (DPOSS) consists of three terabytes of data, which is enough information to fill six million books (Rocke and Dai, 2003). This trend also dominates in Web (Usage) Mining, which makes the web one of the largest repositories of data today (Ganti et al, 1999). Such an explosive growth of databases makes it important for developing data mining techniques that can handle large databases. Ultimately, there will always be a need to improve the performance of Web Usage Mining algorithms, regarding their efficiency in computational terms (Cooley et al, 1997).

Defining large and very large databases is not an easy task. In Kaufman and Rousseeuw (1990), data sets consisting of hundreds or thousands of cases are considered as ‘large databases’. However, in Toivonen (1996), data sets are large if the sample size consists of several ten thousands and even millions of records. In order to avoid misunderstandings, we will define large and very large databases for Web Usage Mining studies. Based on Nautilus Systems inc. (2003), a database can be measured in terms of bytes and number of rows or records. *Large databases* contain one megabyte, or more, up to one gigabyte of data. The number of records in large databases ranges between 10,000 and one million. *Very large databases* store at least one gigabyte of data and consist of minimum one million records. For example, in Keogh et al (2001), experiments are run on large databases of 64 megabytes. In Rocke and Dai (2003) large datasets of 10,000 records are used and in Bradley et al (1998) 13,711 data items in 64 dimensions are considered as a large database. Examples of very large databases are given in Edelstein (2003), where IBM conducted a case study on a total volume of about four terabytes of data. A random sample of the data still included over 900 million records taking up 360 gigabytes of storage. Finally, in Toivonen (1996), very large databases consisting of several millions of records are analysed through sampling several ten thousands of records.

When applying SAM to web usage data in large databases, efficiency problems rise in terms of computational complexity. Therefore, we would like to introduce a SAM heuristic to analyse large databases. Before discussing the heuristic, we first describe computational complexity of SAM in the following section.

7.2 Computational complexity of SAM

From this point forward, we define *computational complexity of SAM* as the total number of SAM distance measures that are calculated between sequences (server sessions) in the analysis. High computational complexity indicates lots of SAM distance measures whereas low computational complexity implicates far less SAM distance measures. An example of high computational complexity is 5 million calculated SAM distance measures. An example of low computational complexity is 5 thousand calculated SAM distance measures. Along with computational complexity, *time complexity of SAM* is the time (expressed in seconds) that is necessary to calculate the total number of SAM distance measures.

Analysing large databases by means of SAM invokes computational complexity because the number of pair-wise comparisons, or the total number of SAM distance measures, is proportional to the total number of sequences in the analysis squared. This means that the number of SAM distance measures explodes for large databases. An illustration is given in equation (7.1). For example, consider the problem of calculating SAM between $N = 10,000$ sequences, the total number of SAM distance measures equals $(10,000 \times 9,999) / 2 = 49,995,000$, resulting in a high computational complexity.

$$\text{Computational complexity of SAM for the SAM application equals} \\ [N \times (N - 1)] / 2 \quad (7.1)$$

where

N = total number of sequences (server sessions) in the analysis;

Table 7.1 shows a simulated study of time and computational complexity for calculating SAM distance measures between sequences when $N = 100$ to 3200 . Note that, for the simulation, a Pentium III, CPU 1000 MHz, 261.56 KB RAM is used. The average length of the sequences used in the simulation studies is 10 with a maximum length of 20. The ratio between computational and time complexity rises until $N = 400$, indicating that computational complexity increments faster than time complexity. However, after this point the ratio between computational and time complexity falls down, indicating that time complexity increments faster than computational complexity. This is also graphically shown in figure 7.1. The figure shows that, if datasets become larger, time complexity exponentially raises with computational complexity.

N	Time complexity (seconds)	Computational complexity	Ratio Computational complexity / Time complexity
100	16	4,950	309.37
200	46	19,900	432.61
400	174	79,800	458.62
800	729	319,600	438.41
1,600	3,312	1,279,200	386.23
3,200	19,608	5,118,400	261.04

Table 7.1: Simulation study for SAM, comparing the number of sequences in the database with time- and computational complexity.

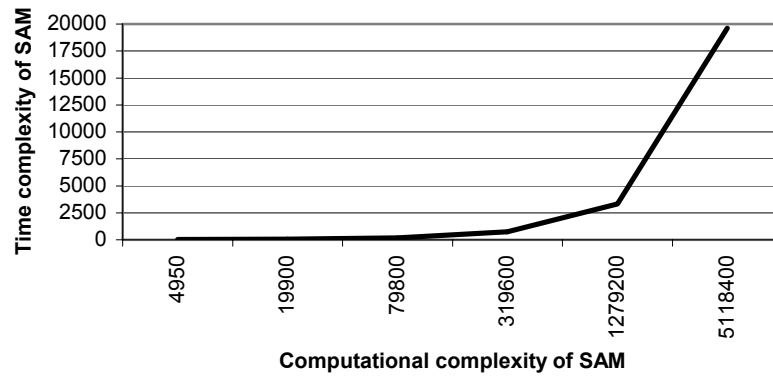


Figure 7.1: Relation between time- and computational complexity of SAM.

7.3 Computational complexity of distance-based clustering

Another type of computational complexity occurs when SAM distance measures are invoked in a distance matrix that is used for clustering. The matrix contains N rows and N columns, representing N sequences. From the matrix the distance between each pair of sequences can be read. This means that the magnitude of the matrix equals $N \times N$ distance measures, which raises the well-known problem of distance-based clustering, unsuitable for large datasets (Fasulo, 1999).

We define *computational complexity of distance-based clustering* as the total number of distance measures in the distance matrix, which need to be considered for clustering. This is illustrated by equation (7.2).

Computational complexity of distance-based clustering for the SAM application equals N^2 (7.2)

where

N = total number of sequences (server sessions) in the analysis;

7.4 SAM heuristic for large databases

To overcome the problems of computational complexity when applying SAM to large databases, we propose a method that clusters server sessions based on SAM distance measures for a small subset of the data. The resulting clusters are used to classify the remaining observations. In Banfield and Raftery (1993), the method is illustrated for model based clustering and large databases. Furthermore, in Fasulo (1999) sampling is mentioned as an important future direction for clustering research. A new general strategy for clustering is suggested as a method of data mining. In this strategy, the assumptions required by the various clustering algorithms can be explicitly parameterised, allowing the computer more freedom to search for the best way to cluster the data. To compensate for the broadening of search space of possible clustering, it is recommended that an implementation of this strategy use sampling to reduce the number of inputs if necessary.

In the subsections that follow, we first describe the algorithm of the SAM heuristic. Then we illustrate how the SAM heuristic reduces the problems of computational complexity.

7.4.1 SAM heuristic for large databases: an algorithm

The SAM heuristic for large databases starts with a random selection of one sequence, called s_k , from the database. Then, ‘one-to-all’ SAM distance measures are calculated between s_k and the remaining (N-1) sequences in the database. The algorithm follows with ordering the sequences in the database, based on their SAM distance measures towards s_k , from low to high. The first sequence of the ordered database is s_k , the second sequence is the one having the smallest SAM distance measure towards s_k and finally the last sequence is the one with the largest SAM distance measure towards s_k .

From the ordered database, a subset of sequences M, expressed as a percentage of the total database, is selected in such a way that the maximum search space is represented by the subset. Practically, this means that the algorithm reads the ordered database top-down and after selecting s_k , every $(N-1) / [((N-1) / 100) \times M]$ lines a sequence is selected. For example, suppose $N = 100,000$ and $M = 1$. After selecting the first sequence in the ordered database, every $99,999 / [(99,999 / 100) \times 1] = 100$ lines a sequence is selected from the ordered database and written in the subset file. The total number of sequences in the subset file equals $1000 (1 + 999)$. This is summarized in figure 7.2.

```
begin
  select  $s_k$                                 //random selection of one sequence from the database//
  max_distance = 0
  for i:=1 to N-1 do                          //calculate one-to-all SAM distance measures between  $s_k$  and
  begin                                       (N-1) remaining sequences in the database//
    calculate  $d_{SAM}(s_i, s_k)$ 
    if  $d_{SAM}(s_i, s_k) > \text{max\_distance}$  then  $\text{max\_distance} := d_{SAM}(s_i, s_k)$ 
  end
  ...
  write(ordered_database,  $s_k$ )              //write  $s_k$  as the first sequence in ordered_database//
  for distance:=0 to max_distance do         //order (N-1) sequences based on their SAM distance
  begin                                       measures towards  $s_k$ , from low to high//
    for i:=1 to N-1 do
      begin
        if distance =  $d_{SAM}(s_i, s_k)$  then
          begin
            write (ordered_database,  $s_i$ )
            writeln (ordered_database)
          end
        end
      end
    end
  end
  ...
  reset(ordered_database)                   //open ordered_database to read top-down//
  read(ordered_database,  $s_k$ )              //read the first sequence in ordered_database//
  write(subset,  $s_k$ )                       //write the first sequence in subset file//
  writeln(subset)
```

```

while not eof(ordered_database) do
begin                                //select sequences and write subset file//
  for j:=1 to (N-1) / [(N-1) / 100] X M] - 1 do
    begin
      readln(ordered_database)
    end
    read (ordered_database, si)
    write(subset, si)
    writeln(subset)
  end
...
end;

```

Figure 7.2: SAM heuristic: selecting a subset of sequences.

We note why we have chosen for one randomly selected sequence s_k , and why we have not chosen for several randomly selected sequences, say s_{k1} , s_{k2} , s_{k3} , s_{k4} , s_{k5} , in the first stage of the SAM heuristic. If we would have chosen more than one random selected sequence, we would risk that the sampled set may not represent the total data set due to the following reasons. Suppose, by chance, two sequences, say s_{k1} and s_{k2} , are very much alike (in the worst case scenario they may even be the same). Instead of one, we now have five ordered databases of which a subset of sequences is selected. If we would select every x lines a sequence in the first database, then proceed to the second, third, fourth and fifth database, we end up with a sampled set holding $2/5^{\text{th}}$ or 40% of its sequences that are very alike, although the total data set does represent the same division of sequences. To assure that the sampled set represents approximately the same division of sequences as in the total data set, we have chosen for one randomly selected sequence s_k and use one ordered database for sequence selection in the sampled set.

After selecting the subset, consisting of $n = [(N) / 100] \times M$ sequences, SAM pair-wise distance measures are calculated between all the sequences in the subset file. The algorithm proceeds with constructing a distance matrix and inserting SAM distance measures into the $n \times n$ matrix. Furthermore, Ward hierarchical clustering is invoked on the matrix and the number of clusters is defined following a consensus among five criteria: pseudo F statistic, T-squared statistic, R-squared, semi-partial R-squared and root mean squared standard deviation. A definition of the criteria is given in chapter four.

In a following step, for each cluster, the centre is defined. The centre of a cluster is represented by the sequence with the minimum sum of the one-to-all SAM distance measures between that sequence and all other sequences within that cluster. The SAM distance measures are read from the $n \times n$ matrix. The method is illustrated in figure 7.3.


```

begin
  for c':=1 to C do                                //C represents the total number of clusters//
    begin
      min_sum_distances = 999,999,999
      for si:=1 to nc do                          //nc represents the number of sequences grouped in
        begin                                      cluster c'//
          go to n x n matrix
          sum_distances[si] = 0
          for sj:=1 to nc - 1 do
            begin
              search dSAM(si, sj)                //search in n x n matrix for the corresponding SAM
                                                    distance measure between sequences si and sj//
              sum_distances[si] := sum_distances[si] + dSAM(si, sj)
            end
            if sum_distances[si] < min_sum_distances then min_sum_distances := sum_distances[si]
          end
          and centre[c'] := si
        end
        write(cluster_centre_file, 'the centre of cluster ', c', ' is represented by sequence ',
          centre[c'])
        writeln(cluster_centre_file)
      end
    end;

```

Figure 7.3: Sam heuristic: defining the cluster centre.

After defining the cluster centres, the remaining (N-n) sequences, which are not selected in the subset, are classified into the clusters based on the minimum distance towards the cluster centres. This means that SAM distance measures are calculated between every sequence s_j with $j = 1, 2, \dots, N-n$ and the cluster centres. Finally, s_j is classified into the cluster where the SAM distance measure between s_j and the centre is minimal. An illustration of how the remaining sequences are classified is given in figure 7.4.

```

begin
  for j:=1 to (N-n) do
    begin
      min_distance := 999,999,999
      for c':=1 to C do
        begin
          calculate dSAM(sj, centre[c'])
          if dSAM(sj, centre[c']) < min_distance then min_distance := dSAM(sj, centre[c']) and
          classification[sj] := c'
        end
      end
    end;

```

Figure 7.4: SAM heuristic: classifying remaining sequences.

7.4.2 Reducing computational complexity

7.4.2.1 Reducing computational complexity of SAM

The heuristic for SAM applied to large databases reduces the computational complexity of SAM considerably. The total number of SAM distance measures that need to be calculated by the heuristic is proportional to the number of sequences in the subset squared (n^2), instead of being proportional to the number of sequences in the analysis squared (N^2). Equation (7.3) illustrates how many calculations of SAM distance measures are necessary for a subset of n sequences.

Computational complexity of SAM for the SAM heuristic equals
 $(N-1) + [n \times (n-1)] / 2 + (N-n) \times C$ (7.3)

where

$(N-1)$ = the number of SAM distance measures to define, for each sequence in the analysis, the distance towards the first randomly selected sequence s_k in the subset;

n = total number of sequences (server sessions) in the subset;

C = total number of clusters;

N = total number of sequences (server sessions) in the analysis;

$(N-n) \times C$ = the number of SAM distance measures to define, for each sequence that is not selected in the subset, the distance towards each cluster centre;

7.4.2.2 Reducing computational complexity of distance-based clustering

If the SAM heuristic is used, computational complexity of distance-based clustering is reduced because $n \times n$ SAM distance measures need to be considered by the clustering algorithm instead of $N \times N$. This is indicated in equation (7.4).

Computational complexity of distance-based clustering for the SAM heuristic equals n^2 (7.4)

where

n = total number of sequences (server sessions) in the subset;

7.4.2.3 Other calculations

Besides calculating SAM distance measures between sequences and considering $n \times n$ SAM distance measures for clustering, other calculations must be taken into account as well when using the SAM heuristic. Equation (7.5) describes the number of calculations in addition to computational complexities.

The number of calculations in addition to computational complexities for the SAM heuristic equals

$$\sum_{i=1}^{N-1} i + (N/100) \times M + \sum_{c=1}^C (n_c + n_c^2) + (N - n) \times C \quad (7.5)$$

where

$\sum_{i=1}^{N-1} i =$	the number of calculations to order $N-1$ sequences in the analysis from low to high SAM distance measure towards the initial randomly selected sequence s_k ;
$(N/100) \times M =$	the number of calculations to select sequences for the subset;
$n_c =$	the number of sequences in cluster c ;
$n_c^2 =$	the number of calculations to define, for each sequence s_i in cluster c , the sum of SAM distance measures between s_i and all the other sequences in cluster c ;
$n_c + n_c^2 =$	the number of calculations to define the minimum sum of SAM distance measures between s_i and all the other sequences in cluster c or the number of calculations to define, for cluster c , the cluster centre;
$\sum_{c=1}^C (n_c + n_c^2) =$	the number of calculations to define, for each cluster, the cluster centre;
$(N - n) \times C =$	the number of calculations to define, for each sequence s_i not selected in the subset, the minimum SAM distance between s_i and the cluster centre;

7.4.2.4 Examples

In table 7.2 some examples of reductions in computational complexities are given. The computational complexities of the SAM heuristic are, generally, several hundreds and even several thousands of times smaller than SAM. The table also shows that the SAM heuristic is able to handle very large databases of up to several hundreds of thousands of cases. Furthermore, the complexity of the SAM heuristic mainly depends on the magnitude of the subset and on the number of clusters. Compared with the information given in table 7.1, the total number of SAM distance measures (given by computational complexity of SAM) gives an indication of the time needed to calculate the SAM distance measures. For example, the SAM heuristic applied to a database of 200,000 cases, using a subset of 1% along with 5 clusters, indicated by the criteria for the number of clusters derived from the subset, will last about 3 hours. This may seem a long time, however, compared with SAM, an improvement in time is obtained with a factor of 462,962. Applying SAM to 200,000 sequences will last about 1,388,888 hours or more than 50,000 days, which is horrifying long!

With regard to computational complexity of distance-based clustering, the distance matrix used by the SAM heuristic, is compressed into a smaller matrix, called $n \times n$ matrix, corresponding with the magnitude of the subset instead of the whole dataset. The $n \times n$ matrix used by the SAM heuristic is handled easily by distance-based clustering.

N	M	n	C	Computational complexity of SAM		Computational complexity of distance-based clustering	
				SAM	SAM heuristic	SAM	SAM heuristic
10,000	0.5	50	5	49,995,000	60,974	100,000,000	2,500
10,000	0.5	50	20	49,995,000	210,224	100,000,000	2,500
10,000	2.0	200	5	49,995,000	78,899	100,000,000	40,000
10,000	2.0	200	20	49,995,000	225,899	100,000,000	40,000
10,000	5.0	500	5	49,995,000	182,249	100,000,000	250,000
10,000	5.0	500	20	49,995,000	324,749	100,000,000	250,000
100,000	0.7	700	5	4,999,950,000	841,149	1e+10	490,000
100,000	0.7	700	20	4,999,950,000	2,330,649	1e+10	490,000
100,000	1.0	1,000	5	4,999,950,000	1,094,499	1e+10	1,000,000
100,000	1.0	1,000	20	4,999,950,000	2,579,499	1e+10	1,000,000
200,000	0.8	1,600	5	19,999,900,000	2,471,199	4e+10	2,560,000
200,000	0.8	1,600	20	19,999,900,000	5,447,199	4e+10	2,560,000
200,000	1.0	2,000	5	19,999,900,000	3,188,999	4e+10	4,000,000
200,000	1.0	2,000	20	19,999,900,000	6,158,999	4e+10	4,000,000

Table 7.2: Comparing computational complexities between SAM and SAM heuristic.

In table 7.3 some examples of additional calculations of the SAM heuristic are given. The same values for N , M , n and C are provided as in table 7.2. The values given to n_c assume that sequences are equally distributed among clusters. Note that most of the additional calculations are due to ordering the data set of $N-1$ sequences (re. column six). In the last column the total number of additional calculations of the SAM heuristic is given.

Table 7.4 provides an overview of the total effort of SAM versus SAM heuristic for the same data sets given in table 7.3 and 7.2. *Total effort for SAM* is equal to the sum of computational complexities. *Total effort for SAM heuristic* is equal to the sum of computational complexities and additional calculations. Practically, this means that the values in the sixth column of table 7.4 represent the sum of column five and seven of table 7.2. The seventh column of table 7.4 represents the sum of column six and eight of table 7.2 and the last column of table 7.3. Comparing total effort of SAM with SAM heuristic, the effort of SAM heuristic is 3 times less than SAM for analysing 10,000 sequences. Moreover, the effort of SAM heuristic is respectively 21 and 40 times less than SAM for analysing 100,000 and 200,000 sequences. This means that, compared to SAM, SAM heuristic becomes more and more appropriate when the total number of sequences in the analysis (N) augments.

Finally, in the next section, an experiment of the SAM heuristic applied to a real large database is provided. Note that this application consists of one-dimensional data, which means that the SAM heuristic is used for analysing visited web pages. However, the SAM heuristic may also be used for analysing two-dimensional data such as visited web pages along with categories of visiting page time (re. chapter four). Future research discusses computational- and time complexities for analysing large data sets of two-dimensional server sessions.

N	M	n	C	n _c	SAM heuristic				
					Additional calculations				Total additional calculations
					$\sum_{i=1}^{N-1} i$	(N/100) x M	$\sum_{c=1}^C (n_c + n_c^2)$	(N - n) x C	
10,000	0.5	50	5	50/5	49,995,000	50	550	49,750	50,045,350
10,000	0.5	50	20	50/20	49,995,000	50	175	199,000	50,194,225
10,000	2.0	200	5	200/5	49,995,000	200	8,200	49,000	50,052,400
10,000	2.0	200	20	200/20	49,995,000	200	2,200	196,000	50,193,400
10,000	5.0	500	5	500/5	49,995,000	500	50,500	47,500	50,093,500
10,000	5.0	500	20	500/20	49,995,000	500	13,000	190,000	50,198,500
100,000	0.7	700	5	700/5	704,982,704	700	98,700	496,500	705,578,604
100,000	0.7	700	20	700/20	704,982,704	700	25,200	1,986,000	706,994,604
100,000	1.0	1,000	5	1,000/5	704,982,704	1,000	201,000	495,000	705,679,704
100,000	1.0	1,000	20	1,000/20	704,982,704	1,000	51,000	1,980,000	707,014,704
200,000	0.8	1,600	5	1,600/5	1,474,936,480	1,600	513,600	992,000	1,476,443,680
200,000	0.8	1,600	20	1,600/20	1,474,936,480	1,600	129,600	3,968,000	1,479,035,680
200,000	1.0	2,000	5	2,000/5	1,474,936,480	2,000	802,000	990,000	1,476,730,480
200,000	1.0	2,000	20	2,000/20	1,474,936,480	2,000	202,000	3,960,000	1,479,100,480

Table 7.3: Additional calculations for the SAM heuristic.

N	M	n	C	n _c	Total effort	
					SAM	SAM heuristic
10,000	0.5	50	5	50/5	149,995,000	50,108,824
10,000	0.5	50	20	50/20	149,995,000	50,406,949
10,000	2.0	200	5	200/5	149,995,000	50,171,299
10,000	2.0	200	20	200/20	149,995,000	50,459,299
10,000	5.0	500	5	500/5	149,995,000	50,525,749
10,000	5.0	500	20	500/20	149,995,000	50,773,249
100,000	0.7	700	5	700/5	1.499995e+10	706,909,753
100,000	0.7	700	20	700/20	1.499995e+10	709,815,253
100,000	1.0	1,000	5	1,000/5	1.499995e+10	707,774,203
100,000	1.0	1,000	20	1,000/20	1.499995e+10	710,594,203
200,000	0.8	1,600	5	1,600/5	5.999990e+10	1,481,474,879
200,000	0.8	1,600	20	1,600/20	5.999990e+10	1,487,042,879
200,000	1.0	2,000	5	2,000/5	5.999990e+10	1,483,919,479
200,000	1.0	2,000	20	2,000/20	5.999990e+10	1,489,259,479

Table 7.4: Comparing total effort between SAM and SAM heuristic.

