Theory of the topical coverage of multiple databases

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

leo.egghe@uhasselt.be

ABSTRACT

We present a model that describes which fraction of the literature on a certain topic we will find when we use *n (n=1,2,…)* databases. It is a generalization of the theory of discovering usability problems.

We prove that, in all practical cases, this fraction is a concave function of *n*, the number of used databases, hereby explaining some graphs that exist in the literature.

We also study limiting features of this fraction for *n* very high and we characterize the case that we find all literature on a certain topic for *n* high enough.

---

[1] Permanent address

Key words and phrases: coverage, multiple databases, usability problems
Acknowledgement: The author is grateful to one anonymous referee for the advise to consider expected values as interpretation of the fractions a_i.

Introduction

The coverage of databases of the literature on a certain topic is an important issue in information retrieval. Only in a few cases one database will cover all the literature on a topic. The smaller the coverage fraction is, the more databases we will have to use in order to cover a certain percentage of the complete literature that exists on this topic. Hereby we can use – besides field-dedicated databases – general databases such as e.g. the Web of Science (WoS).

One database can cover a fraction a (or $100a\%$) of the existing literature on a certain topic. A second database will cover another fraction of the existing literature on a certain topic, but here, in this introduction, we assume that also the second database covers a fraction a of the existing literature on a certain topic. However, using both databases will not yield a fraction $2a$ of the sought literature since several documents will be common to both databases.

What fraction of the existing literature on a certain topic will be found after the use of n $(=1,2,3,\ldots)$ databases? Suppose here that all databases cover the same fraction a of the literature on the topic (this is not very realistic but it is the subject of this paper to extend the theory to different fractions (from the second section onward)). The argument is as follows. The first database yields an expected[2] fraction a of the existing literature on a certain topic, hence it does not yield the complementary fraction $1-a$ of the literature. Using a second database will not yield a fraction $(1-a)^2$ of the literature. Indeed, both databases do not yield a fraction $1-a$ of the sought literature, hence in both databases we have that $1-a$ is the probability to miss a document on the topic. Due to independence we have that after the use of two databases we missed a fraction $(1-a)^2$ of the sought literature. This argument can be repeated to $3,4,\ldots,n$ databases yielding that after using n databases we missed a fraction $(1-a)^n$ of the sought literature and hence we have found a fraction

$$1-(1-a)^n \qquad (1)$$

of the sought literature.

[2] From now on we will delete the adjective "expected" and work with these numbers as probabilities – see also the argument in the next section.

This is similar to the following problem: how many users of a certain service (e.g. a library) must be interviewed to find a certain fraction of the usability problems of that service. Similar to the above we can assume that each user can inform us about a fraction *a* of usability problems. The same argument as above yields a fraction (1) of usability problems after the interviewing of *n* users (see e.g. Nielsen and Landauer (1993) or http://www.useit.com/alertbox/20000319.html (retrieved on January 5, 2012)).

Requiring (1) to be as high as we wish (e.g. 0.9 or 90%) yields the needed number *n* of databases to be used (or users to be interviewed):

$$1-(1-a)^n = 0.9$$

or

$$(1-a)^n = 0.1$$

hence

$$n = \frac{\log(0.1)}{\log(1-a)} \qquad (2)$$

, where any logarithm can be used.

The function (1) is a concavely increasing function of *n* and its limit (for *n* going to ∞) is 1 as is readily seen. This is a partial explanation of graphs as in Hood and Wilson (2001), see Fig. 1 where several topics (indicated in the Figure) are retrieved in 1,2,3,… databases and where the graphs indicate the percentage (fraction) of records retrieved after the use of *n=1,2,3,…* databases.

This partial explanation of Fig.1 is important in information retrieval. It indicates how the recall increases with the number *n* of used databases. As formula (1) and Fig.1 indicate, to reach a recall close to 1 requires the use of a high number of databases and shows the inefficiency of such searches.