7.5 Application

In this section, we illustrate the functionality of the SAM heuristic on a real, large dataset. To this end, we analysed files of logged web usage data from 01/02/1999 to 31/03/1999 on the web site <http://machines.hyperreal.org>. After pre-processing the data using the approach described in chapter four, section 4.5, a total number of 151,712 server sessions, showing visits to 1159 different web pages, are identified. Note that in chapter four, web usage data from 01/02/1999 to 03/02/1999, regarding the same web site, is analysed by means of SAM. Also, in chapter six, SAM incorporating an interestingness measure for web usage data (SAM¹) is applied to logged data from 01/02/1999 to 28/02/1999.

7.5.1 *Applying SAM heuristic to 151,712 server sessions*

In order to examine how sensitive the results are with regard to the randomly selected first server session, ten different runs are executed on the data. Each run starts with a different initial randomly selected server session from the dataset. In table 7.5, the randomly selected first server sessions are given for each run. SAM distance measures are based on the following parameters. Weight values of one are given to deletion and insertion operations while a weight value of two is assigned to reordering. During each run, the data set is ordered based on one-to-all SAM distance measures between the initial selected server session and every remaining server session in the database. Following, a subset of server sessions is sampled with M equal to 0.5. The program reads the ordered database top-down from low to high SAM distance measures and every 200 lines a server session is selected in the subset. This means that, for each run, the subset consists of 759 server sessions, including the initial randomly selected first server session.

SAM heuristic	
Randomly selected first server session	
Run 1	{408, 622, 2}
Run 2	{984}
Run 3	{997, 996}
Run 4	{627, 642}
Run 5	{338, 1153, 574, 469, 86}
Run 6	{316, 714}
Run 7	{496, 509, 574}
Run 8	{403, 497, 574}
Run 9	{163, 894, 906, 947}
Run 10	{452}

Table 7.5: SAM heuristic: randomly selected first server session for ten different runs.

In the following step, for each run, all-to-all SAM distance measures are calculated between the server sessions in the subset. For each run, the SAM distance measures are inserted into a 759×759 distance matrix and Ward hierarchical clustering is invoked on the matrix. In order to define the number of clusters, a consensus among the following information criteria PSF, TST, R-squared, semi-partial R-squared and RMSSTD is used. The criteria are described in chapter four, section 4.7. Equations for calculating the criteria are given in table 4.9. Also, Ward clustering is described and table 4.8 provides equations how to calculate dissimilarities between clusters using different clustering methods.

Figures 7.5, 7.6 and 7.7 graphically present the information criteria for cluster solutions between 1 and 20 for the first three runs. Information criteria for the remaining runs are graphically presented in appendix 7. In the first run, *four clusters* are suggested by a consensus between PSF, TST, R-squared, semi-partial R-squared and RMSSTD. The variance explained by the model equals 85.20%. In the second run, *five clusters* indicate a good solution, given the criteria. TST might suggest three clusters but PSF falls down at this point. Also, the homogeneity of the data in three clusters is not high enough (i.e. the RMSSTD is not low enough), compared to other near by solutions. Moreover, 87.60% of the variance is explained by a model of five clusters. In the third run, *five clusters* are defined. Two clusters explain less than 60% of the variance in the data. Compared to other cluster solutions, the homogeneity of the data in three clusters is relatively low, indicated in figure 7.7 by a relatively high RMSSTD. Finally, table 7.6 provides the number of clusters that are chosen in each run.

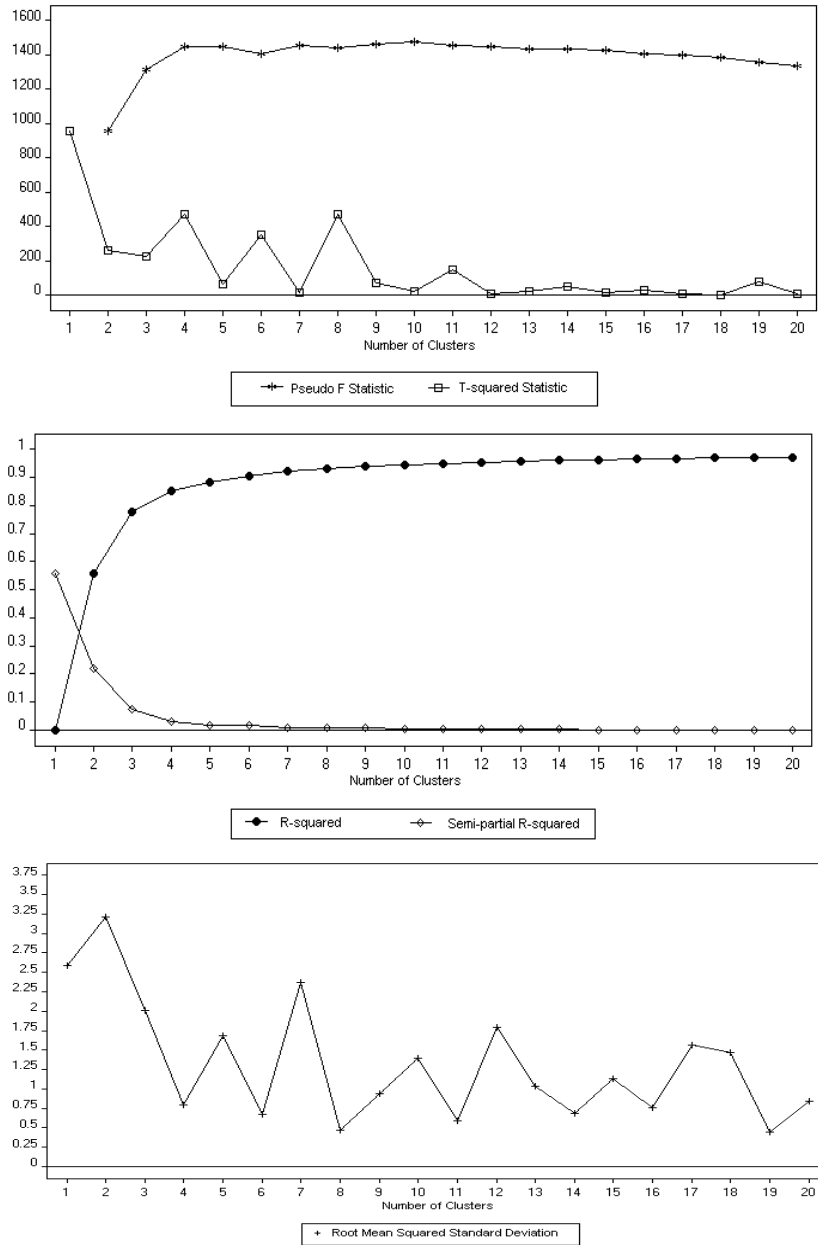


Figure 7.5: Run 1: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.

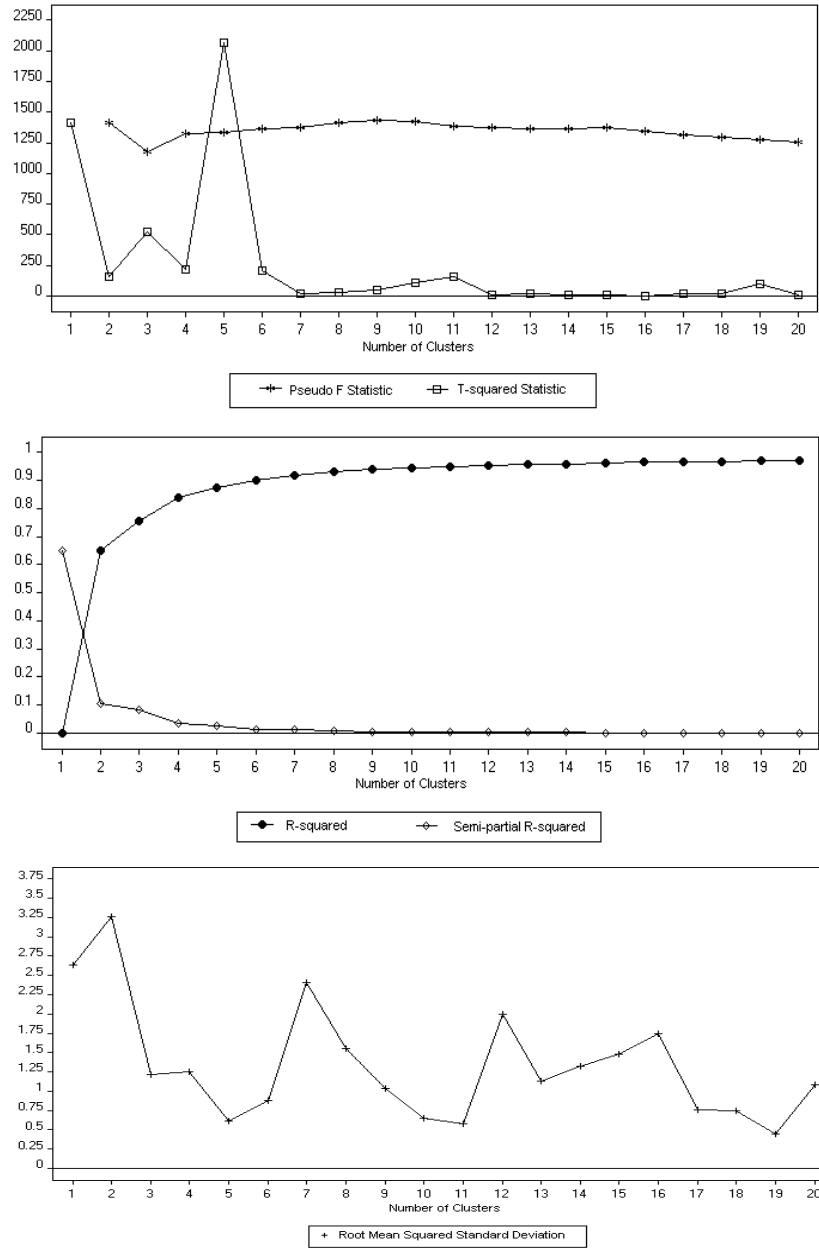


Figure 7.6: Run 2: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.

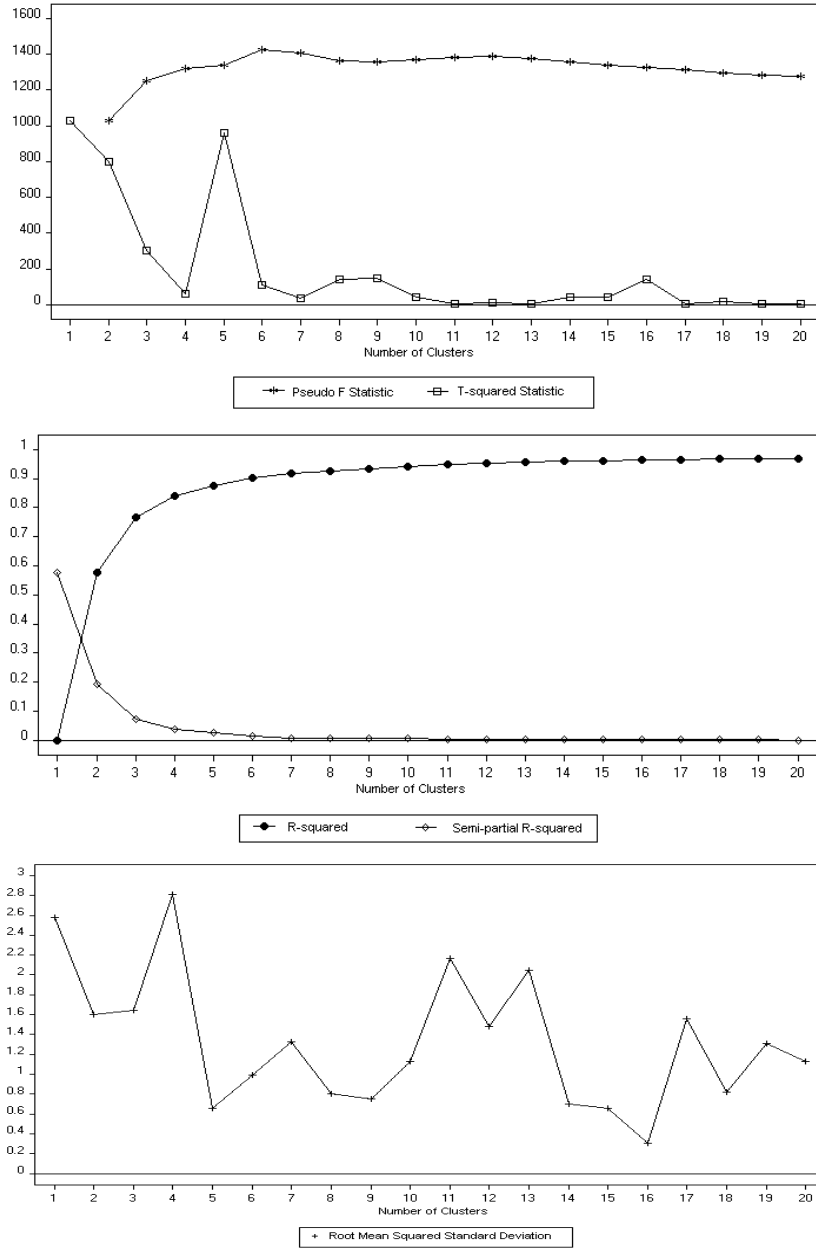


Figure 7.7: Run 3: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.

Run	Total number of clusters
1	4
2	5
3	5
4	3
5	5
6	3
7	4
8	5
9	3
10	3

Table 7.6: SAM heuristic: total number of clusters in 10 different runs.

After defining the number of clusters, the clustering procedure assigns every server session selected in the subset to each cluster. Table 7.7 shows, for each run, how the subset is divided into clusters. In each run, one cluster appears to be relatively large (i.e. holding more server sessions) compared to the others. The reason why this happens is because generally one-page sessions, which are not considered to be strong enough to represent a separate cluster in the corresponding run, are grouped together into this cluster. For example, in the first run, one-page sessions such as 906, 813 or 837 are grouped in cluster one. Yet, 90% of the server sessions in cluster three are one-page sessions to web page 657, which are considered to be strong enough to represent a cluster.

Run	Cluster					TOT
	1	2	3	4	5	
1	474	91	170	24	-	759
2	167	392	134	50	16	759
3	425	225	46	26	37	759
4	373	327	59	-	-	759
5	422	161	104	55	17	759
6	427	64	268	-	-	759
7	454	171	41	93	-	759
8	490	136	39	75	19	759
9	494	241	24	-	-	759
10	495	219	45	-	-	759

Table 7.7: SAM heuristic: subset clustering in 10 different runs.

In the proceeding step of the SAM heuristic the cluster centres are defined for each run. To this end, in each cluster, one-to-all SAM distance measures between each server session and the other server sessions are summed. The 759

x 759 distance matrix is used to search for the corresponding SAM distance measures. In each cluster, the server session with the minimum sum of the SAM distance measures is considered as the centre for that cluster. For example, in cluster one of the first run, the SAM distance measures between the first server session and every other of the remaining 473 server sessions are summed. These distance measures are already defined in a previous step of the SAM heuristic, when all-to-all SAM distance measures were calculated between the server sessions in the subset to construct the 759 x 759 SAM distance matrix. In table 7.8 the centre for each cluster in each run is given. The table shows that each run produces a cluster having one-page session 657 as cluster centre. Other server sessions, which often serve as cluster centre, are one-page sessions 163 and 1129. Comparing table 7.8 with table 7.7, we notice that six of the ten runs provide one relatively large cluster having one-page session 163 as cluster centre. Other cluster centres of relatively large clusters of subset clustering are one-page sessions 1129 (run 4 and run 10), 933 (run 5) and 947 (run 9).

Run	Cluster				
	1	2	3	4	5
1	{163}	{657, 984}	{657}	{657, 802, 657, 802, 657}	-
2	{657}	{163}	{1129}	{657, 815, 657}	{338, 1153}
3	{163}	{657}	{1129}	{984}	{657, 996, 657, 1025, 657}
4	{1129}	{657}	{657, 1026, 657}	-	-
5	{933}	{657}	{984}	{1129}	{657, 813, 657, 1134, 657}
6	{163}	{1129}	{657}	-	-
7	{163}	{657}	{1129}	{657, 815, 657}	-
8	{163}	{657}	{1129}	{657, 984}	{657, 947, 657, 984, 1000, 657, 933, 657}
9	{947}	{657}	{657, 1018, 657, 1026, 657}	-	-
10	{1129}	{657}	{657, 972, 657}	-	-

Table 7.8: SAM heuristic: cluster centres in 10 different runs.

In the last step of the SAM heuristic application, for each run, the remaining server sessions that were not selected in the subset are assigned to the clusters based on the minimum SAM distance measure between server session and cluster centre. Table 7.9 provides information about the number of server sessions in the final clusters for each run. We notice that, in each run, one cluster appears to be relatively large (i.e. holding much more server sessions) compared to the others. The reason why these large clusters appear is because they generally group together one-page server sessions to page_id 657, representing the home page of the web site, along with other one-page server sessions, which are not represented in other clusters. For example, in run one, cluster three groups together approximately 90% of the server sessions in the analysis. 23% of the server sessions in cluster three are one-page sessions to page_id 657. Examples of other server sessions grouped in cluster three are one-page sessions to page_id 1129, 984, 947 and 933, because these types of server sessions are not represented as cluster centres in any other cluster of the first run. Yet, cluster one represents one-page sessions to page_id 163. In run two, cluster one groups together approximately 86% of the server sessions in the analysis. Likewise, one fifth of the server sessions in cluster one are one-page sessions to page_id 657. Other server sessions grouped in cluster one are one-page sessions to page_id 984, 947 and 933. Unlike the previous run, one-page sessions to page_id 1129 are grouped in a separate cluster (i.e. cluster three). In run three, one-page server sessions to page_id 657 along with one-page server sessions to page_id 947 and 933 are represented by cluster two.

In general, in each run, server sessions are shifted dependant on their distance with cluster centres. For example, in run three, server sessions 657, 815, 657, 810 and 657, 802, 657, 802 both are grouped, based on their highest similarity with cluster centre {657}, in cluster two. On the other hand, the same server sessions are clustered differently in the two previous runs. In run one, server session 657, 815, 657, 810 is grouped in cluster three while server session 657, 802, 657, 802 is grouped in cluster four. In run two, server session 657, 815, 657, 810 is grouped in cluster four while server session 657, 802, 657, 802 is grouped in cluster one.

Run	Cluster					TOT
	1	2	3	4	5	
1	6,293	8,164	136,911	344	-	151,712
2	131,804	6,197	8,492	2,984	2,235	151,712
3	6,249	133,433	8,395	3,089	546	151,712
4	8,811	141,572	1,329	-	-	151,712
5	3,068	136,395	3,270	8,490	489	151,712
6	6,243	8,423	137,046	-	-	151,712
7	6,274	134,014	8,397	3,027	-	151,712
8	6,313	128,109	8,402	8,039	849	151,712
9	3,645	147,253	814	-	-	151,712
10	8,941	142,309	462	-	-	151,712

Table 7.9: SAM heuristic: dataset clustering in 10 different runs.

Figures 7.8, 7.9 and 7.10 provide, for the first three runs, graphical presentations of the clusters. On the horizontal axis, 1,159 distinct web pages are represented by means of 50 groups. Each group reflects 23 web pages, except for the last group. For example, group 1 reflects page 1 to 23, group 2 reflects page 24 to 46, group 3 reflects page 47 to 69 etc. Finally, group 50 reflects page 1,128 to 1,159. On the vertical axis, frequency values (number of requests of the page_ids within the corresponding group divided by the total number of requests in the file or cluster, multiplied by 100) are given. We remark that the same scales are used in the presentations of the graphical figures in chapter four.

In each of the three runs, figures 7.8, 7.9 and 7.10 show that every cluster represents a different distribution of visited pages, which indicates that the clusters are well separated. Also, the distribution of visited pages within clusters is relatively similar across different runs, which may argue for stable results. For example, in run one, two and three, respectively clusters one, two and one are very alike, representing a peak at group 8, holding page_ids of or between 162 and 184. In run one, two and three, respectively clusters three, one and two are very alike, representing a peak at group 29, holding page_ids of or between 645 and 667.

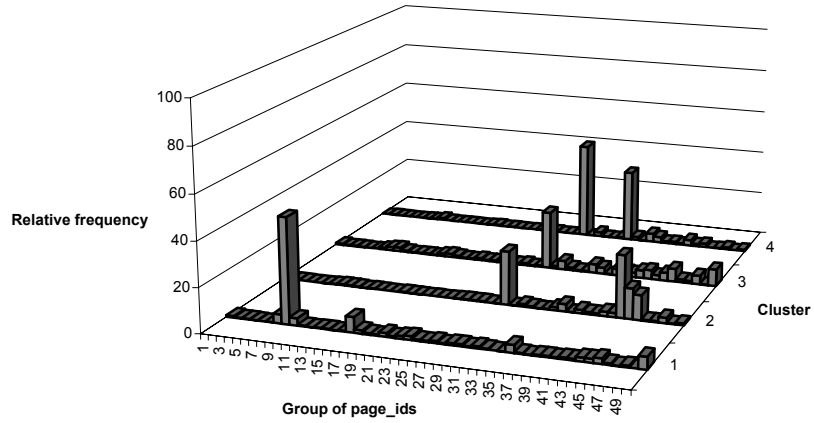


Figure 7.8: Run 1: SAM heuristic applied to 151,712 server sessions of <http://machines.hyperreal.org>: visited web pages in four clusters.

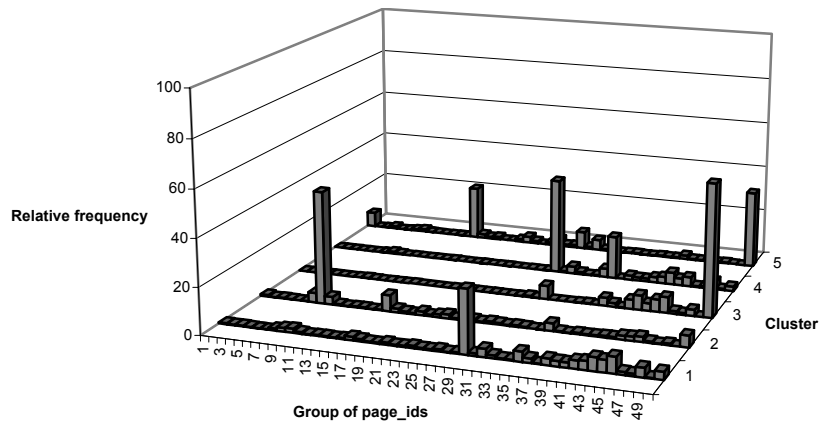


Figure 7.9: Run 2: SAM heuristic applied to 151,712 server sessions of <http://machines.hyperreal.org>: visited web pages in five clusters.