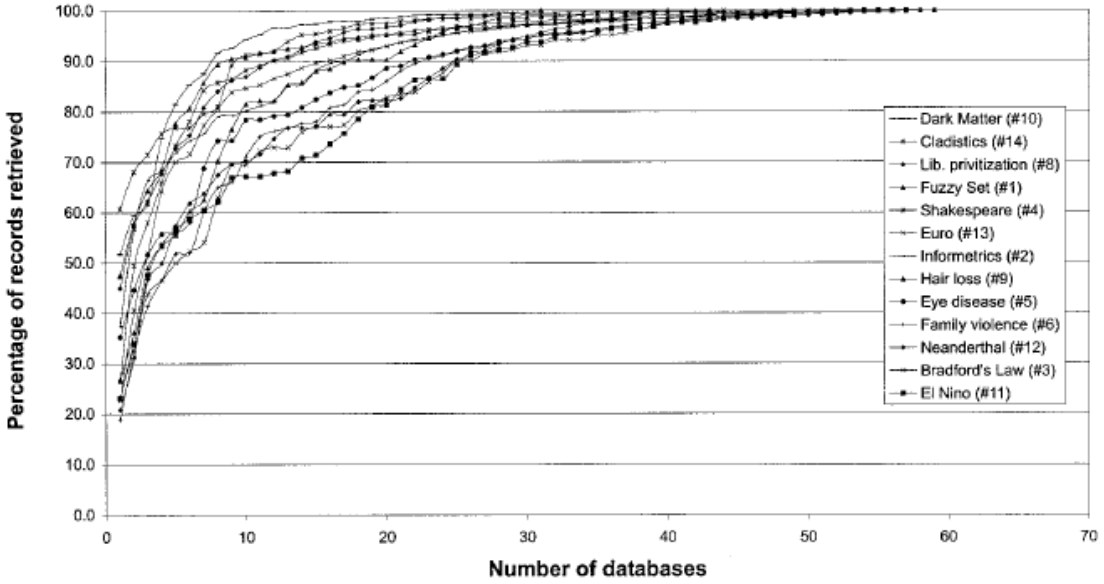

Fig.1. Distribution of records from 13 search statements shown as percentage of records retrieved over the number of databases searched. (Hood and Wilson (2001)).

However the explanation above is partial since we assumed that all databases yield a fixed fraction *a* of the sought literature. Similarly, it is not realistic to assume that all users of a certain service yield a fixed fraction of usability problems. Hence both applications need variable fractions per database or per user. This is the topic of this paper. We will, henceforth, use the information retrieval terminology but the application to the detection of usability problems is similar.

In the next section we present the general formula for the fraction of sought literature after the use of *n* databases. We prove under which conditions we have a concavely increasing curve (in function of *n*) and we show that in case of Fig.1 these conditions are satisfied, hence yielding a complete explanation of these graphs.

In the third section we study limiting problems of this formula for the fraction of sought literature. We give necessary and sufficient conditions for this fomrula (function of *n*) to go to 1 for *n* going to ∞. Only in this case we can be as close as we want to retrieving 100% of the sought literature, if we use enough databases. An example where this is the case and an example where this is not the case is given.

The paper closes with some final remarks and suggestions for further research.

The fraction of sought literature after the use of *n* databases

Let us have a non-specified number of databases that we can use for retrieving documents on a certain topic. The order in which we use these databases is important in practice but is not specified at this moment. We come back to this issue later on in this section.

We denote by $a_i \ \ (0 < a_i < 1)$ the expected[3] fraction of sought documents in database *i* $(i = 1, 2, 3, ...)$. Here we assume that when we use database *i*, we can retrieve the complete fraction a_i of sought documents (otherwise the value a_i is reduced which is not important at this stage). In analogy with the argument yielding formula (1) we have now that, using only database 1, there is a fraction $1 - a_1$ of sought documents that is not retrieved. After using the first two databases we have a fraction $(1 - a_1)(1 - a_2)$ of sought documents that is not retrieved (due to independence). After using the first *n* databases we hence have a fraction $(1 - a_1)(1 - a_2)...(1 - a_n)$ of sought documents that is not retrieved. Consequently, after using the first *n* databases we hence have a fraction

$$f(a_1, ..., a_n) = 1 - \prod_{i=1}^{n} (1 - a_i) \qquad (3)$$

of sought documents that is retrieved, where $\prod_{i=1}^{n} (1 - a_i)$ denotes the product $(1 - a_1)(1 - a_2)...(1 - a_n)$.