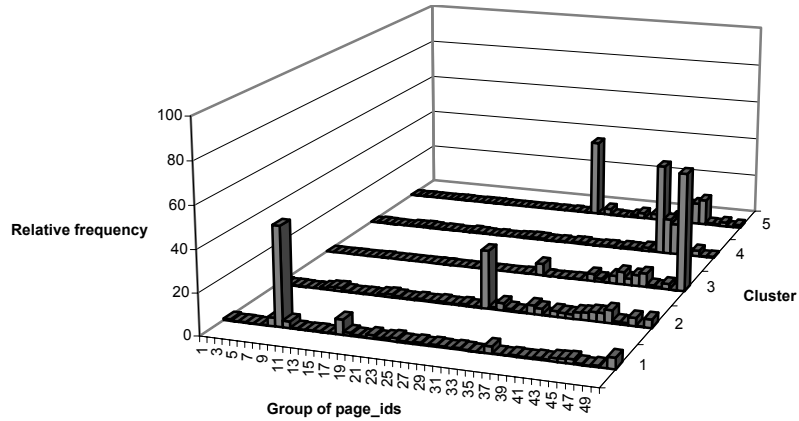


Figure 7.10: Run 3: SAM heuristic applied to 151,712 server sessions of <http://machines.hyperreal.org>: visited web pages in five clusters.

Comparing figures 7.8, 7.9 and 7.10 with cluster centres of the first three runs in table 7.8, identical clusters across different runs show equal cluster centres. For example, clusters one, two and one of respectively run one, two and three present server session {163} as cluster centre. Also, clusters three, one and two of respectively run one, two and three present server session {657} as cluster centre.

Comparing figures 7.8, 7.9 and 7.10 with the results of dataset clustering in the first three runs of table 7.9, the number of server sessions in equal clusters across different runs is more or less the same. For example, clusters one, two and one of respectively run one, two and three hold approximately 6,000 server sessions, representing the third largest cluster. Also, clusters three, one and two of respectively run one, two and three hold approximately 130,000 server sessions, representing the largest cluster.

7.5.2 Validating the results

The results of the ten different runs of the SAM heuristic, applied to a large dataset consisting of 151,712 server sessions, are validated by comparing equalities within the division of server sessions in clusters for each pair of runs. Table 7.10 presents information about the number of server sessions that are equally grouped together across different runs. For example, comparing the final clustering results of run 1 with run 2, 90.61% of the server sessions are equally grouped together.

RUN	1	2	3	4	5	6	7	8	9	10
1	100	90.61	91.71	93.31	89.90	94.05	92.06	93.34	97.06	93.80
2		100	97.15	98.77	95.37	99.54	99.41	93.67	97.06	99.28
3			100	98.80	97.39	99.58	97.57	95.72	97.06	99.30
4				100	95.36	95.71	93.69	89.81	97.06	99.29
5					100	95.84	93.82	91.91	97.06	99.35
6						100	97.59	93.72	97.06	99.29
7							100	93.69	97.06	99.30
8								100	97.06	99.30
9									100	93.80
10										100

Table 7.10: % of cases (server sessions) that are equally grouped together, for each pair of runs.

Details of table 7.10 are given in appendix 7, where clusters are presented vertically and horizontally in equality tables. An *equality table* shows equalities between server sessions grouped in each cluster, for each pair of runs. The total number of server sessions for each cluster are presented vertically (last column) and horizontally (last row). The total number of server sessions in the analysis are written in the corner at the bottom right of each table. The cells printed in bold show the maximum number of server sessions (in each row) that are equally grouped together. We remark that we use the words ‘*equally grouped together*’ to refer to the fact that, in run two, cluster one groups server sessions of cluster two, three and four of run one. This means that run two does not recognize three different groups and instead these server sessions are grouped together. With regard to table 7.10 we may also say that 85.42% (i.e. $[(123,660 + 5,937) / 151,712] * 100$) of the cases in run one and two are *equally clustered*. This means that, in run one and two, not only the same server sessions are equally grouped together but also the same clusters are distinguished.

Comparing the equality tables with the cluster centres in table 7.8 and with the graphical presentations in figures 7.8, 7.9 and 7.10, the same observations are given. Cluster one of the first run is most identical with cluster two of the second run, clearly indicated by the same cluster centre {163} in both runs. In figure 7.8 cluster one shows a peak for group 8, because page_id 163, which is an element of group 8, is frequently visited within the server sessions that are grouped in cluster one. In figure 7.9, cluster two shows a peak for group 8, indicating server sessions, which are merely concentrated on page_id 163. Furthermore, clusters two, three and four of the first run are grouped together in cluster one of the second run, due to differences in cluster centres. Also, cluster three, four and five of the second run are grouped together in cluster

three of the first run. The reason why this occurs is because run one does not recognize cluster centres {1129}, {657, 815, 657} and {338, 1153}. Server sessions are assigned to cluster three in run one because the SAM distance measure towards cluster centre {657} is smaller compared with other cluster centres. Also, cluster three of run one generally groups together one-page server sessions to page_id 657, representing the home page of the web site, along with other one-page server sessions, which are not represented by other cluster centres. Such as for example one-page server sessions 1129, which are grouped in cluster three of run one, together with one-page sessions 657. This is also shown in cluster three of figure 7.8, where group 29 including frequently visited page_id 657 shows a peak and in cluster three of figure 7.9, where group 50 including frequently visited page_id 1129 shows a peak.

The shifts of server sessions between clusters in run one and two, along with cluster centres and total number of server sessions in each cluster is given in figure 7.11. Double arrows indicate approximately equal clustering solutions between run one and two. Above each arrow, the maximum number of server sessions that is shifted from one cluster to another, between run one and two, is given. For example, 5,937 out of 6,293 server sessions from cluster one of run one are grouped in cluster two of run two. Likewise, 5,937 out of 6,197 server sessions from cluster two of run two are grouped in cluster one of run one. Also, 8,391 out of 8,492 server sessions in cluster three of run two are grouped in cluster three of run one. The same information is also given in appendix 7 (re. equality table comparing run 1 with run 2).

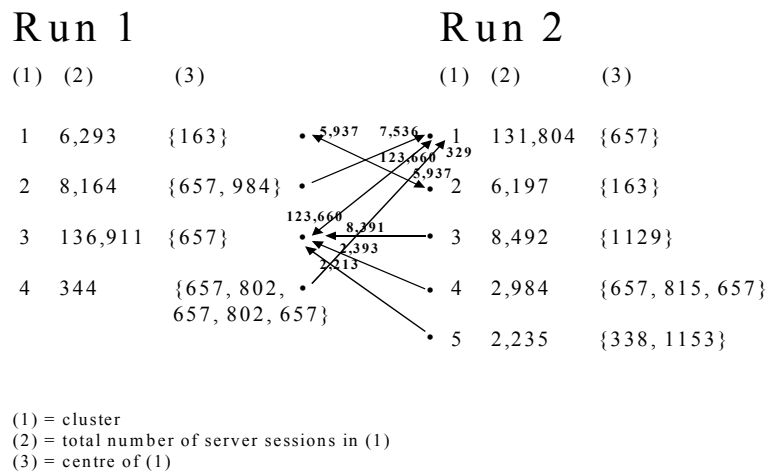


Figure 7.11: Shifts of server sessions between clusters in run one and two.

Finally, it is important to notice that 151,712 server sessions are analysed by the SAM heuristic within a time period of approximately one hour. If SAM had been applied to the dataset, the analysis would have lasted for more than 25,000 days or 68 years. This means that SAM is not able to analyse such a large data set.

7.6 Conclusion and Future Research

When applying SAM to web usage data in large databases (i.e. databases storing at least one megabyte of data or containing at least 10,000 records), efficiency problems rise in terms of computational complexity (i.e. total number of SAM distance measures). For this reason, in this chapter, a SAM heuristic is introduced, developed and applied to a large database.

Computational complexity of SAM is proportional to the number of sequences in the analysis squared, whereas computational complexity of the SAM heuristic is proportional to the number of sequences in the subset squared. This means that, taking into account other additional calculations as well, the total effort for analysing large databases of, for example, 10,000 records by means of SAM heuristic is approximately 3 times less than SAM. Moreover, the effort of SAM heuristic is respectively 21 and 40 times less than SAM for analysing 100,000 and 200,000 sequences. This means that, compared to SAM, SAM heuristic becomes more and more appropriate when the total number of sequences in the analysis (N) augments.

To illustrate the functionality of the SAM heuristic for Web Usage Mining studies on a real, large dataset, we analysed files of logged web usage data from 01/02/1999 to 31/03/1999 on the web site <http://machines.hyperreal.org>. After pre-processing the data using the approach described in chapter four, section 4.5, a total number of 151,712 server sessions, showing visits to 1159 different web pages, are identified. In order to examine how sensitive the results are with regard to the randomly selected first server session, ten different runs are executed on the data. Each run starts with a different initial randomly selected server session from the dataset. After defining, for each run, a subset of 759 server sessions (i.e. 0.5% of the original database) along with clustering information for the subset and cluster centres, the remaining 150,953 server sessions are assigned to a cluster based on the minimum SAM distance between server session and cluster centre.

The clustering results for each of the ten runs are validated by means of equality tables, showing equalities between server sessions grouped in each cluster of two different runs. The tables provide information about the total

number of server sessions that are *equally grouped together* or *equally clustered* between two different runs. ‘Equally grouped together’ refers to the number of server sessions that are grouped together between two different runs. ‘Equally clustered’ refers to the number of server sessions that are grouped together and assigned to the same type of clusters, between two different runs. For example, 90.60% of the cases in run one are equally grouped together in run two and 91.71% of the cases in run one are equally grouped together in run three. Yet, 85.42% of the cases in run one are equally clustered in run two and 88.28% of the cases in run one are equally clustered in run three. After investigating the equality tables of each pair of runs, we may conclude that the amount of server sessions that are equally clustered across different runs lies always above 80%. This means that, although the SAM heuristic starts with randomly selecting a first server session from the data set, the final clustering results are relatively stable, given the results of the experimental tests.

Differences in clustering results across different runs, which are maximally 20% of the server sessions in the data set, are generally due to differences in cluster centres, which occur due to differences in subset selection, which is in fact a consequence of the first randomly selected server session. Given the results of our experiments, we may also conclude that the length of the first randomly selected server session (one-page session, two-page session or more) does not influence the results.

Finally, it is important to notice that 151,712 server sessions are analysed by the SAM heuristic within a time period of approximately one hour. If SAM had been applied to the dataset, the analysis would have lasted for more than 25,000 days or 68 years. This means that, within our approach of Web Usage Mining, SAM is unable to analyse large datasets. With regard to SAM heuristic, given the results of our experiments, we may conclude that, although not all of the server sessions are equally clustered across different runs, most of them do. Minimum 80% of the results of the SAM heuristic are stable while maximally 20% may be sensitive to the initial randomly selected first starting value. Despite the fact that maximally 20% of the results of the SAM heuristic may be unstable, a huge advantage in processing time is accomplished and therefore we may suggest SAM heuristic for analysing large databases in Web Usage Mining studies.

Further research is necessary to examine the effect of the magnitude of the subset on the final results. This means that, given N , the optimal value for M , with minimum sensitivity to the randomly selected starting value, must be defined. Also, computational and time complexities for analysing large data sets of two-dimensional server sessions, such as visited web pages along with categories of visiting page time (re. chapter four), should be studied in order to extent SAM heuristic for two-dimensional web usage data. In order to

generalize the performance of SAM heuristic to Web Usage Data in general, more experimental tests are necessary on log files of different web sites. Finally, it might be interesting to investigate how sensitivity to the first randomly selected sequence may be reduced.

REFERENCES

- R.C. Agarwal, C.C. Aggarwal, V.V.V. Prasad. A tree projection algorithm for generation of frequent itemsets. *Journal of parallel and distributed computing*, 61(3): 350-371, 2001.
- R. Agrawal. Data Mining: Crossing the chasm. Presented at 5th ACM SIGKDD'99 International Conference on Knowledge Discovery and Data Mining (KDD-99), 1999.
- R. Agrawal, T. Imielinski, A. Swami. Database Mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering, Special Issue on Learning and Discovery in Knowledge-Based Databases*, 5(6): 914-925, 1993.
- R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 307-328. MIT Press, 1996.
- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke, C. Zaniolo, editors, *Proceedings 20th International Conference on Very Large Databases*, pages 487-499. Morgan Kaufmann, 1994a.
- R. Agrawal and R. Srikant. Mining sequential patterns. In P.S. Yu, A.L.P. Chen, editors, *Proceedings 11th International Conference on Data Engineering*, pages 3-14. IEEE Computer Society, 1995. Expanded version IBM Research Report RJ 9910, 1994b.
- G.O. Arocena, A.O. Mendelzon. WebSQL: Restructuring documents, databases and webs. In *Proceedings 14th International Conference on Data Engineering*, pages 24-33. IEEE Computer Society, 1998.
- J.F. Baldwin. Evidential support logic programming. *Fuzzy sets and systems*, 24(1): 1-26, 1987.
- A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. In *Proceedings Web Mining workshop at the 1st SIAM Conference on Data Mining*, pages 33-40, 2001.
- J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49: 803-821, 1993.
- B. Berendt, B. Mobasher, M. Spiliopoulou, J. Wiltshire. Measuring the accuracy of sessionizers for web usage analysis. In *Proceedings Web Mining workshop at the 1st SIAM Conference on Data Mining*, pages 7-14, 2001.

T. Berners-Lee. The World Wide Web: past, present and future. <http://www.w3.org/People/Berners-Lee/1996/ppf.html>, 1996.

Biz/ed, http://www.bized.ac.uk/timeweb/crunching/crunch_relate_expl.htm, 2003.

H.H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2: 77-108, 1985.

C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In W. Härdle and B. Rönz, editors, *Proceedings 15th Conference on Computational Statistics*, pages 395-400. Physica-Verlag, 2002.

J. Borges and M. Levene. A fine grained heuristic to capture web navigation patterns. *SIGKDD Explorations*, 2(1): 40-50, 2000a.

J. Borges and M. Levene. Data Mining of user navigation patterns. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Usage Analysis and User Profiling*, volume 1836 of *Lecture Notes in Computer Science*, pages 92-111. Springer-Verlag, 2000b.

P.J. Boyle and R. Flowerdew. Improving distance estimates between areal units in migration models. *Geographical Analysis*, 29: 93-107, 1997.

P.S. Bradley, U. Fayyad, C. Reina. Scaling clustering algorithms to large databases. In *Proceedings International KDD Conference on Knowledge Discovery and Data Mining*, pages 9-15. AAAI Press, 1998.

S. Brin, R. Motwani, C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Peckham, editor, *Proceedings International ACM SIGMOD Conference on Management of Data*, pages 265-276. Sigmod Record, 1997.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In H. Ashman and P. Thistlewaite, editors, *Proceedings 7th International World Wide Web Conference*. Elsevier, 1998. *Computer Networks*, 30(1-7): 107-117, 1998.

A.G. Büchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J.G. Highes. Navigation pattern discovery from internet data. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Usage Analysis and User Profiling*, pages 25-30. Springer-Verlag, 1999.

A.G. Büchner and M.D. Mulvenna. Discovering Internet marketing intelligence through online analytical web usage mining. *ACM SIGMOD Record*, 27(4): 54-61, 1998.

- I. Cadez, D. Heckerman, C. Meek, P. Smyth, S. White. Visualization of navigation patterns on a web site using model-based clustering. In E. Simoudis, J. Han, U. Fayyad, editors, *Proceedings 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 280-284. AAAI Press, 2000. Technical Report MSR-TR-00-18, Microsoft Research, 2000.
- T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3: 1-27, 1974.
- Capri. Generic sequence discovery product. <http://www.mineit.com/products>, 2001.
- L. Catledge and J. Pitkow. Characterizing browsing behaviors on the world wide web. *Computer Networks and ISDN Systems*, 27(6): 1065-1073, 1995.
- CERN, Conseil Européenne pour la Recherche Nucléaire <http://public.web.cern.ch>, 2002.
- S. Chakrabarti, B. Dom, D. Gibson, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. Experiments in topic distillation. In *ACM SIGIR workshop on Hypertext Information Retrieval on the Web*, 1998.
- F.K. Chan. A non-invasive learning approach to building web user profiles. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Usage Analysis and User Profiling*, pages 7-12. Springer-Verlag, 1999.
- G. Chang, M.J. Healey, J.A.M. McHugh, J.T.L. Wang. *Mining the world wide web. An information search approach*. Kluwer Academic Publishers, 2001.
- E.H. Chi, P. Piroli, J. Pitkow. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of ACM CHI 2000 Conference on Human Factors in Computing Systems*, pages 161-168. ACM Press, 2000.
- R. Cooley. Web Usage Mining: discovery and application of interesting patterns from web data. PhD thesis, faculty of the graduate school of the University of Minnesota. <http://www-users.cs.umn.edu/~cooley/pubs.html>, 2000.
- R. Cooley, B. Mobasher, J. Srivastava. Web Mining: information and pattern discovery on the world wide web. A survey paper. In *Proceedings 9th International Conference on Tools with Artificial Intelligence*, pages 558-567. IEEE, 1997.
- R. Cooley, B. Mobasher, J. Srivastava. Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems*, 1(1): 5-32, 1999a.

- R. Cooley, P.-N. Tan, J. Srivastava. Discovery of interesting usage patterns from web data. Technical Report TR 99-022 University of Minnesota, 1999b.
- R. Cooley, P.-N. Tan, J. Srivastava. WebSIFT: The web site information filter system. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Usage Analysis and User Profiling*. Springer-Verlag, 1999c.
- M.C. Cooper and G.W. Milligan. The effect of error on determining the number of clusters. In *Proceedings International Workshop on Data Analysis, Decision Support and Expert Knowledge Representation in Marketing and Related Areas of Research*, pages 319-328, 1988.
- M. W. Craven and J.W. Shavlik. Using neural networks for data mining. *Future Generation Computer Systems, special issue on data mining*, 13: 211-229, 1998.
- H. Dai and B. Mobasher. A road map to more effective web personalization: Integrating domain knowledge with Web Usage Mining. In *Proceedings 4th International Conference on Internet Computing*, pages. CSREA Press, 2003.
- L. Dehaspe and H. Toivonen. Discovery of relational association rules. In S. Dzeroski and N. Lavrac, editors, *Relational data mining*, pages 189-212. Springer, 2001.
- R.B. Doorenbos, O. Etzioni, D.S. Weld. A scalable comparison-shopping agent for the World Wide Web. In *Proceedings 1st International Conference on Autonomous Agents*, pages 39-48. ACM Press, 1997.
- R.O. Duda and P.E. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, 1973.
- H. Edelstein. Mining large databases – a case study. <http://www.twocrows.com/largedb.pdf>, 2003.
- M. Ester, H.P. Kriegel, M. Schubert. Web Site Mining: A new way to spot competitors, customers and suppliers in the World Wide Web. In *Proceedings 8th International Conference on Knowledge Discovery and Data Mining*. ACM SIGKDD, 2002.
- M.P. Etgen and J. Cantor. What does getting wet (web event-logging tool) mean for web usability? In *Proceedings 5th Conference on Human Factors and the Web*, 1999.
- B.S. Everitt. *Cluster analysis*. Halsted Press, 1980.
- B.S. Everitt. Unresolved problems in cluster analysis. *Biometrics*, 35: 169-181, 1979.
- D. Fasulo. An analysis of recent work on clustering algorithms. Technical Report 01-03-02 University of Washington, 1999.

U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. *Advances in knowledge discovery and data mining*. MIT Press, 1996.

A. Foss, W. Weinan, O.R. Zaïane. A non-parametric approach to web log analysis. In *Proceedings Web Mining workshop at the 1st SIAM Conference on Data Mining*, pages 41-50, 2001.

Y. Fu, K. Sandhu, M-Y Shih. Clustering of web users based on access patterns. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Usage Analysis and User Profiling*. Springer-Verlag, 1999.

V. Ganti, J. Gehrke, R. Ramakrishnan. Mining very large databases. *IEEE Computer*, 32(8): 38-45, 1999.

D. Gusfield. Core string edits, alignments, and dynamic programming. In *Algorithms on Strings, Trees, and Sequences: Computer Sciences and Computational Biology*, pages 215-246. Cambridge University Press, 1997.

B. Goethals. *Efficient Frequent Pattern Mining*. PhD thesis, transnationale Universiteit Limburg, School voor Informatietechnologie, Kennistechnologie, Informatica, Wiskunde, ICT. Universiteit Maastricht, Limburgs Universitair Centrum (LUC), 2002.

J. Hair, R. Andersen, R. Tatham, W. Black. *Multivariate Data Analysis*. Prentice Hall, 1998.

J. Han, O.R. Zaïane, Y. Fu. Resource and knowledge discovery in global information systems: A scalable multiple layered database approach. In *Proceedings 1st International Conference on Advances in Digital Libraries*. ACM Press, 1995.

D. Hand, H. Mannila, P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

J.A. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2: 63-76, 1985.

B. Hay, G. Wets, K. Vanhoof. A position and page-sensitive sequence alignment method illustrated for visiting patterns within web sites. *Expert Update – Knowledge Based Systems and Applied Artificial Intelligence*, 3(3): 15-19, 2001a.

B. Hay, G. Wets, K. Vanhoof. Clustering navigation patterns on a web site using a sequence alignment method. In *Proceedings Intelligent Techniques for Web Personalization Workshop (ITWP'01) at the 17th International Conference on Artificial Intelligence*, pages 1-6, 2001b.

B. Hay, G. Wets, K. Vanhoof. Discovering interesting navigations on a web site using Sequence Alignment Method extended with an Interestingness Measure. Accepted at

Intelligent Techniques for Web Personalization Workshop (ITWP'03) at the 18th International Conference on Artificial Intelligence, 2003a.

B. Hay, G. Wets, K. Vanhoof. Mining navigation patterns using a sequence alignment method. *Knowledge and Information Systems*, to appear, 2003b.

B. Hay, G. Wets, K. Vanhoof. Multi-dimensional sequence alignment methods: Discovering navigation patterns on web sites presenting page- and time information. In *Proceedings of International Conference on Machine Learning and Applications (ICMLA'02)*, pages 57-63, 2002a.

B. Hay, G. Wets, K. Vanhoof. Segmentation of visiting patterns on web sites using a sequence alignment method. *Journal of Retailing and Consumer Services*, 10: 145-153, 2003c.

B. Hay, G. Wets, K. Vanhoof. Web Usage Mining by means of multi-dimensional sequence alignment methods. In *Proceedings of WEBKDD, Web Mining for Usage Patterns and User Profiles at the ACM-SIGKDD Conference on Knowledge Discovery in Databases*, pages 44 -52. ACM, 2002. Extended version in *Lecture Notes in Artificial Intelligence*, to appear, 2003d.

B. Hay, G. Wets, K. Vanhoof. Web Usage Mining by means of multi-dimensional sequence alignment methods: preliminary study. In *Proceedings of Belgian-Dutch Conference on Artificial Intelligence (BNAIC'02)*, pages 123-130, 2002b.

J. Heer and E.H. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *Proceedings Web Mining Workshop at the 1st SIAM Conference on Data Mining*, pages 51-58, 2001.

J. Heer and E.H. Chi. Separating the swarm: Categorization methods for user sessions on the web. In *Proceedings of ACM CHI 2002 Conference on Human Factors in Computing Systems*, pages 243-250. ACM Press, 2002.

F.S. Hillier and G.J. Lieberman. *Introduction to Operations Research*. McGraw-Hill International Editions, 1990.

Hitmill. http://www.hitmill.com/internet/web_history.asp, 2003

J.I. Hong and J.A. Landay. WebQuilt: A framework for capturing and visualizing the web experience. In *Proceedings 10th International World Wide Web Conference*, pages 717-724. ACM Press, 2001.

ISPA, Belgium. Organization of Belgian Internet providers. <http://www.ispa.be>, 2003.

- T. Joachims, D. Freitag, T.M. Mitchell. Web Watcher: A tour guide for the World Wide Web. In *Proceedings 15th International Joint Conference on Artificial Intelligence*, pages 770-777. Morgan Kaufmann, 1997.
- C.H. Joh, T.A. Arentze, F. Hofman, H.J.P. Timmermans. Activity-travel pattern similarity: A multidimensional alignment method. *Transportation Research B*, 36: 385-403, 2002.
- C.H. Joh, T.A. Arentze, H.J.P. Timmermans. A position-sensitive sequence-alignment method illustrated for space-time activity-diary data. *Environment and planning A*, 33(2): 313-338, 2001.
- C.H. Joh, T.A. Arentze, H.J.P. Timmermans. Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms. *Geographical Analysis*, 33(3): 247-270, 2001.
- C.H. Joh, T.A. Arentze, H.J.P. Timmermans, P. Popkowski-Leszczyc. Identifying purchase-history sensitive shopper segments using scanner panel data and sequence alignment methods. *Journal of Retailing and Consumer Services*, 10: 135-144, 2003.
- H. Kato, T. Nakayama, Y. Yamane. Navigation analysis tool based on the correlation between contents distribution and access patterns. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Mining for E-Commerce – Challenges and Opportunities*, Springer-Verlag, 2000.
- L. Kaufman and P.J. Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, 1990.
- E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3): 263-286, 2001.
- R. Kohavi and F. Provost. Applications of data mining to electronic commerce. *Data Mining and Knowledge Discovery*, 5(1/2): 5-10, 2001.
- R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations*, 2(1): 1-15, 2000.
- C. Kwok, O. Etzioni and D.S. Weld. Scaling question answering to the web. In *Proceedings 10th International World Wide Web Conference*, pages 150-161. IW3C2, 2001.
- M. Les and C. Maher. Measuring diversity: Choice in local housing markets. *Geographical Analysis*, 30(2): 172-190, 1998.

- B. Liu, W. Hsu, S. Chen. Using general impressions to analyze discovered classification rules. In *Proceedings 3rd International Conference on Knowledge Discovery and Data Mining*, pages 31-36. AAAI Press, 1997.
- P. Macnaughton-Smith, W.T. Williams, M.B. Dale, L.G. Mockett. Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature*, 202: 1034-1035, 1964.
- H. Mannila and P. Ronkainen. Similarity of event sequences. In *Proceedings 4th International Workshop on Temporal Representation and Reasoning*, pages 136-139. TIME, IEEE Computer Society, 1997.
- H. Mannila, H. Toivonen, A.I. Verkamo. Discovering frequent episodes in sequences. In *Proceedings 1st International Conference on Knowledge Discovery and Data Mining*, pages 210-215. AAAI Press, 1995.
- B. Masand and M. Spiliopoulou. *Advances in Web Usage Mining and User Profiling: Proceedings of the webkdd'99 workshop*. LNAI 1836. Springer Verlag, 2000.
- J. Mena. *Data mining your website*. Digital Press, 1999.
- J. Mena. *Webmining for profit. E-business optimization*. Digital Press, 2001.
- G.W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45: 325-342, 1980.
- G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50: 159-179, 1985.
- B. Mobasher, R. Cooley, J. Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8): 142-151, 2000.
- B. Mobasher, H. Dai, T. Luo, M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proceedings 3rd International Workshop on Web Information and Data Management*, pages 9-15. ACM, 2001.
- B. Mobasher, H. Dai, T. Luo, M. Nakagawa, J. Wiltshire. Discovery of aggregate usage profiles for web personalization. In B. Masand and M. Spiliopoulou, editors, *Proceedings WEBKDD Workshop on Web Mining for E-Commerce – Challenges and Opportunities*, Springer-Verlag, 2000.
- B. Mobasher, N. Jain, E. Han, J. Srivastava. Web mining: Pattern discovery from world wide web transactions. Technical Report TR 96-050, University of Minnesota, 1996.
- M.D. Mulvenna, S.S. Anand, A.G. Büchner. Personalization on the net using web mining: introduction. *Communications of the ACM*, 43(8): 122-125, 2000.

A.T. Murray. Spatial characteristics and comparisons of interaction and median clustering models. *Geographical Analysis*, 32(1): 1-18, 2000.

O. Nasraoui, H. Frigui, A. Joshi, R. Krishnapuram. Mining web access logs using relational competitive fuzzy clustering. In *Proceedings 8th International Fuzzy Systems Association World Congress*, 1999.

Nautilus Systems inc. Competitive edge from information. <http://www.nautilus-systems.com/vlbd.html>, 2003.

Netcraft. <http://news.netcraft.com>, 2003.

D.S.W. Ngu and X. Wu. Sitehelper: A localized agent that helps incremental exploration of the World Wide Web. *International journal of computer and telecommunications networking*, 30, 1997.

Northwest Missouri State University,
<http://www.nwmissouri.edu/nwcourses/martin/methods/peascorr.html>, 2003.

Packet Sniffing Tools.
http://cs.ecs.baylor.edu/~donahoo/tools/sniffer/packet_sniffing_tools.htm, 2003.

B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings 4th International Conference on Knowledge Discovery and Data Mining*, pages 94-100. AAAI Press, 1998.

N. Pendse. What is OLAP?, the OLAP report. <http://www.olapreport.com>, 2003.

M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118 (1-2), 2000.

M. Perkowitz and O. Etzioni. Adaptive web sites: An AI challenge. In *Proceedings 15th International Joint Conference on Artificial Intelligence*, pages 16-23. Morgan Kaufmann, 1997.

G. Piatetsky-Shapiro, U.M. Fayyad, P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1-35. AAAI/MIT Press, 1996.

G. Piatetsky-Shapiro and C.J. Matheus. The interestingness of deviations. In *Proceedings Workshop on Knowledge Discovery in Databases*, pages 25-36. AAAI, 1994.

J. Pitkow. Summary of www characterizations. *Computer Networks and ISDN Systems*, 30(1-7): 551-558, 1998.

J. Pitkow and K. Bharat. Webviz: A tool for world-wide-web access log analysis. Technical Report GIT-GVU-94-20, Georgia Institute of Technology, 1994.

J. Pitkow. In search of reliable usage data on the www. In *Proceedings 6th International World Wide Web Conference*, pages 451-463. Elsevier, 1997.

D.M. Rocke and J. Dai. Sampling and subsampling for cluster analysis in data mining with applications to sky survey data. <http://myprofile.cos.com/rocked16>, 2003.

D. Sankoff and J.B. Kruskal. *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*. Addison Wesley, 1983.

C. Shahabi and F. Banaei-Kashani. A framework for efficient and anonymous web usage mining based on client-side tracking. In R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava, editors, *Proceedings WEBKDD Workshop on Mining Web Log Data Across All Customer Touch Points*, volume 2356 of *Lecture Notes in Computer Science*, pages 113-145. Springer-Verlag, 2001.

C. Shahabi, A. Faisal, F. Banaei-Kashani, J. Faruque. INSITE: A tool for real-time knowledge discovery from users web navigation. In *Proceedings 26th International Conference on Very Large Databases*, pages 635-638. Morgan Kaufman, 2000.

C. Shahabi, A. Zarkesh, J. Adibi, V. Shah. Knowledge discovery from users web-page navigation. In *Proceedings 7th International Workshop on Research Issues in Data Engineering*, pages 20-31. IEEE, 1997.

J. Shakes, M. Langheinrich, O. Etzioni. O. Ahoy! The home page finder. In *Proceedings 6th International World Wide Web Conference*. Elsevier, 1997.

A. Sieg, B. Mobasher, S. Lytinen, R. Burke. Concept based query enhancement in the ARCH search agent. In *Proceeding of the 4th International Conference on Internet Computing*. CSREA Press, 2003.

A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6): 970-974, 1996.

E.A. Silver and R. Peterson. *Decision systems for inventory management and production planning*. John Wiley & Sons, 1985.

M. Spiliopoulou, L.C. Faulstich. WUM: A web utilization miner. In *Proceedings Workshop EDBT WebDB98*, volume 1590 of *Lecture Notes in Computer Science*, pages 109-115. Springer Verlag, 1998.

M. Spiliopoulou, C. Pohle, L.C. Faulstich. Improving the effectiveness of a web site with web usage mining. In B. Masand and M. Spiliopoulou, editors, pages 139-159. Springer Verlag, 2000.

SPSS Tutorial, University of Scranton,
<http://academic.uofs.edu/departement/psych/methods/common99/level2a.html>, 2003.

R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In P.M.G. Apers, M. Bouzeghoub, G. Gardarin, editors, *Proceedings 5th International Conference on Extending Database Technology*, pages 3-17. Springer-Verlag, 1996. Expanded version, IBM Research Report RJ 9994, 1995.

R. Srikant and Y. Yang. Mining web logs to improve web site organization. In *Proceedings 10th International World Wide Web Conference*, pages 430-437. Elsevier, 2001.

J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2): 12-23, 2000.

D.J. States and M.S. Boguski. Similarity and homology. In M. Gribskov and J. Devereux, editors, *Sequence Analysis Primer*, pages 92-124. Stockton Press, 1991.

Texas A&M University, News and Information from the Department of Statistics,
<http://stat.tamu.edu/stat30x/notes/node41.html>, 2003.

K. Thearling. An introduction to data mining. Discovering hidden value in your data warehouse. White paper, <http://www.thearling.com/index.htm#wps>, 2003.

H. Toivonen. Sampling large databases for association rules. In J.B. Bocca, M. Jarke, C. Zaniolo, editors, *Proceedings 22nd International Conference on Very Large Databases*, pages 134-145. Morgan Kaufmann, 1996.

W. Van Baelen. Web Usage Mining: Sensitiviteitsanalyse van de Sequence Alignment Method. Eindverhandeling voorgedragen tot het bekomen van de graad Handelsingenieur in de Beleidsinformatica, LUC, Diepenbeek, 2003.

W. Wang and O.R. Zaïane. Clustering web sessions by sequence alignment. In *Proceedings 3rd International Workshop on Management of Information on the Web*, pages 394-398, 2002.

M.S. Waterman. *Introduction to Computational Biology*. Chapman and Hall, 1995.

Webopedia. Online dictionary for computer and internet terms. <http://webopedia.internet.com>, 2002.

R. Weiss, B. Velez, M.A. Sheldon, C. Namprempre, P. Szilagyi, A. Duda, D.K. Gifford. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings 7th International ACM Conference on Hypertext*, pages 180-193, 1996.

W.C. Wilson. Activity pattern analysis by means of sequence alignment methods. *Environment and planning*, A(30): 1017-1038, 1998.

I. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, 2000.

K. Wu, P.S. Yu, A. Ballman. Speedtracer: A Web Usage Mining and analysis tool. *IBM systems journal*, 37(1): 89-105, 1998.

W3C. World Wide Web Consortium. <http://www.w3c.org>, 2003.

T. Yan, M. Jacobsen, H. Garcia-Molina, U. Dayal. From user access patterns to dynamic hypertext linking. In H. Ashman and P. Thistlewaite, editors, *Proceedings 5th International World Wide Web Conference*, pages 1007-1014. Elsevier, 1996.

O.R. Zaïane. From resource discovery to knowledge discovery on the Internet. <http://www.cs.ualberta.ca/~zaïane>, 1998.

O.R. Zaïane. Conference tutorial notes. Web mining: Concepts, practices and research. Presented at 14th Brazilian Symposium on Databases, pages 410-474. ACM SIGMOD, 2000.

O.R. Zaïane and J. Han. Resource and knowledge discovery in global information systems: A preliminary design and experiment. In U.M. Fayyad and R. Uthurusamy, editors, *Proceedings 1st International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1995.

O.R. Zaïane and J. Luo. Towards evaluating learners' behaviour in a web-based distance learning environment. In T. Okamoto, R. Hartley, J. Kinshuk & Klus, editors, *Proceedings International Conference on Advanced Learning Technologies*, pages 357-360. IEEE Press, 2001.

O.R. Zaïane, M. Xin, J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Proceedings Advances in Digital Libraries*, pages 19-29. IEEE Press, 1998.

O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *Proceedings 21st International SIGIR Conference on Research and Development in Information Retrieval*, pages 46-54. ACM, 1998.

APPENDIX 1

The basic data fields in log files (re. table 2.1) are explained below.

IP address Internet Protocol address, also known as remote hostname, is the Internet address of the machine that made the request and may be represented by the address of a proxy server. For most users accessing the Internet from a dial-up Internet service provider (ISP), the IP address will be different every time they log on. The format of an IP address is a 32-bit numeric address written as four numbers separated by periods. Each number can be zero to 255. For example, 1.168.12.243.

User ID User Identification or remote login name of the user. If authentication is required to access password protected WWW pages, the user id is filled in.

Date Date and time of the request. Time refers to the moment the request was received by the web server.

Request This field records the method, URI and protocol for the object that is retrieved by the client.

Method This may be GET (requests an object from the web server), POST (sends information to the web server) or HEAD (requests just the HTTP header for an object).

URI The Uniform Resource Identifier can either be a static file in the local file system, or the name of an executable program that is called in response to a request.

Protocol Examples of protocol systems are File Transfer Protocol (FTP) and Hyper Text Transfer Protocol (HTTP). FTP is the protocol used on the Internet for sending files. HTTP is the underlying protocol used by the World Wide Web. HTTP defines how messages are formatted and transmitted and what actions web servers and browsers should take in response to various commands. For example, when you enter a URL in your browser, this actually sends an HTTP command to the web server directing it to fetch and transmit the requested web page.

Status This is the HTTP response code returned to the client. It indicates whether or not the file was successfully retrieved. If not, an error message is returned. Different codes are used, for example, codes ranging from 200 to 299 indicate success, 300 to 399 indicate some form of redirection, 400 to 499 indicate an error serving the particular request and finally codes ranging from 500 to 599 indicate a problem with the web server. A list of frequently used HTTP response codes is presented in the table below.

Code	Meaning
200	OK.
201	Created.
202	Accepted.
204	No content.
301	Moved permanently.
302	Moved temporarily.
304	Not modified.
400	Bad request.
401	Unauthorized.
403	Forbidden.
404	Not found.
500	Internal server error.
501	Not implemented.
502	Bad gateway.
503	Service unavailable.

Frequently used HTTP response codes (status).

Bytes The number of bytes transferred.

Referrer The url that was visited before making this particular request. The referrer field will be null for url's that are typed in or for an access through a bookmark (Cooley, 2000).

User agent The operating system and browser software the client is using.

APPENDIX 4

Commonly used distance measures, which do not take into account the order of elements within sequences (SPSS Tutorial, 2003).

- Euclidean (straight-line) distance

$$d(x, y) = \sqrt{(x - y)(x - y)}$$

- Minkowski metric

$$d(x, y) = \left[\sum_{i=1}^p |x_i - y_i|^m \right]^{1/m} \quad \text{and} \quad \begin{array}{l} p = \text{number of dimensions;} \\ m = 1 \text{ for city-block distance;} \\ m = 2 \text{ for Euclidean distance;} \end{array}$$

- Jaccard coefficient

$$d(x, y) = X \cap Y / X \cup Y$$

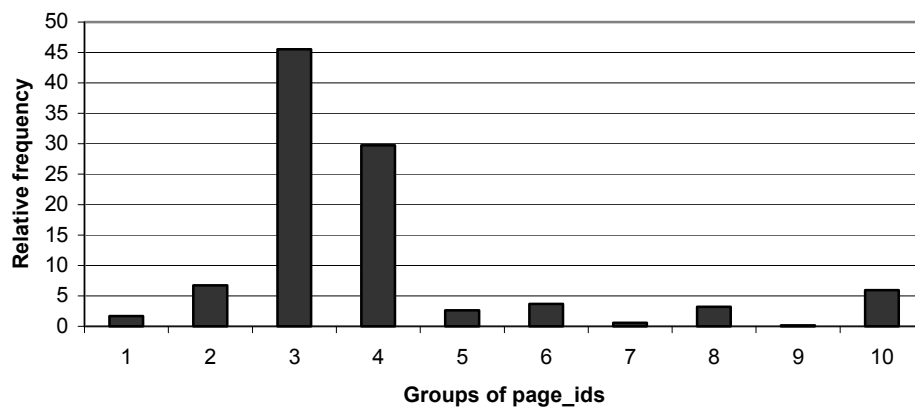
- Simple matching coefficient

$$d(x, y) = X \cap Y$$

Groups of page_ids based on classes for data set 2 (<http://machines.hyperreal.org>).

Class	Group
Categories	1
Machines	2
Manufacturers	3
Music	4
Do-it-yourself	5
Drum-machines	6
Samples	7
Software	8
Incoming	9
Remaining	10

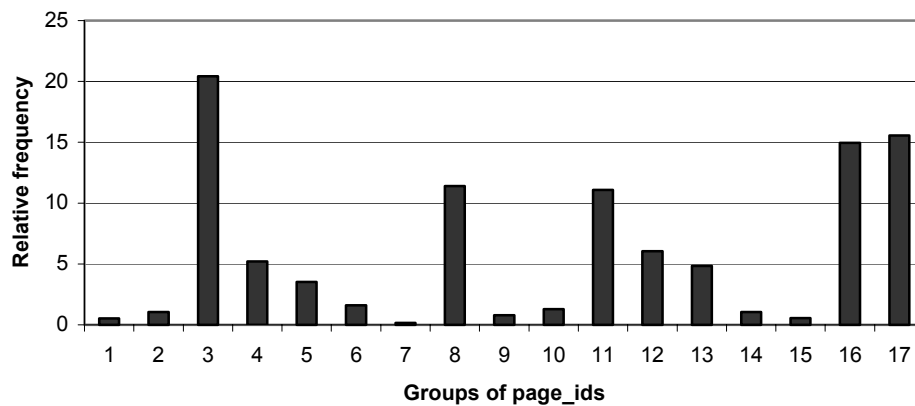
Distribution of 1,159 distinct page_ids in data set 2, represented in ten different groups, based on ten different classes.



Groups of page_ids based on classes for data set 3 (Belgian telecom provider).

Class		Group
French language	Service X	1
	FAQ	2
	Main	3
	New	4
	Products	5
	Range	6
	Sales	7
	Prices	8
Dutch language	Service X	9
	FAQ	10
	Main	11
	New	12
	Products	13
	Range	14
	Sales	15
	Prices	16
French + Dutch language	Remaining	17

Distribution of 492 distinct page_ids in data set 3, represented in seventeen different groups, based on seventeen different classes.



Number of server sessions in each cluster of data set 1, 2 and 3.

Cluster	Dataset		
	1	2	3
1	467	1894	230
2	650	584	262
3	337	362	230
4	1008	185	51
5	277	106	-
6	25	-	-
Total	2764	3131	773

SAM applied to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages: Distribution of web pages in six clusters.

Page_id	Relative frequency					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1	20,48	1,72	97,78	1,8	1,08	0,67
2	11,35	0,68	0,44	0,5	1,68	0,27
3	0,56	0,05	0,22	1,36	0	0,53
4	0,93	0,05	0	2,17	0,54	0,67
5	1,85	0,1	0	1,49	0,27	1,07
6	1,44	0,1	0	1,74	0,07	0,53
7	1,25	0,1	0	1,67	0,2	0,53
8	0,14	0,05	0	3,22	0	0,4
9	14,83	0,94	0,44	0,56	3,3	0,27
10	2,55	0,16	0,44	5,46	0,61	0,4
11	1,81	0,31	0	5,58	0,34	0,53
12	1,25	0,42	0	1,86	0,47	0,4
13	0	0,21	0	0,74	0,07	1,6
14	0,09	0,1	0	0,87	0	1,46
15	0	0	0	0,62	0	1,6
16	0	0,1	0	0,81	0	1,46
17	0	0,05	0	0,87	0	1,6
18	0	0	0	0,68	0	1,6
19	0,14	1,56	0	1,3	0,67	1,73
20	0	0,05	0	0,93	0	1,46
21	0	0	0	0,68	0	1,46
22	0	0	0	0,62	0	1,6
23	0	0	0	0,68	0	1,6
24	0	0,05	0	0,87	0	1,6
25	0	0	0	0,68	0	1,33
26	0	0,94	0	1,24	0,07	2,13
27	0,09	0,68	0	0,81	0	2
28	0,05	0,47	0	0,93	0	2,26
29	0,05	0,21	0	0,99	0	1,6
30	0	0,21	0	1,05	0	2,13
31	0	0,05	0	0,99	0	1,73
32	0	0,62	0	1,12	0,07	1,86
33	0	0,78	0	0,99	0,07	2,13
34	0,05	0,73	0	0,93	0	1,6
35	0	0,36	0	0,81	0,07	2,13
36	0	0,57	0	0,74	0	2,13
37	0	0,26	0	0,68	0	2,13
38	0	1,61	0	0,93	0,13	2,13
39	0,09	1,87	0	1,05	0	2,13
40	0,05	1,72	0	1,36	0,07	2,13
41	0	0,94	0	1,12	0,67	1,73
42	0	2,03	0	1,05	0,2	1,6
43	1,95	9,94	0	1,05	2,76	0,53
44	0,23	1,35	0	1,74	0,61	1,86
45	0,37	1,41	0	1,18	0,2	1,86
46	0	0,47	0	1,43	0,2	1,33
47	0,19	1,98	0	1,05	0,54	1,86
48	0,09	1,41	0	1,36	0,07	1,33
49	1,58	6,66	0	0,43	1,15	0,53
50	0	0	0	0,87	0,4	1,6
51	0	0	0	0,99	0,13	1,46
52	0,05	0	0	0,81	0,2	1,6
53	0,09	0,05	0	0,87	0,07	1,6
54	0	0	0	0,81	0	1,46

55	1,25	1,56	0	1,61	9,43	1,86
56	0,32	2,45	0	0,62	2,16	1,86
57	0,46	0,52	0	1,18	2,96	1,46
58	0,14	0,42	0	2,6	2,9	2
59	0,28	0,21	0	2,05	2,09	2,4
60	0,05	0,05	0	2,23	0,67	2,53
61	0,23	0,1	0	2,23	1,21	2,4
62	0,09	0,05	0	2,29	0,47	2,26
63	0,14	0,21	0	2,17	1,21	2,66
64	0,05	0,05	0	1,49	0,67	1,46
65	6,21	2,86	0,22	1,92	17,92	0,53
66	1,16	0,21	0	2,48	0,34	0,13
67	3,1	0,73	0	1,55	2,49	0,13
68	17,61	38,73	0,44	2,29	28,1	0,4
69	2,13	0,94	0	2,36	1,08	0,27
70	0,51	0,68	0	2,85	5,39	0,53
71	2,69	6,09	0	0,93	3,91	0,13

SAM applied to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages: Exclusivity of web pages in six clusters.