As one referee points out, the above argument is not completely correct (as is the one that proves (1)) and can be made more correct by considering expected values. This can be done as follows. Let *S* be the set of the literature on a topic and *A(j)* be the subset covered by database *j*, $j = 1, 2, ...$. Denote by *B(j)* the complement of *A(j).* For each element $w \in S$ and each $n = 1, 2, ...$, define the indicator function $I(w; n) = 1$ if *w* is included in at least one of the

[3] As in the previous section, we will henceforth delete the adjective "expected" and work with these numbers as probabilities – see also the argument below.

first *n* databases (i.e. if $w \in \bigcup_{j=1}^{n} A(j)$) and $I(w;n)=0$ otherwise (i.e. if $w \in \bigcap_{j=1}^{n} B(j)$). The function *I* is a random variable of which we want to know the expected proportion in the first *n* databases (as in (3)).

This is

$$
\begin{aligned}
&P\left[I(w;n)=1\right] \\
&=1-P\left[I(w;n)=0\right] \\
&=1-P\left[w \in B(1), w \in B(2), ..., w \in B(n)\right] \\
&=1-\prod_{i=1}^{n} P\left(w \in B(i)\right) \\
&=1-\prod_{i=1}^{n}\left(1-a_i\right)
\end{aligned}
$$

by the assumed independence of compilation of different databases. Hence we refind (3).

Note: One referee remarks that the above model assumes that "every item in the literature has the same chance of being included in a particular database as any other item". This is not assumed in the above model. It is clear that some sought documents have a higher chance to be included in a database than others. But that does not prevent us of assuming that a_i is the (expected) fraction of sought documents in database *i*. In fact we simply extend the well-established model (1) of Nielsen and Landauer (1993).

Function (3) generalizes function (1) and it will turn out that it does not always have the property that it increases concavely nor that is goes to 1 for *n* going to ∞. The latter problem will be studied in the next section on the limiting properties of $f(a_1,...,a_n)$ for *n* going to ∞; the former property will be studied here. We have the following Proposition.

Proposition 1:

The function $f(a_1,...,a_n)$ is always increasing and is concave if and only if, for every $n=2,3,....$

$$a_n - a_{n-1} - a_{n-1}a_n < 0 \qquad (4)$$

Proof:

The function $f(a_1,...,a_n)$ clearly increases (strictly) since $0 < a_n < 1$ for all $n = 1,2,...$. It is concave (in *n*, with fixed a_i-values) if and only if, at each $n = 2,3,...$, we have that

$$f(a_1,...,a_n) - f(a_1,...,a_{n-1}) < f(a_1,...,a_{n-1}) - f(a_1,...,a_{n-2}) \qquad (5)$$

where we define $f(a_1,...,a_{n-2}) = 0$ for $n = 2$ (the starting point of *f* when zero databases are used). But (5) boils down to, for $n = 3,\ldots$

$$1 - \prod_{j=1}^{n}(1-a_j) - 1 + \prod_{j=1}^{n-1}(1-a_j) < 1 - \prod_{j=1}^{n-1}(1-a_j) - 1 + \prod_{j=1}^{n-2}(1-a_j)$$

or

$$\prod_{j=1}^{n-1}(1-a_j)a_n < \prod_{j=1}^{n-2}(1-a_j)a_{n-1}$$

or

$$(1-a_{n-1})a_n < a_{n-1}$$

from which (4) follows. This condition is also found if $n = 2$ (using that $f(a_1,...,a_{n-2}) = 0$). □

Cases in which (4) are valid are many.

(i) Requirement (4) is valid if the sequence $(a_n)_{n=1,2,...}$ is decreasing. This is the case in Fig. 1: per search, databases are used in decreasing order of their fraction of sought documents (see Hood and Wilson (2001), p.1246, search procedure (3)). So Proposition 1 gives a full explanation of the shapes of the curves in Fig.1 – the small deviations of the concavity in the curves are due to

the fact that an information retrieval process is a sample in the sought documents.

(ii) If the a_i-values are large (i.e. close to 1); then for every $n = 2,3,\ldots, a_n \approx a_{n-1}$ and $a_{n-1}a_n \approx 1$ making (4) valid. Here any order in which the databases are used yields a concave function $f(a_1,\ldots,a_n)$. This case will occur often in practice for the following reasons. When trying to retrieve documents on a certain topic one uses only databases in the field of this topic or general databases (such as the WoS). In both cases the fraction of the sought documents in these databases is high. Make distinction with the fraction of the documents in the database which are sought, which is usually low but these are not the a_i-values: they are the fraction of the sought documents that are in database *i*. This is common sense: a mathematical topic will not be searched in e.g. a medical database and vice-versa.

(iii) There are even cases where the sequence $(a_n)_{n=1,2,\ldots}$ is increasing and where (4) is valid.