Page_id	Exclusivity					
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1	0,46	0,03	0,46	0,03	0,02	0,01
2	0,83	0,04	0,01	0,03	0,08	0,01
3	0,3	0,03	0,03	0,55	0	0,1
4	0,29	0,01	0	0,51	0,12	0,07
5	0,51	0,03	0	0,31	0,05	0,1
6	0,47	0,03	0	0,42	0,02	0,06
7	0,43	0,03	0	0,43	0,05	0,06
8	0,05	0,02	0	0,88	0	0,05
9	0,8	0,05	0,01	0,02	0,12	0,01
10	0,34	0,02	0,01	0,55	0,06	0,02
11	0,27	0,04	0	0,63	0,03	0,03
12	0,36	0,11	0	0,4	0,09	0,04
13	0	0,14	0	0,41	0,03	0,41
14	0,07	0,07	0	0,48	0	0,38
15	0	0	0	0,45	0	0,55
16	0	0,08	0	0,5	0	0,42
17	0	0,04	0	0,52	0	0,44
18	0	0	0	0,48	0	0,52
19	0,04	0,39	0	0,27	0,13	0,17
20	0	0,04	0	0,56	0	0,41
21	0	0	0	0,5	0	0,5
22	0	0	0	0,45	0	0,55
23	0	0	0	0,48	0	0,52
24	0	0,04	0	0,52	0	0,44
25	0	0	0	0,52	0	0,48
26	0	0,33	0	0,36	0,02	0,29
27	0,05	0,3	0	0,3	0	0,35
28	0,02	0,21	0	0,36	0	0,4
29	0,03	0,12	0	0,48	0	0,36
30	0	0,11	0	0,46	0	0,43
31	0	0,03	0	0,53	0	0,43
32	0	0,27	0	0,4	0,02	0,31
33	0	0,31	0	0,33	0,02	0,33
34	0,02	0,33	0	0,36	0	0,29
35	0	0,19	0	0,35	0,03	0,43

36	0	0,28	0	0,31	0	0,41
37	0	0,16	0	0,34	0	0,5
38	0	0,48	0	0,23	0,03	0,25
39	0,03	0,51	0	0,24	0	0,23
40	0,01	0,45	0	0,3	0,01	0,22
41	0	0,31	0	0,31	0,17	0,22
42	0	0,55	0	0,24	0,04	0,17
43	0,14	0,65	0	0,06	0,14	0,01
44	0,06	0,32	0	0,34	0,11	0,17
45	0,11	0,38	0	0,27	0,04	0,2
46	0	0,2	0	0,51	0,07	0,22
47	0,05	0,47	0	0,21	0,1	0,17
48	0,03	0,44	0	0,35	0,02	0,16
49	0,18	0,67	0	0,04	0,09	0,02
50	0	0	0	0,44	0,19	0,38
51	0	0	0	0,55	0,07	0,38
52	0,03	0	0	0,45	0,1	0,41
53	0,07	0,03	0	0,47	0,03	0,4
54	0	0	0	0,54	0	0,46
55	0,11	0,13	0	0,11	0,59	0,06
56	0,06	0,43	0	0,09	0,29	0,13
57	0,11	0,11	0	0,2	0,47	0,12
58	0,03	0,07	0	0,38	0,39	0,14
59	0,07	0,04	0	0,36	0,34	0,2
60	0,01	0,01	0	0,54	0,15	0,28
61	0,06	0,03	0	0,46	0,23	0,23
62	0,03	0,02	0	0,58	0,11	0,27
63	0,04	0,05	0	0,44	0,23	0,25
64	0,02	0,02	0	0,51	0,21	0,23
65	0,27	0,11	0	0,06	0,54	0,01
66	0,33	0,05	0	0,53	0,07	0,01
67	0,47	0,1	0	0,17	0,26	0,01
68	0,24	0,47	0	0,02	0,26	0
69	0,38	0,15	0	0,32	0,13	0,02
70	0,07	0,08	0	0,3	0,52	0,03
71	0,23	0,47	0	0,06	0,23	0

SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of groups of page_ids in five clusters.

Page_id	Relative frequency				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	1,01	0	0,18	0	0
2	0,27	0	0	0,25	0
3	0,66	0,39	0	0	0
4	0,42	0	0,06	0	0
5	0,04	0	0,06	0	0
6	0,21	0	0	0	0
7	1,63	0,26	0,37	0	0
8	4,56	0,52	1,29	0,74	0,62
9	2,53	0,65	0,43	0,75	0,2
10	1,08	0	0,31	0,25	0
11	0,21	0	0,06	0	0
12	0,28	0,13	0,3	0	0
13	0,8	0,13	0,24	0,25	0,62
14	0,33	0,39	0,06	0	0
15	1,56	0,26	0,37	0	0
16	2,14	0,26	0,06	0	0,31
17	1,26	0,13	0,18	0	0
18	0,67	0	0,12	0	0,1
19	0,91	0,13	0	0	0
20	2,1	0,13	0,12	0	0
21	0,99	0,13	0	0	0,3
22	0,85	0	0	0	0,2
23	1,59	0,26	0,25	0	0,1
24	0,85	0,9	0,12	0	0,1
25	1,31	0	0,06	0	0
26	1,95	0,52	0,12	0,5	0,92
27	0,19	0,13	0	0	0,2
28	0,93	0,52	0	0	0,52
29	11,57	73,53	30,19	14,71	41,07
30	1,18	0	1,35	0,25	0,51
31	2,42	0,52	5,82	15,7	2,07
32	1,85	0,52	0,86	0	0,2
33	0,52	0,13	0,24	0	0,31
34	0,15	0	0,18	0	0,41
35	2,93	9,26	9,23	1,23	1,96
36	1,72	0,65	5,26	0,99	6,6
37	0,36	0	0	0	0
38	2,67	1,56	2,58	1,73	3,82
39	2,43	0,65	1,23	0,99	1,55
40	3,16	0,39	1,5	0	1,96
41	3,65	0,78	1,31	2,22	5,04
42	4,44	1,16	1,8	3,44	5,98
43	3,4	0,78	10,83	2,94	6,4
44	5,76	0,65	7	4,42	4,75
45	7,63	1,17	10,7	4,92	5,16
46	1,85	0,39	0,24	0,49	0,41
47	1,03	0,39	1,03	0,49	1,13
48	4,28	0,78	1,66	1,49	3,82
49	0,24	0,13	0,37	0	0,51
50	4,83	0,91	1,75	41,43	2,06

SAM applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Exclusivity of groups of page_ids in five clusters.

Page_id	Exclusivity				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	0,93	0	0,07	0	0
2	0,92	0	0	0,08	0
3	0,9	0,1	0	0	0
4	0,95	0	0,05	0	0
5	0,67	0	0,33	0	0
6	1	0	0	0	0
7	0,89	0,03	0,08	0	0
8	0,85	0,02	0,1	0,01	0,03
9	0,86	0,04	0,06	0,02	0,02
10	0,88	0	0,1	0,02	0
11	0,9	0	0,1	0	0
12	0,67	0,06	0,28	0	0
13	0,73	0,02	0,09	0,02	0,13
14	0,78	0,17	0,06	0	0
15	0,89	0,03	0,08	0	0
16	0,94	0,02	0,01	0	0,03
17	0,93	0,02	0,05	0	0
18	0,9	0	0,06	0	0,03
19	0,98	0,03	0	0	0
20	0,97	0,01	0,02	0	0
21	0,91	0,02	0	0	0,07
22	0,95	0	0	0	0,05
23	0,9	0,03	0,05	0	0,01
24	0,78	0,15	0,04	0	0,02
25	0,98	0	0,02	0	0
26	0,83	0,04	0,02	0,02	0,09
27	0,73	0,09	0	0	0,18
28	0,81	0,08	0	0	0,1
29	0,24	0,29	0,24	0,03	0,2
30	0,63	0	0,29	0,01	0,07
31	0,35	0,01	0,33	0,23	0,07
32	0,79	0,04	0,15	0	0,02
33	0,74	0,03	0,13	0	0,1
34	0,5	0	0,21	0	0,29
35	0,33	0,2	0,41	0,01	0,05
36	0,31	0,02	0,37	0,02	0,28
37	1	0	0	0	0
38	0,53	0,06	0,2	0,03	0,18
39	0,69	0,03	0,14	0,03	0,1
40	0,74	0,02	0,14	0	0,11
41	0,64	0,03	0,09	0,04	0,21
42	0,62	0,03	0,1	0,05	0,2
43	0,35	0,02	0,44	0,03	0,16
44	0,56	0,01	0,27	0,04	0,11
45	0,55	0,02	0,31	0,04	0,09
46	0,85	0,03	0,04	0,02	0,04
47	0,57	0,04	0,22	0,03	0,14
48	0,7	0,02	0,11	0,02	0,15
49	0,45	0,05	0,27	0	0,23
50	0,47	0,02	0,07	0,4	0,05

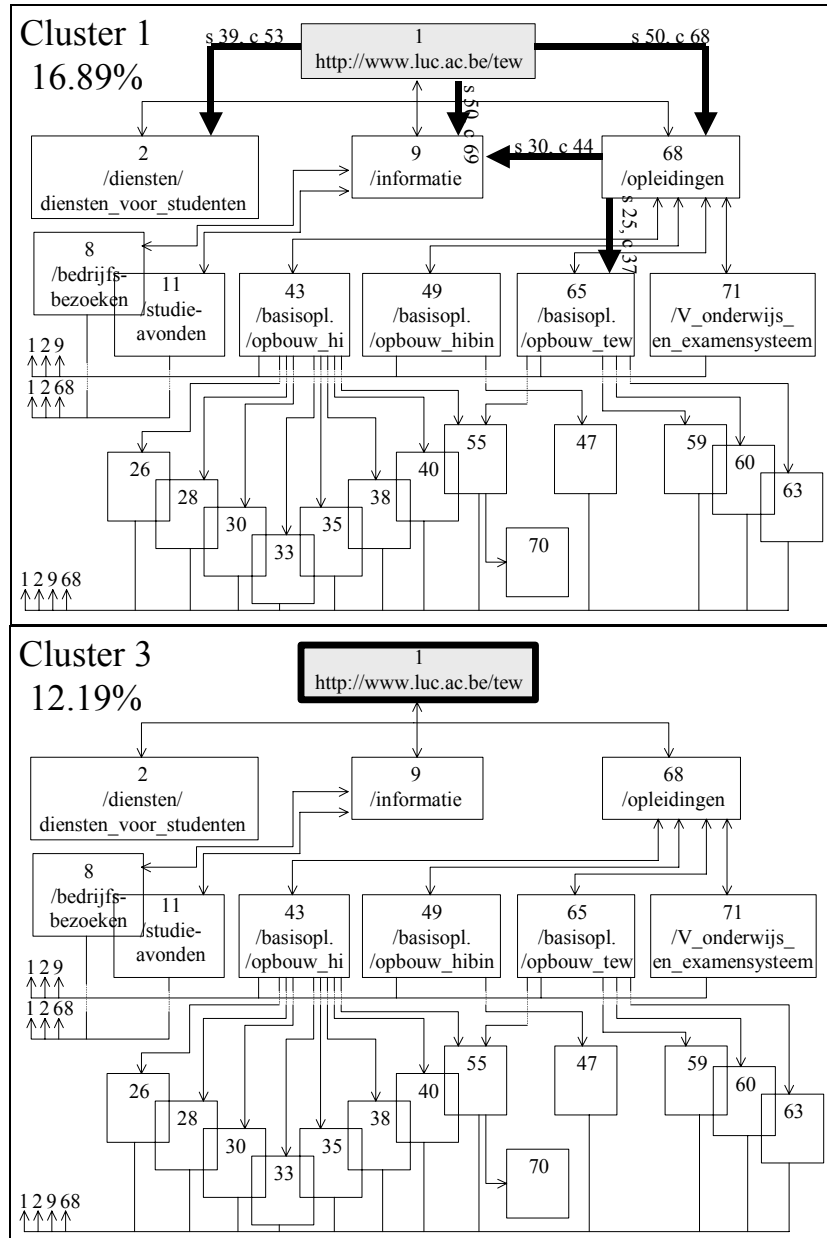
SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of groups of page_ids in four clusters.

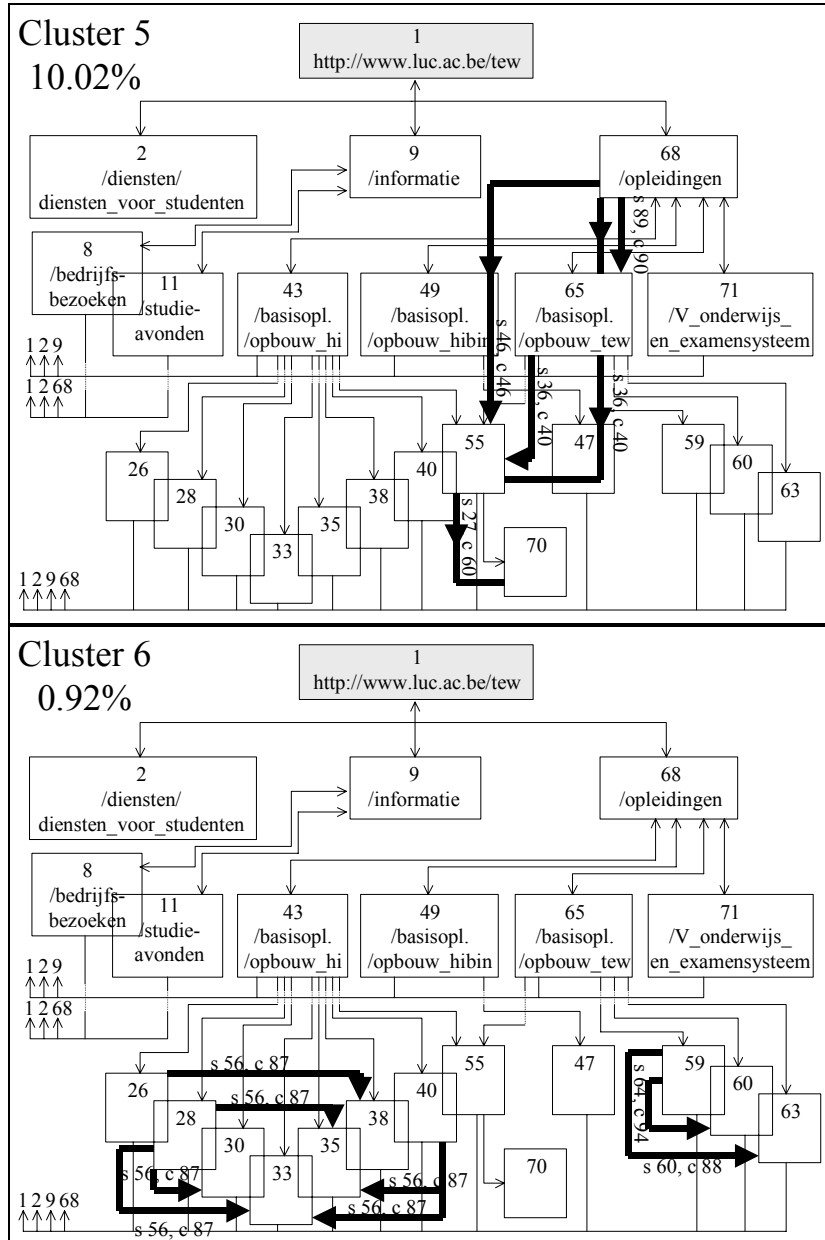
Page_id	Relative frequency			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	1,44	5,25	0	0
2	30,02	6,51	0,34	0
3	3,36	2,79	0	0
4	1,26	1,78	0	0
5	27,91	5,95	0,22	0
6	1,86	4,03	0	0
7	0,72	2,99	0	0
8	3,12	6,7	0,06	0
9	1,98	1,55	0	0
10	26,69	8,07	0,22	0
11	0,24	3,99	2,6	2,24
12	0,36	8,4	28,39	34,08
13	0,06	8,54	1,14	0
14	0	2,01	1,73	0
15	0,18	9,65	28,74	33,43
16	0	2,02	1,37	1,28
17	0	2,78	2,17	0,64
18	0	2,56	2,28	0,32
19	0,24	6,16	6,13	3,2
20	0,42	7,74	24,62	24,76

SAM applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Exclusivity of groups of page_ids in four clusters.

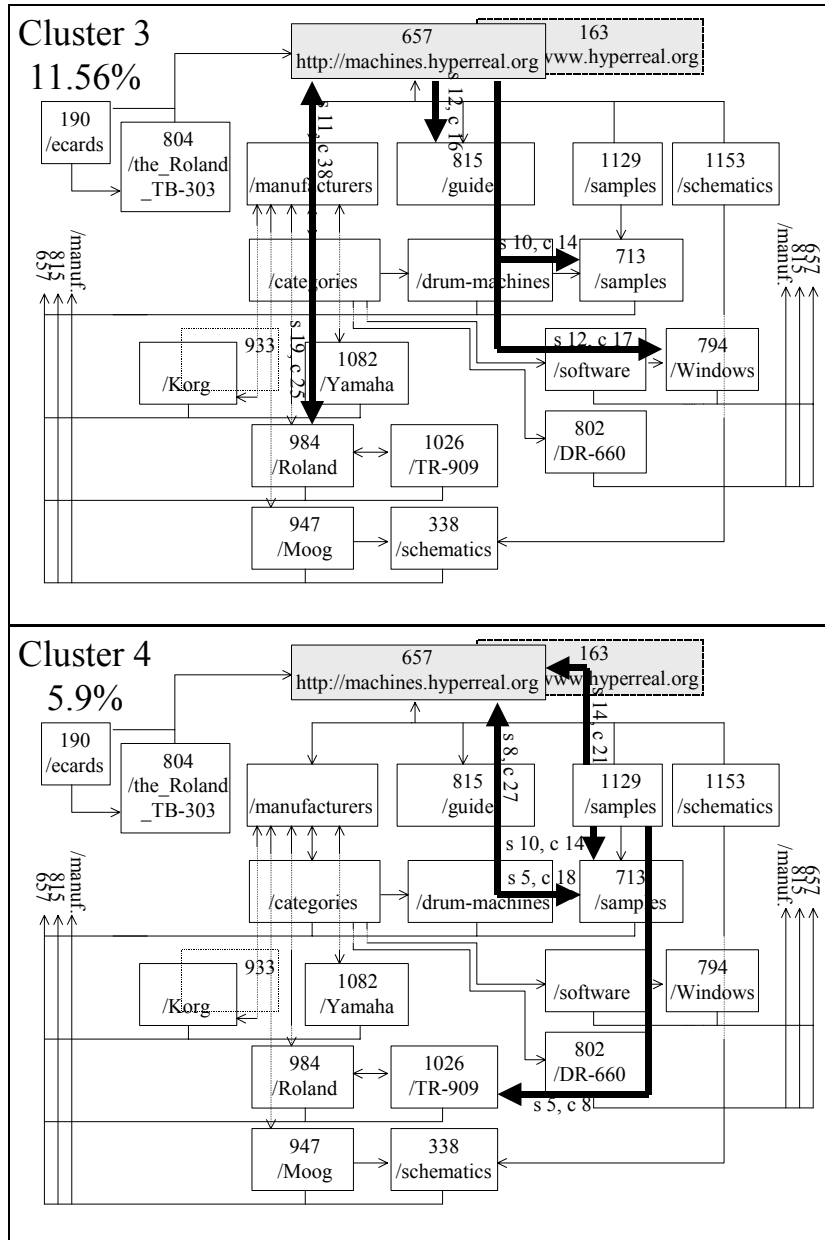
Page_id	Exclusivity			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0,34	0,66	0	0
2	0,89	0,1	0,01	0
3	0,69	0,31	0	0
4	0,57	0,43	0	0
5	0,89	0,1	0,01	0
6	0,46	0,54	0	0
7	0,31	0,69	0	0
8	0,46	0,53	0,01	0
9	0,7	0,3	0	0
10	0,85	0,14	0,01	0
11	0,04	0,39	0,49	0,08
12	0,01	0,11	0,73	0,16
13	0,01	0,78	0,21	0
14	0	0,38	0,63	0
15	0	0,12	0,72	0,15
16	0	0,39	0,52	0,09
17	0	0,38	0,58	0,03
18	0	0,36	0,63	0,02
19	0,02	0,31	0,61	0,06
20	0,01	0,12	0,74	0,13

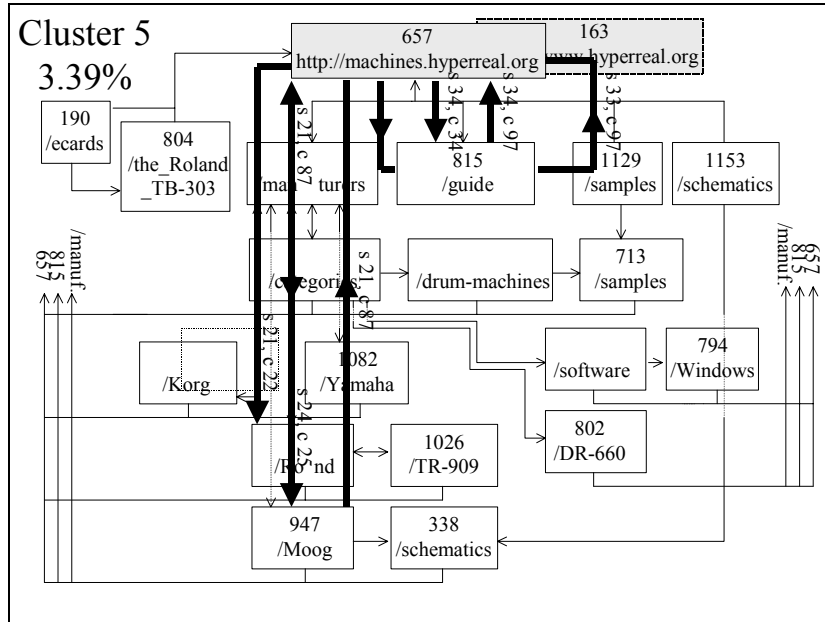
SAM applied to data set 1: Surfing behaviour at <http://www.luc.ac.be/tew>: navigation patterns, providing page and order-based information.



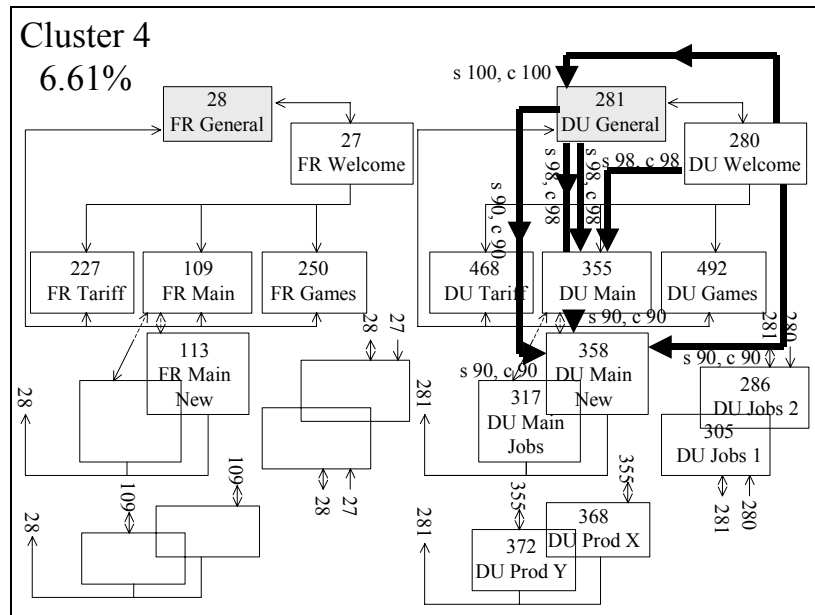


SAM applied to data set 2: Surfing behaviour at <http://machines.hyperreal.org>: navigation patterns, providing page and order-based information.





SAM applied to data set 3: Surfing behaviour at the web site of a Belgian telecom provider: navigation patterns, providing page and order-based information.



Association distance applied to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages: Distribution of web pages in three clusters.

Page_id	Relative frequency		
	Cluster 1	Cluster 2	Cluster 3
1	13,32	17,44	4,14
2	4,44	1,4	3,01
3	0,41	0,63	0,5
4	0,58	1,75	0,73
5	0,98	0,91	0,82
6	0,77	1,05	0,64
7	0,7	0,77	0,82
8	0,41	2,38	0,23
9	6,43	1,54	3,46
10	1,13	5,74	1,09
11	1,3	4,34	0,87
12	0,68	1,12	1,18
13	0,11	0,49	0,73
14	0,06	0,63	0,73
15	0,06	0,35	0,59
16	0,15	0,28	0,64
17	0,09	0,63	0,59
18	0,04	0,49	0,59
19	0,73	0,63	1,55
20	0,02	0,7	0,68
21	0	0,7	0,5
22	0,09	0,35	0,55
23	0,09	0,42	0,55
24	0,09	0,56	0,64
25	0,11	0,42	0,41
26	0,38	0,56	1,32
27	0,15	0,63	1,18
28	0,21	0,49	1,14
29	0,09	0,7	0,82
30	0,17	0,7	0,87
31	0,11	0,49	0,77
32	0,21	0,56	1,18
33	0,32	0,63	1,05
34	0,28	0,84	0,73
35	0,28	0,35	0,82
36	0,19	0,56	0,96
37	0,13	0,42	0,87
38	0,38	0,56	1,68
39	0,51	0,63	1,73
40	0,62	0,84	1,46
41	0,38	0,7	1,37
42	0,64	0,42	1,55
43	4,14	1,33	3,73
44	0,66	1,26	1,46
45	0,73	0,84	1,09
46	0,26	0,84	0,91
47	0,85	0,77	1,37
48	0,66	0,91	0,82
49	2,84	1,05	1,87
50	0,17	0,56	0,68
51	0,17	0,35	0,68
52	0,06	0,56	0,77
53	0,13	0,7	0,59
54	0,02	0,49	0,68

55	2,99	1,33	3,55
56	1,37	0,77	1,55
57	1	0,56	1,73
58	1,07	1,75	1,55
59	0,79	1,05	1,73
60	0,38	1,19	1,37
61	0,56	1,26	1,5
62	0,38	1,26	1,18
63	0,62	1,33	1,37
64	0,41	0,84	0,68
65	8,3	0,07	4,55
66	0,7	1,54	0,87
67	1,94	0,84	1,82
68	24,72	14,29	9,97
69	1,2	1,33	2
70	1,52	2,73	1,96
71	3,52	2,38	2,23

Association distance applied to data set 1 (<http://www.luc.ac.be>), server sessions consisting of visited pages: Exclusivity of web pages in three clusters.

Page_id	Exclusivity		
	Cluster 1	Cluster 2	Cluster 3
1	0,65	0,26	0,09
2	0,71	0,07	0,22
3	0,49	0,23	0,28
4	0,4	0,37	0,24
5	0,6	0,17	0,23
6	0,55	0,23	0,22
7	0,53	0,18	0,29
8	0,33	0,59	0,09
9	0,75	0,06	0,19
10	0,33	0,52	0,15
11	0,43	0,44	0,13
12	0,43	0,22	0,35
13	0,18	0,25	0,57
14	0,11	0,32	0,57
15	0,14	0,24	0,62
16	0,28	0,16	0,56
17	0,15	0,35	0,5
18	0,09	0,32	0,59
19	0,44	0,12	0,44
20	0,04	0,38	0,58
21	0	0,48	0,52
22	0,19	0,24	0,57
23	0,18	0,27	0,55
24	0,15	0,31	0,54
25	0,25	0,3	0,45
26	0,33	0,15	0,53
27	0,17	0,21	0,62
28	0,24	0,17	0,6
29	0,13	0,31	0,56
30	0,22	0,27	0,51
31	0,17	0,24	0,59
32	0,23	0,18	0,59
33	0,32	0,19	0,49
34	0,32	0,29	0,39
35	0,36	0,14	0,5
36	0,24	0,21	0,55

37	0,19	0,19	0,61
38	0,29	0,13	0,59
39	0,34	0,13	0,54
40	0,4	0,16	0,44
41	0,31	0,17	0,52
42	0,43	0,09	0,49
43	0,66	0,06	0,28
44	0,38	0,22	0,4
45	0,49	0,17	0,34
46	0,27	0,27	0,45
47	0,49	0,14	0,37
48	0,5	0,21	0,29
49	0,7	0,08	0,22
50	0,26	0,26	0,48
51	0,29	0,18	0,54
52	0,11	0,29	0,61
53	0,21	0,34	0,45
54	0,04	0,3	0,65
55	0,59	0,08	0,33
56	0,59	0,1	0,31
57	0,51	0,09	0,41
58	0,46	0,23	0,31
59	0,41	0,17	0,42
60	0,28	0,26	0,46
61	0,34	0,23	0,43
62	0,29	0,29	0,42
63	0,37	0,24	0,38
64	0,41	0,26	0,33
65	0,79	0	0,2
66	0,45	0,3	0,26
67	0,64	0,08	0,28
68	0,73	0,13	0,14
69	0,47	0,16	0,37
70	0,46	0,25	0,28
71	0,67	0,14	0,2

Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Distribution of groups of page_ids in three clusters.

Page_id	Relative frequency		
	Cluster 1	Cluster 2	Cluster 3
1	0,69	0,61	0,47
2	0,05	0,04	0,34
3	0,23	0,58	0,34
4	0,05	0,16	0,43
5	0	0,04	0,06
6	0,05	0,08	0,18
7	0,69	1,35	0,87
8	5,01	3,02	1,12
9	1,74	2,52	0,64
10	0,23	1,38	0,34
11	0,05	0,28	0,06
12	0,14	0,24	0,27
13	0,45	0,78	0,5
14	0,09	0,53	0,09
15	0,54	1,51	0,71
16	0,28	1,67	1,46
17	0,33	1,05	0,7
18	0,1	0,81	0,27
19	0,46	0,4	0,61
20	0,46	1,34	1,49
21	0,47	0,64	0,57
22	0,29	0,32	0,7
23	1	0,56	1,13
24	0,57	0,48	0,67
25	0,65	0,68	0,77
26	1,52	1,42	0,92
27	0,14	0,16	0,12
28	0,69	0,89	0,34
29	27,7	22,43	25,8
30	1,54	0,52	0,89
31	3,77	3,51	3,51
32	1,56	1,18	1,02
33	0,47	0,52	0,24
34	0,2	0,04	0,27
35	4,7	5,49	3,95
36	2	2,46	3,89
37	0,27	0,16	0,15
38	1,62	2,73	3,31
39	1,69	2,4	1,47
40	2,98	2,44	1,54
41	3,12	2,9	2,93
42	3,7	3,71	3,72
43	3,75	3,98	6,66
44	6,8	4,35	5,01
45	7,83	4,86	8,54
46	1,04	0,96	1,32
47	0,84	0,85	1,15
48	3,06	3,01	3,45
49	0,18	0,24	0,37
50	5,02	7,28	4,25

Association distance applied to data set 2 (<http://machines.hyperreal.org>), server sessions consisting of visited pages: Exclusivity of groups of page_ids in three clusters.

Page_id	Exclusivity		
	Cluster 1	Cluster 2	Cluster 3
1	0,33	0,33	0,33
2	0,08	0,08	0,85
3	0,17	0,47	0,37
4	0,05	0,21	0,74
5	0	0,33	0,67
6	0,11	0,22	0,67
7	0,2	0,43	0,37
8	0,5	0,33	0,16
9	0,31	0,51	0,17
10	0,1	0,68	0,22
11	0,1	0,7	0,2
12	0,17	0,33	0,5
13	0,22	0,42	0,36
14	0,11	0,72	0,17
15	0,17	0,51	0,32
16	0,06	0,44	0,5
17	0,13	0,46	0,41
18	0,06	0,65	0,29
19	0,25	0,25	0,5
20	0,11	0,36	0,53
21	0,22	0,36	0,42
22	0,16	0,22	0,62
23	0,3	0,19	0,51
24	0,26	0,26	0,48
25	0,25	0,3	0,45
26	0,34	0,36	0,31
27	0,27	0,36	0,36
28	0,31	0,46	0,23
29	0,31	0,27	0,42
30	0,45	0,17	0,38
31	0,3	0,31	0,4
32	0,35	0,3	0,34
33	0,32	0,42	0,26
34	0,29	0,07	0,64
35	0,28	0,37	0,35
36	0,19	0,26	0,55
37	0,4	0,27	0,33
38	0,17	0,32	0,51
39	0,26	0,41	0,33
40	0,38	0,34	0,28
41	0,29	0,3	0,4
42	0,28	0,31	0,41
43	0,21	0,25	0,54
44	0,36	0,26	0,39
45	0,31	0,21	0,49
46	0,25	0,27	0,48
47	0,24	0,28	0,49
48	0,27	0,29	0,44
49	0,18	0,27	0,55
50	0,26	0,42	0,32

Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Distribution of groups of page_ids in four clusters.

Page_id	Relative frequency			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	3,08	2,03	1,12	0
2	10,83	8,57	20,98	0,2
3	3,57	3,45	1,01	0
4	1,54	2,01	0,45	0
5	11,53	7,72	18,23	0,2
6	2,17	2,3	1,59	0
7	0,84	3,45	0,87	0
8	3,99	10,31	1,17	0
9	2,24	2,01	0,45	0
10	13,7	4,31	17,11	0,2
11	3,15	4,33	0,85	1,64
12	8,38	9,16	10,63	33,01
13	1,26	4,31	3,3	0,49
14	2,03	2,59	0,46	0,19
15	10,41	8,3	10,75	31,46
16	1,4	2,01	0,5	0,97
17	2,31	4,04	0,69	0,58
18	1,26	4,32	1,25	0,87
19	5,95	6,58	1,59	3,87
20	10,48	8,57	7,32	26,31

Association distance applied to data set 3 (Belgian telecom provider), server sessions consisting of visited pages: Exclusivity of groups of page_ids in four clusters.

Page_id	Exclusivity			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0,62	0,1	0,28	0
2	0,28	0,05	0,67	0
3	0,63	0,15	0,22	0
4	0,59	0,19	0,22	0
5	0,32	0,05	0,63	0
6	0,46	0,12	0,42	0
7	0,31	0,31	0,38	0
8	0,5	0,32	0,18	0
9	0,68	0,15	0,17	0
10	0,38	0,03	0,59	0
11	0,49	0,16	0,16	0,18
12	0,18	0,05	0,28	0,5
13	0,19	0,15	0,61	0,05
14	0,6	0,19	0,17	0,04
15	0,21	0,04	0,28	0,47
16	0,43	0,15	0,2	0,22
17	0,51	0,22	0,18	0,09
18	0,28	0,23	0,34	0,14
19	0,48	0,13	0,16	0,23
20	0,26	0,05	0,22	0,47

APPENDIX 5

Algorithm, used by SAM¹, for transforming server sessions into sessions with interesting combinations of pages, respecting the order of pages.

```

begin
  while not eof(server_sessions_original) do           //read input file//
    begin
      i:=0;
      k:=0;
      while not eoln(server_sessions_original) do
        begin
          i:=i+1;
          read(server_sessions_original,page[i]); //read the server session as array page[i]//
        end;
        readln(server_sessions_original);           //proceed to the next line of the input file//
        if i>1 then //original server sessions must consist of minimum two pages in order to
          begin be relevant for interestingness based on the support logic framework//
            Reset(interesting_related_pages); //open and read file with interesting related
            while not eof(interesting_related_pages) do pages//
              begin
                j:=0;
                while not eoln(interesting_related_pages) do
                  begin
                    j:=j+1;
                    read(interesting_related_pages,intpage[j]); //read the interesting related pages
                  end; //as array intpage[j]//
                readln(interesting_related_pages); //proceed to the next line of the file//
                if i>=j then //if i<j, the original server session has less pages than the
                  begin interesting frequent item set and cannot hold the interesting
                    for x:=1 to j do related pages// //the interesting frequent item set consists of
                      begin j related pages//
                        teller[x]:=0;
                      end
                    for a:=1 to i do
                      begin
                        for x:=1 to j do
                          begin
                            if page[a]=intpage[x] then teller[x]:=teller[x]+1;
                          end;
                        end;
                      end;
                    present:=0;
                    for x:=1 to j do
                      begin
                        if (teller[x]<>0) then present:= present + 1;
                      end;
                    if present = j then
                      begin //the interesting frequent item set is found in the server session//

```

```

        for a:=1 to j do
        begin
            k:=k+1;
            interesting[k]:=intpage[a]; //the interesting related pages found in the input
        end; //file are written in array interesting[k]//
        end;
        end;
        CloseFile(interesting_related_pages);
end;
if k<>0 then //k<>0 indicates that at least one interesting frequent item set
begin //of related pages is found in the original server session//
    for a:=1 to i do
    begin
        S:=0;
        for b:=1 to k do
        begin
            if (page[a]=interesting[b]) and (S=0) then
            begin
                S:=1;
                write(server_sessions_transformed,',page[a]); //write the transformed
            end; //server session, including
            end; //only interesting related
            end; //pages with respect of the
            writeln(server_sessions_transformed); //order, in the output file//
        end;
    end;
    end;
    CloseFile(server_sessions_original);
    CloseFile(server_sessions_transformed);
end;

```


Interesting beliefs of related pages on <http://machines.hyperreal.org>.

Interesting beliefs of related pages	Evidence			$IM_{\beta_i} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
(657, 162)	[0.0101; 0.0101]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3999 (-)	1.3999 (-)	-
(657, 159)	[0.0102; 0.0102]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3997 (-)	1.3997 (-)	-
(815, 657, 810)	[0.0123; 0.0123]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3968 (-)	1.3968 (-)	-
(657, 349)	[0.0167; 0.0167]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3905 (-)	1.3905 (-)	-
(657, 810)	[0.0210; 0.0210]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3845 (-)	1.3845 (-)	-
(815, 163)	[0.0273; 0.0273]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3756 (-)	1.3756 (-)	-
(657, 813)	[0.0345; 0.0345]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3654 (-)	1.3654 (-)	-
(984, 163)	[0.0356; 0.0356]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3638 (-)	1.3638 (-)	-
(163, 162)	[0.0417; 0.0417]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3552 (-)	1.3552 (-)	-
(984, 985)	[0.0473; 0.0473]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3473 (-)	1.3473 (-)	-
(984, 987)	[0.0538; 0.0538]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3381 (-)	1.3381 (-)	-
(984, 1012)	[0.0550; 0.0550]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3364 (-)	1.3364 (-)	-
(984, 1000)	[0.0650; 0.0650]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3222 (-)	1.3222 (-)	-
(984, 1022)	[0.0651; 0.0651]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3221 (-)	1.3221 (-)	-
(984, 995)	[0.0698; 0.0698]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3155 (-)	1.3155 (-)	-
(1082, 1093)	[0.0730; 0.0730]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3109 (-)	1.3109 (-)	-
(984, 1024)	[0.0737; 0.0737]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3099 (-)	1.3099 (-)	-
(1082, 1091)	[0.0746; 0.0746]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3087 (-)	1.3087 (-)	-
(163, 159)	[0.0748; 0.0748]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3084 (-)	1.3084 (-)	-
(984, 1001)	[0.0770; 0.0770]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3053 (-)	1.3053 (-)	-
(984, 1013)	[0.0834; 0.0834]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2962 (-)	1.2962 (-)	-
(984, 997)	[0.0876; 0.0876]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2903 (-)	1.2903 (-)	-

Interesting beliefs of related pages	Evidence			$IM_{\beta_i} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
(984, 990)	[0.0878; 0.0878]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2900 (-)	1.2900 (-)	-
(984, 1025)	[0.0922; 0.0922]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2838 (-)	1.2838 (-)	-
(984, 1021)	[0.0938; 0.0938]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2815 (-)	1.2815 (-)	-
(984, 999)	[0.0938; 0.0938]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2815 (-)	1.2815 (-)	-
(815, 810)	[0.0951; 0.0951]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2797 (-)	1.2797 (-)	-
(1082, 1083)	[0.0995; 0.0995]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2734 (-)	1.2734 (-)	-
(452, 349)	[0.1149; 0.1149]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2517 (-)	1.2517 (-)	-
(1026, 984)	[0.1151; 0.1151]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2514 (-)	1.2514 (-)	-
(163, 349)	[0.1221; 0.1221]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2415 (-)	1.2415 (-)	-
(349, 627)	[0.1239; 0.1239]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2389 (-)	1.2389 (-)	-
(984, 998)	[0.1242; 0.1242]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2385 (-)	1.2385 (-)	-
(984, 1018)	[0.1300; 0.1300]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2303 (-)	1.2303 (-)	-
(815, 657)	[0.1338; 0.1338]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2249 (-)	1.2249 (-)	-
(159, 162)	[0.1418; 0.1418]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2136 (-)	1.2136 (-)	-
(349, 162)	[0.1588; 0.1588]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.1896 (-)	1.1896 (-)	-
(868, 857)	[0.1645; 0.1645]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.1815 (-)	1.1815 (-)	-
(984, 996)	[0.1855; 0.1855]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.1518 (-)	1.1518 (-)	-
(815, 657, 813)	[0.0189; 0.0189]	[0.8334; 0.8334]	[0.0879; 0.0879]	1.1518 (-)	-	1.0543 (+)
(349, 159)	[0.2092; 0.2092]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.1183 (-)	1.1183 (-)	-
(1034, 1040)	[0.2349; 0.2349]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0820 (-)	1.0820 (-)	-
(349, 524)	[0.2367; 0.2367]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0794 (-)	1.0794 (-)	-
(62, 171)	[0.2372; 0.2372]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0787 (-)	1.0787 (-)	-
(1034, 1041)	[0.2798; 0.2798]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.0185 (-)	1.0185 (-)	-
(857, 859)	[0.3382; 0.3382]	[1.0000; 1.0000]	[1.0000; 1.0000]	0.9359 (-)	0.9359 (-)	-
(882, 883)	[0.3388; 0.3388]	[1.0000; 1.0000]	[1.0000; 1.0000]	0.9350 (-)	0.9350 (-)	-
(1134, 815, 657)	[0.0069; 0.0069]	[0.6667; 0.6667]	[0.0137; 0.0137]	0.9330 (-)	-	0.9235 (+)

Interesting beliefs of related pages	Evidence			$IM_{bt} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
(815, 657, 698)	[0.0072; 0.0072]	[0.6667; 0.6667]	[0.0143; 0.0143]	0.9326 (-)	-	0.9226 (+)
(657, 984, 985)	[0.0072; 0.0072]	[0.6667; 0.6667]	[0.0143; 0.0143]	0.9326 (-)	-	0.9226 (+)
(657, 984, 1028)	[0.0081; 0.0081]	[0.6667; 0.6667]	[0.0161; 0.0161]	0.9314 (-)	-	0.9201 (+)
(657, 984, 810)	[0.0084; 0.0084]	[0.6667; 0.6667]	[0.0167; 0.0167]	0.9310 (-)	-	0.9192 (+)
(815, 657, 1026, 984)	[0.0084; 0.0084]	[0.6667; 0.6667]	[0.0167; 0.0167]	0.9309 (-)	-	0.9192 (+)
(815, 657, 1129)	[0.0087; 0.0087]	[0.6667; 0.6667]	[0.0173; 0.0173]	0.9305 (-)	-	0.9184 (+)
(657, 984, 992)	[0.0090; 0.0090]	[0.6667; 0.6667]	[0.0178; 0.0178]	0.9301 (-)	-	0.9177 (+)
(657, 984, 987)	[0.0093; 0.0093]	[0.6667; 0.6667]	[0.0184; 0.0184]	0.9297 (-)	-	0.9168 (+)
(815, 657, 1018)	[0.0093; 0.0093]	[0.6667; 0.6667]	[0.0184; 0.0184]	0.9297 (-)	-	0.9168 (+)
(657, 984, 991)	[0.0096; 0.0096]	[0.6667; 0.6667]	[0.0190; 0.0190]	0.9292 (-)	-	0.9160 (+)
(657, 868, 857)	[0.0096; 0.0096]	[0.6667; 0.6667]	[0.0190; 0.0190]	0.9292 (-)	-	0.9160 (+)
(657, 984, 1012)	[0.0102; 0.0102]	[0.6667; 0.6667]	[0.0202; 0.0202]	0.9284 (-)	-	0.9143 (+)
(657, 984, 1024)	[0.0105; 0.0105]	[0.6667; 0.6667]	[0.0208; 0.0208]	0.9280 (-)	-	0.9134 (+)
(657, 984, 995)	[0.0105; 0.0105]	[0.6667; 0.6667]	[0.0208; 0.0208]	0.9280 (-)	-	0.9134 (+)
(657, 882, 883)	[0.0108; 0.0108]	[0.6667; 0.6667]	[0.0214; 0.0214]	0.9275 (-)	-	0.9126 (+)
(657, 882, 886)	[0.0108; 0.0108]	[0.6667; 0.6667]	[0.0214; 0.0214]	0.9275 (-)	-	0.9126 (+)
(657, 984, 1022)	[0.0111; 0.0111]	[0.6667; 0.6667]	[0.0220; 0.0220]	0.9271 (-)	-	0.9117 (+)
(657, 984, 1000)	[0.0117; 0.0117]	[0.6667; 0.6667]	[0.0231; 0.0231]	0.9263 (-)	-	0.9102 (+)
(815, 657, 786)	[0.0117; 0.0117]	[0.6667; 0.6667]	[0.0231; 0.0231]	0.9263 (-)	-	0.9102 (+)
(657, 1082, 1092)	[0.0123; 0.0123]	[0.6667; 0.6667]	[0.0243; 0.0243]	0.9254 (-)	-	0.9085 (+)
(657, 984, 1001)	[0.0129; 0.0129]	[0.6667; 0.6667]	[0.0255; 0.0255]	0.9246 (-)	-	0.9068 (+)
(815, 657, 933)	[0.0132; 0.0132]	[0.6667; 0.6667]	[0.0261; 0.0261]	0.9241 (-)	-	0.9059 (+)
(815, 657, 1026)	[0.0138; 0.0138]	[0.6667; 0.6667]	[0.0272; 0.0272]	0.9233 (-)	-	0.9044 (+)
(657, 984, 1021)	[0.0138; 0.0138]	[0.6667; 0.6667]	[0.0272; 0.0272]	0.9233 (-)	-	0.9044 (+)
(657, 984, 999)	[0.0144; 0.0144]	[0.6667; 0.6667]	[0.0284; 0.0284]	0.9224(-)	-	0.9027 (+)
(657, 984, 997)	[0.0144; 0.0144]	[0.6667; 0.6667]	[0.0284; 0.0284]	0.9224 (-)	-	0.9027 (+)

Interesting beliefs of related pages	Evidence			$IM_{bt} \geq 0.75$		
	Usage	Structure	Combined	Usage - Structure	Usage - Combined	Structure - Combined
(657, 984, 990)	[0.0141; 0.0141]	[0.6667; 0.6667]	[0.0278; 0.0278]	0.9224 (-)	-	0.9027 (+)
(815, 657, 663)	[0.0150; 0.0150]	[0.6667; 0.6667]	[0.0296; 0.0296]	0.9216 (-)	-	0.9010 (+)
(657, 984, 1013)	[0.0162; 0.0162]	[0.6667; 0.6667]	[0.0319; 0.0319]	0.9199 (-)	-	0.8977 (+)
(815, 657, 1082)	[0.0162; 0.0162]	[0.6667; 0.6667]	[0.0319; 0.0319]	0.9199 (-)	-	0.8977 (+)
(657, 984, 1025)	[0.0168; 0.0168]	[0.6667; 0.6667]	[0.0330; 0.0330]	0.9190 (-)	-	0.8962 (+)
(815, 657, 713)	[0.0183; 0.0183]	[0.6667; 0.6667]	[0.0359; 0.0359]	0.9169 (-)	-	0.8921 (+)
(815, 794, 657)	[0.0189; 0.0189]	[0.6667; 0.6667]	[0.0371; 0.0371]	0.9161 (-)	-	0.8904 (+)
(815, 657, 947)	[0.0192; 0.0192]	[0.6667; 0.6667]	[0.0377; 0.0377]	0.9157 (-)	-	0.8895 (+)
(657, 984, 1006)	[0.0195; 0.0195]	[0.6667; 0.6667]	[0.0383; 0.0383]	0.9152 (-)	-	0.8887 (+)
(657, 984, 998)	[0.0210; 0.0210]	[0.6667; 0.6667]	[0.0411; 0.0411]	0.9131 (-)	-	0.8847 (+)
(815, 1026, 984)	[0.0237; 0.0237]	[0.6667; 0.6667]	[0.0463; 0.0463]	0.9093 (-)	-	0.8774 (+)
(657, 984, 1018)	[0.0252; 0.0252]	[0.6667; 0.6667]	[0.0492; 0.0492]	0.9072 (-)	-	0.8733 (+)
(657, 1026, 984)	[0.0297; 0.0297]	[0.6667; 0.6667]	[0.0577; 0.0577]	0.9008 (-)	-	0.8613 (+)
(815, 657, 984)	[0.0354; 0.0354]	[0.6667; 0.6667]	[0.0684; 0.0684]	0.8927 (-)	-	0.8461 (+)
(386, 206, 385)	[0.0372; 0.0372]	[0.6667; 0.6667]	[0.0717; 0.0717]	0.8902 (-)	-	0.8415(+)
(657, 984, 996)	[0.0372; 0.0372]	[0.6667; 0.6667]	[0.0717; 0.0717]	0.8902 (-)	-	0.8415(+)
(815, 794, 657, 786)	[0.0116; 0.0116]	[0.5834; 0.5834]	[0.0162; 0.0162]	0.8086 (-)	-	0.8021 (+)

APPENDIX 6

Algorithm for sampled set.

```
...
var                                     //define variables used throughout the data
t, a, b, N, l, s: integer               sampling program//
function integerrandom(a, b: integer): integer //define equal probability function for
                                         integer values i of or between a and b//
...
begin
  for N:= 1 to 50 do                     //repeat following command 50 times to end
    begin                                up with 500 sequences in output file//
      for s:= 1 to 6 do                  //sample six sequences with length = 1//
        begin
          t = trunc(random * (b-a+1)) + a
          integerrandom = t
          write(output, t)
          writeln(output)
        end
        for s:= 1 to 1 do                //sample one sequence with length = 2//
          begin
            for l:= 1 to 2 do
              begin
                t = trunc(random * (b-a+1)) + a
                integerrandom = t
                if l < 2 then write(output, t, ' ')
                else write(output, t)
              end
            end
            writeln(output)
          end
          for s:= 1 to 1 do ...           //sample one sequence with length = 3//
          for s:= 1 to 1 do ...           //sample one sequence with length = 4//
          for s:= 1 to 1 do ...           //sample one sequence with length = 5//
        end
      end
    end
  ...
begin
  integerrandom(1, 20)                   //a and b are identified by 1 and 20//
end;
```

Dissimilarity table for experiment 100 1 2.

Experiment 100_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	0	0.6265	0.729	0.704	0.4385	0.7335
	20 12 15 10		0	0.5525	0.557	0.4525	0.6245
	20 7 15 1			0	0.201	0.8035	0.1395
	100 2 5 1				0	0.739	0.1555
	100 12 15 10					0	0.8185
	100 7 15 1						0

Correlation table for experiment 100 1 2.

Experiment 100_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	1	-0.253	-0.458	-0.408	0.683	-0.467
	20 12 15 10		1	-0.105	-0.114	0.095	-0.249
	20 7 15 1			1	0.598	-0.607	0.721
	100 2 5 1				1	-0.478	0.689
	100 12 15 10					1	-0.637
	100 7 15 1						1

Sensitivity table for experiment 100 1 2.

Experiment 100_1_2		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	ZERO	MAJOR	MAJOR	MAJOR	MAJOR	MAJOR
	20 12 15 10		ZERO	MAJOR	MAJOR	MAJOR	MAJOR
	20 7 15 1			ZERO	CONS	MAJOR	CONS
	100 2 5 1				ZERO	MAJOR	CONS
	100 12 15 10					ZERO	MAJOR
	100 7 15 1						ZERO

Dissimilarity table for experiment 2 4 10.

Experiment 2_4_10		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	0	0.572	0.608	0.689	0.2085	0.8495
	20 12 15 10		0	0.48	0.481	0.482	0.5745
	20 7 15 1			0	0.2695	0.6695	0.2755
	100 2 5 1				0	0.6305	0.3745
	100 12 15 10					0	0.75
	100 7 15 1						0

Correlation table for experiment 2 4 10.

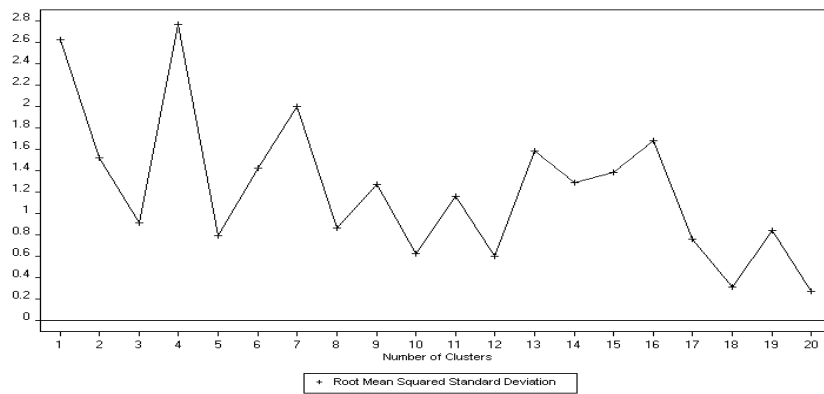
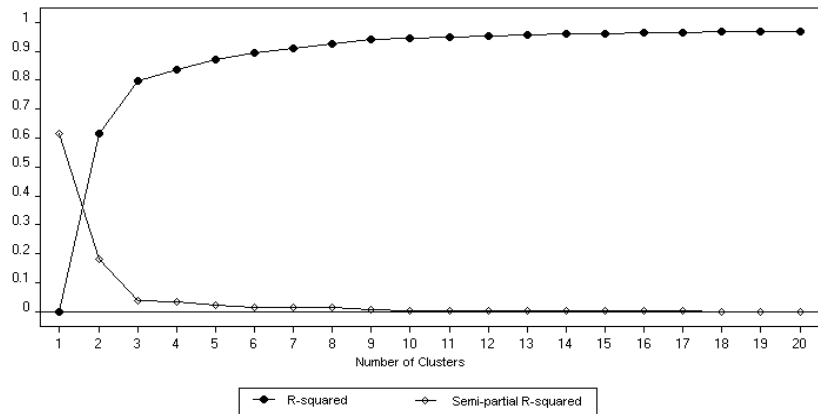
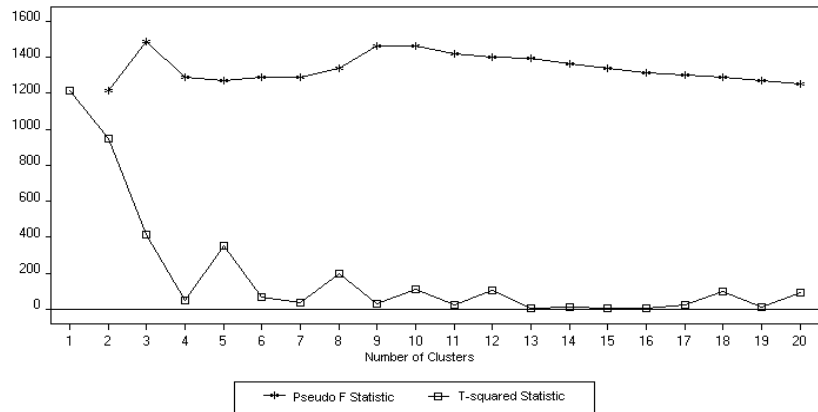
Experiment 2_4_10		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	1	-0.144	-0.216	-0.378	0.583	-0.699
	20 12 15 10		1	0.04	0.038	0.036	-0.149
	20 7 15 1			1	0.461	-0.339	0.449
	100 2 5 1				1	-0.261	0.251
	100 12 15 10					1	-0.5
	100 7 15 1						1

Sensitivity table for experiment 2 4 10.

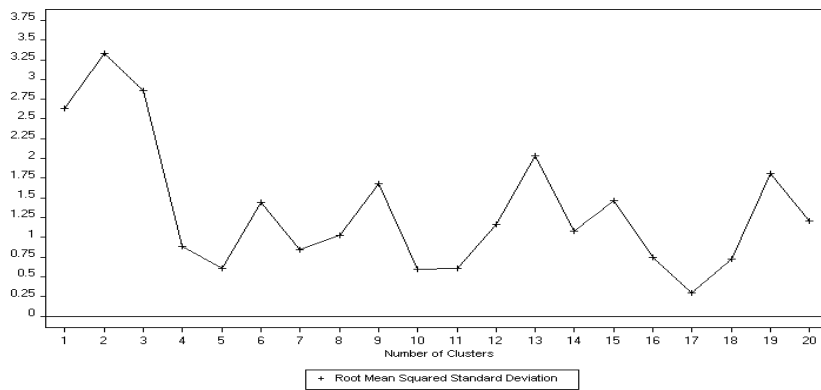
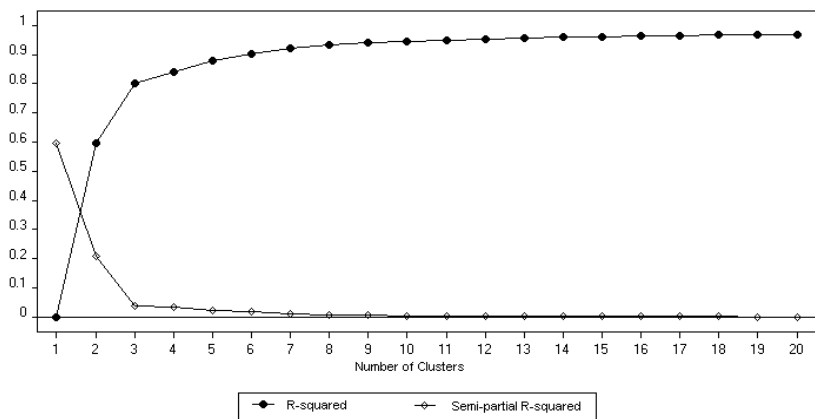
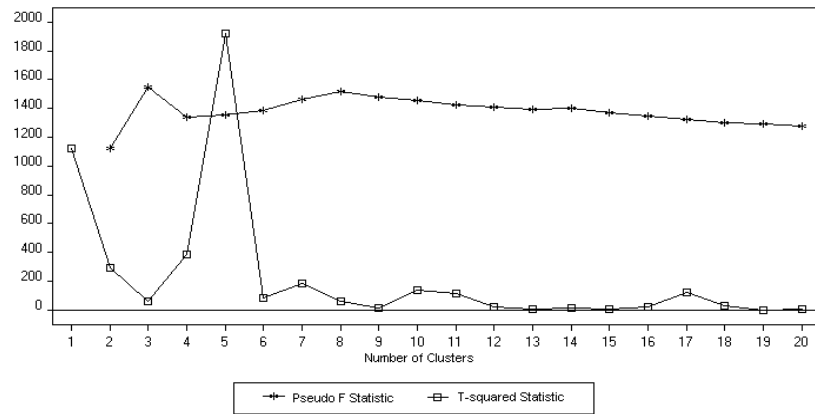
Experiment 2_4_10		S A M P L E D S E T					
		20 2 5 1	20 12 15 10	20 7 15 1	100 2 5 1	100 12 15 10	100 7 15 1
SAMPLED SET	20 2 5 1	ZERO	MAJOR	MAJOR	MAJOR	CONS	MAJOR
	20 12 15 10		ZERO	MAJOR	MAJOR	MAJOR	MAJOR
	20 7 15 1			ZERO	MAJOR	MAJOR	MAJOR
	100 2 5 1				ZERO	MAJOR	MAJOR
	100 12 15 10					ZERO	MAJOR
	100 7 15 1						ZERO

APPENDIX 7

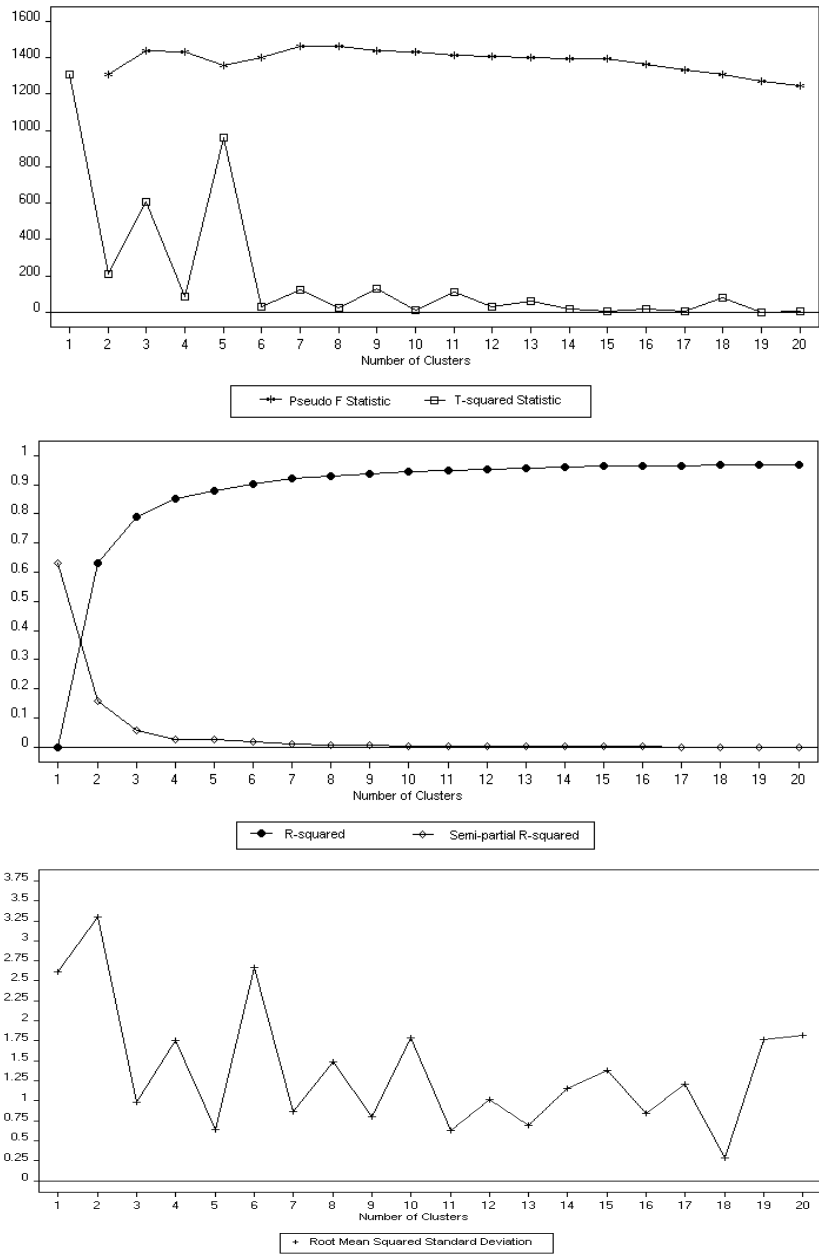
Run 4: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



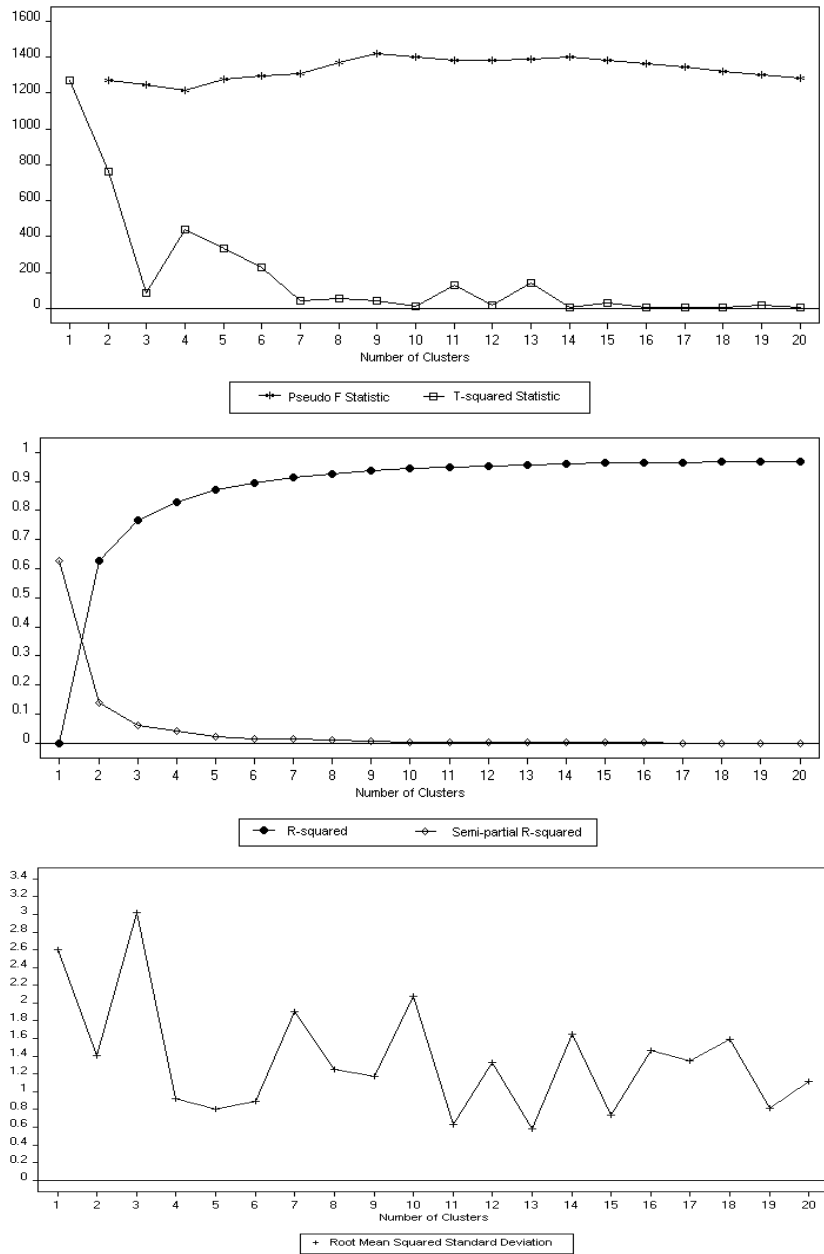
Run 5: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



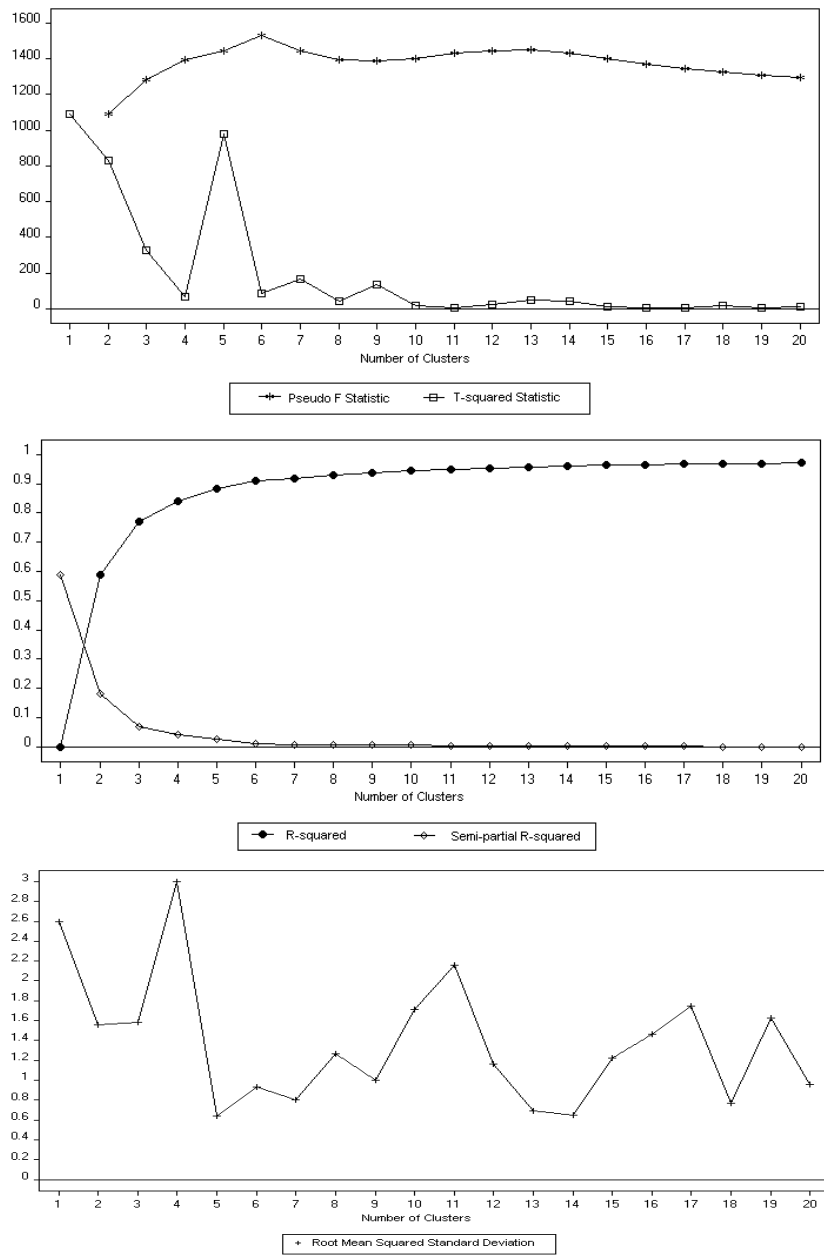
Run 6: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



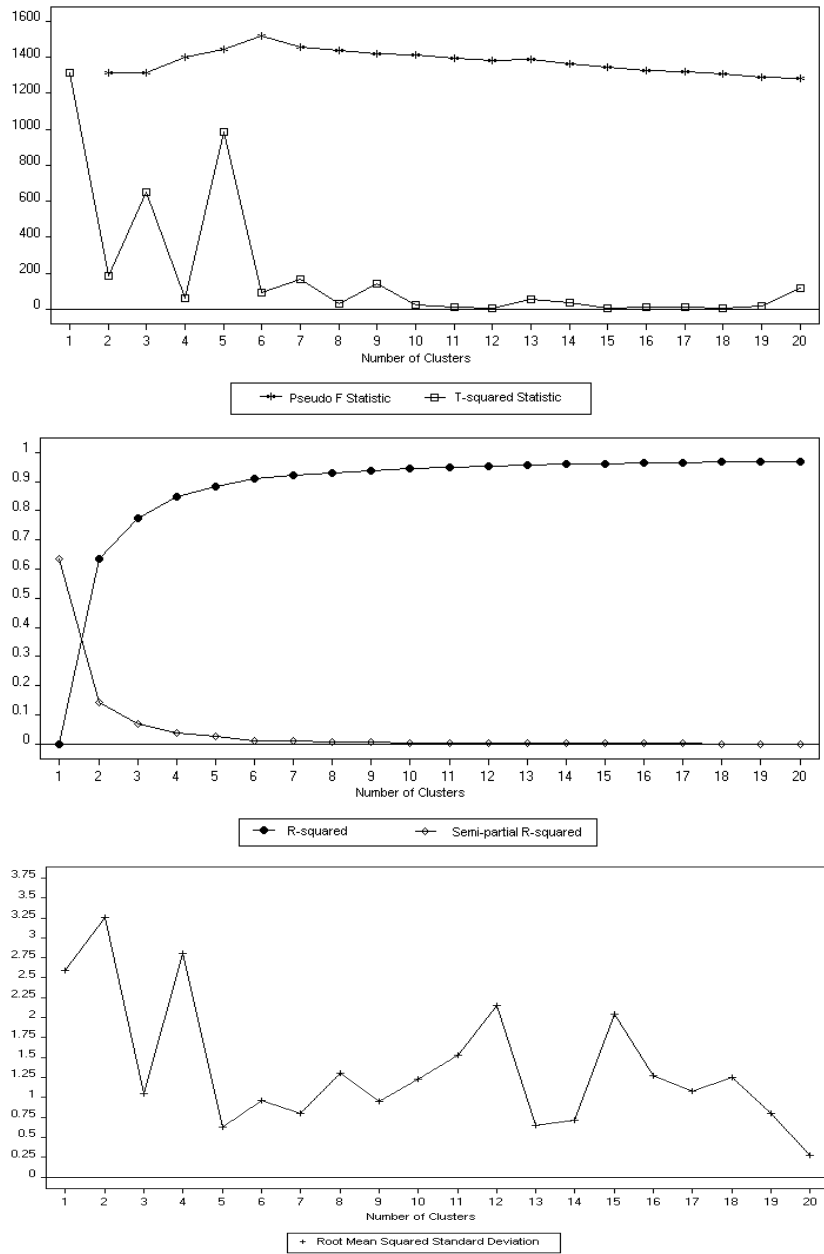
Run 7: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



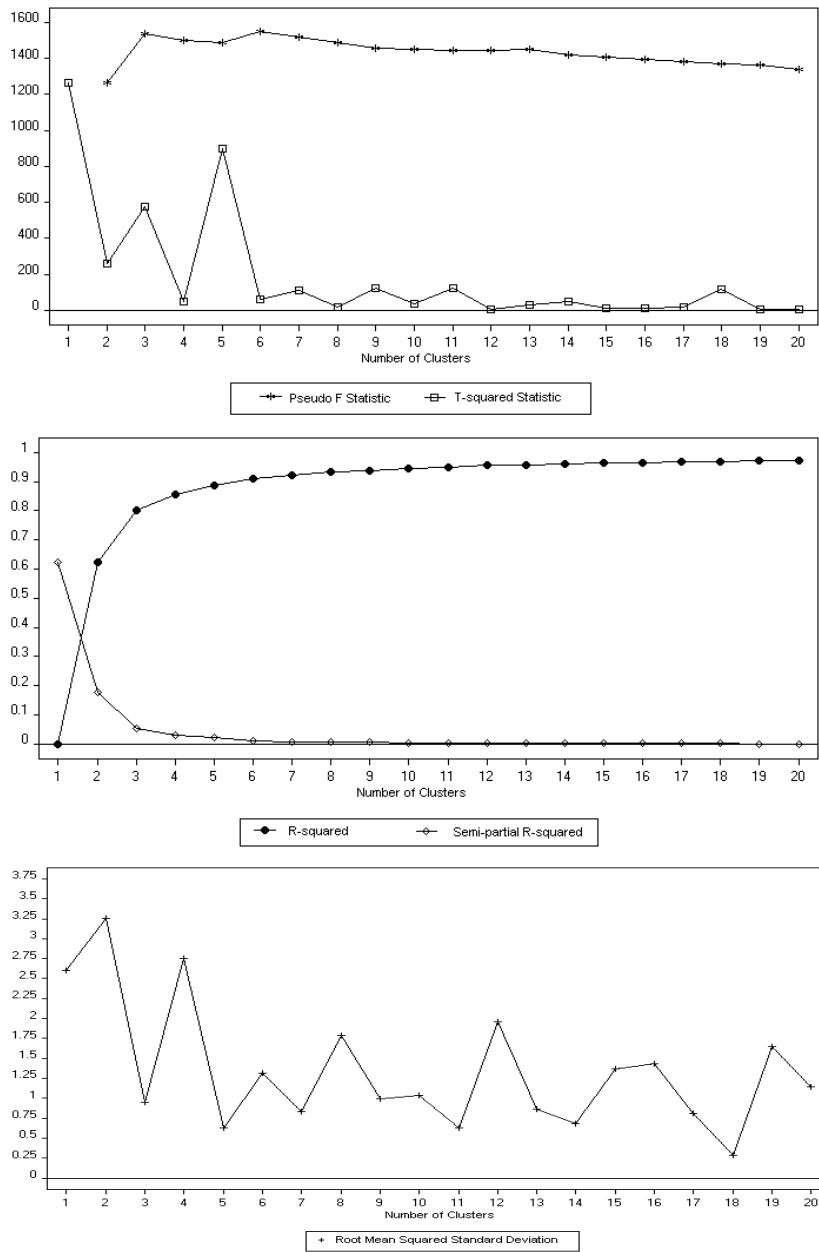
Run 8: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



Run 9: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



Run 10: Information criteria for defining the number of clusters, using SAM distance measures between 759 server sessions selected in the subset.



Equality table comparing run 1 with run 2.

Run 1: 4 clusters	Run 2: 5 clusters					TOT
	1	2	3	4	5	
1	279	5,937	61	0	16	6,293
2	7,536	6	40	576	6	8,164
3	123,660	254	8,391	2,393	2,213	136,911
4	329	0	0	15	0	344
TOT	131,804	6,197	8,492	2,984	2,235	151,712

Equality table comparing run 1 with run 3.

Run 1: 4 clusters	Run 3: 5 clusters					TOT
	1	2	3	4	5	
1	5,980	270	38	5	0	6,293
2	3	4,871	30	3,070	190	8,164
3	266	127,952	8,327	14	352	136,911
4	0	340	0	0	4	344
TOT	6,249	133,433	8,395	3,089	546	151,712

Equality table comparing run 1 with run 4.

Run 1: 4 clusters	Run 4: 3 clusters			TOT
	1	2	3	
1	240	6,049	4	6,293
2	33	7,767	364	8,164
3	8,538	127,418	955	136,911
4	0	338	6	344
TOT	8,811	141,572	1,329	151,712

Equality table comparing run 1 with run 5.

Run 1: 4 clusters	Run 5: 5 clusters					TOT
	1	2	3	4	5	
1	189	5,862	121	121	0	6,293
2	3	4,922	3,081	29	129	8,164
3	2,876	125,271	68	8,340	356	136,911
4	0	340	0	0	4	344
TOT	3,068	136,395	3,270	8,490	489	151,712

Equality table comparing run 1 with run 6.

Run 1: 4 clusters	Run 6: 3 clusters			TOT
	1	2	3	
1	5,951	40	302	6,293
2	9	31	8,124	8,164
3	283	8,352	128,276	136,911
4	0	0	344	344
TOT	6,243	8,423	137,046	151,712

Equality table comparing run 1 with run 7.

Run 1: 4 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	5,956	297	38	2	6,293
2	15	7,511	30	608	8,164
3	303	125,875	8,329	2,404	136,911
4	0	331	0	13	344
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 1 with run 8.

Run 1: 4 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	5,963	287	36	7	0	6,293
2	15	90	30	7,922	107	8,164
3	335	127,423	8,336	85	732	136,911
4	0	309	0	25	10	344
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 1 with run 9.

Run 1: 4 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	206	6,087	0	6,293
2	52	7,853	259	8,164
3	3,387	132,972	552	136,911
4	0	341	3	344
TOT	3,645	147,253	814	151,712

Equality table comparing run 1 with run 10.

Run 1: 4 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	300	5,991	2	6,293
2	44	8,082	38	8,164
3	8,597	127,894	420	136,911
4	0	342	2	344
TOT	8,941	142,309	462	151,712

Equality table comparing run 2 with run 3.

Run 2: 5 clusters	Run 3: 5 clusters					TOT
	1	2	3	4	5	
1	259	127,984	26	3,080	455	131,804
2	5,939	253	3	2	0	6,197
3	37	89	8,366	0	0	8,492
4	0	2,896	0	1	87	2,984
5	14	2,211	0	6	4	2,235
TOT	6,249	133,433	8,395	3,089	546	151,712

Equality table comparing run 2 with run 4.

Run 2: 5 clusters	Run 4: 3 clusters			TOT
	1	2	3	
1	18	130,661	1,125	131,804
2	409	5,787	1	6,197
3	8,383	104	5	8,492
4	0	2,786	198	2,984
5	1	2,234	0	2,235
TOT	8,811	141,572	1,329	151,712

Equality table comparing run 2 with run 5.

Run 2: 5 clusters	Run 5: 5 clusters					TOT
	1	2	3	4	5	
1	2,845	125,552	3,123	33	251	131,804
2	169	5,820	119	89	0	6,197
3	33	78	13	8,368	0	8,492
4	0	2,739	8	0	237	2,984
5	21	2,206	7	0	1	2,235
TOT	3,068	136,395	3,270	8,490	489	151,712

Equality table comparing run 2 with run 6.

Run 2: 5 clusters	Run 6: 3 clusters			TOT
	1	2	3	
1	259	44	131,501	131,804
2	5,929	7	261	6,197
3	43	8,372	77	8,492
4	1	0	2,983	2,984
5	11	0	2,224	2,235
TOT	6,243	8,423	137,046	151,712

Equality table comparing run 2 with run 7.

Run 2: 5 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	283	131,390	30	101	131,804
2	5,929	264	4	0	6,197
3	43	84	8,363	2	8,492
4	0	61	0	2,923	2,984
5	19	2,215	0	1	2,235
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 2 with run 8.

Run 2: 5 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	314	123,296	27	7,482	685	131,804
2	5,942	248	3	4	0	6,197
3	48	71	8,372	1	0	8,492
4	0	2,282	0	546	156	2,984
5	9	2,212	0	6	8	2,235
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 2 with run 9.

Run 2: 5 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	2,634	128,538	632	131,804
2	172	6,025	0	6,197
3	801	7,690	1	8,492
4	0	2,807	177	2,984
5	38	2,193	4	2,235
TOT	3,645	147,253	814	151,712

Equality table comparing run 2 with run 10.

Run 2: 5 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	273	131,115	416	131,804
2	256	5,939	2	6,197
3	8,405	86	1	8,492
4	0	2,943	41	2,984
5	7	2,226	2	2,235
TOT	8,941	142,309	462	151,712

Equality table comparing run 3 with run 4.

Run 3: 5 clusters	Run 4: 3 clusters			TOT
	1	2	3	
1	211	6,037	1	6,249
2	236	131,978	1,219	133,433
3	8,362	32	1	8,395
4	2	3,085	2	3,089
5	0	440	106	546
TOT	8,811	141,572	1,329	151,712

Equality table comparing run 3 with run 5.

Run 3: 5 clusters	Run 5: 5 clusters					TOT
	1	2	3	4	5	
1	206	5,832	120	91	0	6,249
2	2,858	129,995	88	36	456	133,433
3	2	27	3	8,363	0	8,395
4	2	30	3,057	0	0	3,089
5	0	511	2	0	33	546
TOT	3,068	136,395	3,270	8,490	489	151,712

Equality table comparing run 3 with run 6.

Run 3: 5 clusters	Run 6: 3 clusters			TOT
	1	2	3	
1	5,968	8	273	6,249
2	270	49	133,114	133,433
3	5	8,366	24	8,395
4	0	0	3,089	3,089
5	0	0	546	546
TOT	6,243	8,423	137,046	151,712

Equality table comparing run 3 with run 7.

Run 3: 5 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	5,958	286	5	0	6,249
2	310	130,178	26	2,919	133,433
3	3	26	8,366	0	8,395
4	3	3,068	0	18	3,089
5	0	456	0	90	546
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 3 with run 8.

Run 3: 5 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	5,978	263	5	3	0	6,249
2	330	127,527	30	4,799	747	133,433
3	3	25	8,367	0	0	8,395
4	2	19	0	3,067	1	3,089
5	0	275	0	170	101	546
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 3 with run 9.

Run 3: 5 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	187	6,062	0	6,249
2	2,672	130,050	711	133,433
3	756	7,639	0	8,395
4	30	3,058	1	3,089
5	0	444	102	546
TOT	3,645	147,253	814	151,712

Equality table comparing run 3 with run 10.

Run 3: 5 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	272	5,976	1	6,249
2	295	132,710	428	133,433
3	8,370	25	0	8,395
4	4	3,085	0	3,089
5	0	513	33	546
TOT	8,941	142,309	462	151,712

Equality table comparing run 4 with run 5.

Run 4: 3 clusters	Run 5: 5 clusters					TOT
	1	2	3	4	5	
1	117	233	11	8,450	0	8,811
2	2,947	134,913	3,257	40	415	141,572
3	4	1,312	2	0	74	1,329
TOT	3,068	136,395	3,270	8,490	489	151,712

Equality table comparing run 4 with run 6.

Run 4: 3 clusters	Run 6: 3 clusters			TOT
	1	2	3	
1	219	8,375	217	8,811
2	6,024	48	135,500	141,572
3	0	0	1,329	1,329
TOT	6,243	8,423	137,046	151,712

Equality table comparing run 4 with run 7.

Run 4: 3 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	213	235	8,363	0	8,811
2	6,061	132,656	34	2,821	141,572
3	0	1,123	0	206	1,329
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 4 with run 8.

Run 4: 3 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	221	222	8,364	3	1	8,811
2	6,090	127,020	38	7,686	738	141,572
3	2	867	0	350	110	1,329
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 4 with run 9.

Run 4: 3 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	873	7,938	0	8,811
2	2,771	138,559	242	141,572
3	1	756	572	1,329
TOT	3,645	147,253	814	151,712

Equality table comparing run 4 with run 10.

Run 4: 3 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	8,566	244	1	8,811
2	375	140,824	373	141,572
3	0	1,241	88	1,329
TOT	8,941	142,309	462	151,712

Equality table comparing run 5 with run 6.

Run 5: 5 clusters	Run 6: 3 clusters			TOT
	1	2	3	
1	206	5	2,857	3,068
2	5,831	39	130,525	136,395
3	116	3	3,151	3,270
4	90	8,375	25	8,490
5	0	1	488	489
TOT	6,243	8,423	137,046	151,712

Equality table comparing run 5 with run 7.

Run 5: 5 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	189	2,874	2	3	3,068
2	5,868	127,740	30	2,757	136,395
3	123	3,116	3	28	3,270
4	94	34	8,362	0	8,490
5	0	250	0	239	489
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 5 with run 8.

Run 5: 5 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	215	2,848	2	3	0	3,068
2	5,875	124,840	31	4,846	803	136,395
3	128	71	3	3,068	0	3,270
4	95	29	8,366	0	0	8,490
5	0	321	0	128	46	489
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 5 with run 9.

Run 5: 5 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	224	2,844	0	3,068
2	2,613	133,028	754	136,395
3	47	3,223	0	3,270
4	761	7,729	0	8,490
5	0	429	60	489
TOT	3,645	147,253	814	151,712

Equality table comparing run 5 with run 10.

Run 5: 5 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	162	2,906	0	3,068
2	293	135,654	448	136,395
3	32	3,238	0	3,270
4	8,454	36	0	8,490
5	0	475	14	489
TOT	8,941	142,309	462	151,712

Equality table comparing run 6 with run 7.

Run 6: 3 clusters	Run 7: 4 clusters				TOT
	1	2	3	4	
1	5,976	261	5	1	6,243
2	15	39	8,369	0	8,423
3	283	133,714	23	3,026	137,046
TOT	6,274	134,014	8,397	3,027	151,712

Equality table comparing run 6 with run 8.

Run 6: 5 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	5,992	244	5	2	0	6,243
2	14	38	8,370	1	0	8,423
3	307	127,827	27	8,036	849	137,046
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 6 with run 9.

Run 6: 3 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	201	6,042	0	6,243
2	764	7,659	0	8,423
3	2,680	133,552	814	137,046
TOT	3,645	147,253	814	151,712

Equality table comparing run 6 with run 10.

Run 6: 3 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	277	5,966	0	6,243
2	8,377	46	0	8,423
3	287	136,297	462	137,046
TOT	8,941	142,309	462	151,712

Equality table comparing run 7 with run 8.

Run 7: 4 clusters	Run 8: 5 clusters					TOT
	1	2	3	4	5	
1	5,985	280	3	6	0	6,274
2	323	125,510	36	7,454	691	134,014
3	3	31	8,363	0	0	8,397
4	2	2,288	0	579	158	3,027
TOT	6,313	128,109	8,402	8,039	849	151,712

Equality table comparing run 7 with run 9.

Run 7: 4 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	194	6,080	0	6,274
2	2,689	130,690	635	134,014
3	762	7,635	0	8,397
4	0	2,848	179	3,027
TOT	3,645	147,253	814	151,712

Equality table comparing run 7 with run 10.

Run 7: 4 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	260	6,013	1	6,274
2	315	133,279	420	134,014
3	8,366	31	0	8,397
4	0	2,986	41	3,027
TOT	8,941	142,309	462	151,712

Equality table comparing run 8 with run 9.

Run 8: 5 clusters	Run 9: 3 clusters			TOT
	1	2	3	
1	215	6,098	0	6,313
2	2,636	125,020	453	128,109
3	762	7,640	0	8,402
4	32	7,773	234	8,039
5	0	722	127	849
TOT	3,645	147,253	814	151,712

Equality table comparing run 8 with run 10.

Run 8: 5 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	269	6,044	0	6,313
2	292	127,473	344	128,109
3	8,374	28	0	8,402
4	6	8,004	29	8,039
5	0	760	89	849
TOT	8,941	142,309	462	151,712

Equality table comparing run 9 with run 10.

Run 9: 3 clusters	Run 10: 3 clusters			TOT
	1	2	3	
1	925	2,717	3	3,645
2	8,016	138,840	397	147,253
3	0	752	62	814
TOT	8,941	142,309	462	151,712