Example: take $a_n = \dfrac{n}{n+1}$ for all $n = 1,2,\ldots$. Then the sequence $(a_n)_{n=1,2,\ldots}$ increases strictly but condition (4) is valid:

$$a_n - a_{n-1} - a_{n-1}a_n = \frac{n}{n+1} - \frac{n-1}{n} - \frac{n-1}{n}\frac{n}{n+1} = \frac{1+n-n^2}{n(n+1)} < 0$$

since $n \geq 2$ in condition (4). Here

$$\begin{aligned} f(a_1,\ldots,a_n) &= 1 - \prod_{j=1}^{n}(1-a_j) \\ &= 1 - \left(1-\frac{1}{2}\right)\left(1-\frac{2}{3}\right)\ldots\left(1-\frac{n}{n+1}\right) \\ &= 1 - \frac{1}{2}.\frac{1}{3}\ldots\frac{1}{n+1} \end{aligned}$$

a concave function of *n*. Indeed

$$
\begin{aligned}
& f\left(a_1,...,a_{n+1}\right)-f\left(a_1,...,a_n\right) \\
&=1-\frac{1}{2}.\frac{1}{3}...\frac{1}{n+1}-1+\frac{1}{2}.\frac{1}{3}...\frac{1}{n} \\
&=\frac{1}{2}.\frac{1}{3}...\frac{1}{n}\left(1-\frac{1}{n+1}\right) \\
&=\frac{1}{2}.\frac{1}{3}...\frac{1}{n+1}
\end{aligned}
$$

which is decreasing in *n* and hence $f\left(a_1,...,a_n\right)$ is concave in *n*.

(iv) However, not every increasing sequence $\left(a_n\right)_{n=1,2,...}$ yields a concave $f\left(a_1,...,a_n\right)$. Indeed, take $n=3,\ a_1=0.1,\ a_2=0.2,\ a_3=0.3$. Then condition (4) is not satisfied:

$$a_3-a_2-a_2a_3=0.3-0.2-(0.3)(0.2)>0$$

and indeed:

$$
\begin{aligned}
& f\left(a_1\right)=0.1 \\
& f\left(a_1,a_2\right)=1-(1-0.1)(1-0.2)=0.28 \\
& f\left(a_1,a_2,a_3\right)=1-(1-0.1)(1-0.2)(1-0.3)=0.496
\end{aligned}
$$

Hence *f* is not concave since 0.496 - 0.28 = 0.216 > 0.28 – 0.1 = 0.18. In fact, *f* is even convex in this case.

Limiting properties of the function

$f\left(a_1,...,a_n\right)=1-\prod_{i=1}^{n}\left(1-a_i\right)$ for *n=1,2,…*

This is an important issue. More specifically we are interested in when

$$\lim_{n\to\infty} f\left(a_1,...,a_n\right)=1-\prod_{i=1}^{\infty}\left(1-a_i\right)=1 \tag{6}$$

, in other words when

$$\prod_{i=1}^{\infty}\left(1-a_i\right)=0 \tag{7}$$

If (6) is the case we are in a situation that, when using sufficient databases, we can reach (almost) complete coverage of the sought documents. Note that this is the case for all searches in Fig. 1 of Hood and Wilson (2001). We will, however, see that (6) (or (7)) is not always valid. In case (6) (or(7)) is not always valid, we have that

$$\lim_{n\to\infty} f\left(a_1,...,a_n\right)<1 \tag{8}$$

and in this case, no matter how many databases we are searching, we will never come close to complete coverage of the sought documents. In the sequel we will give an example of both cases: one where we have (6) and one where we have (8). Note that in the special case (1) we always have (6) which shows that our extension of f to formula (3) has its merits.

First we will give some definitions on convergent or divergent products. They can be found in Apostol (1974) (p. 206-209). We limit our definitions to our case studied here.

Definition 1:

Denote by p_n the product

$$p_n=\prod_{i=1}^{n}\left(1-a_i\right) \tag{9}$$

Then we say that this product converges if there exists a number $p\neq 0$ such that $\lim_{n\to\infty} p_n = p$.

The number *p* is then denoted

$$p = \prod_{i=1}^{\infty} (1 - a_i) \tag{10}$$

If $p = \lim_{n \to \infty} p_n = 0$ we say that the product diverges to 0 (hence the case (7) or (6), the most interesting case since we are able to retrieve most documents on the topic by taking *n* high enough).

We can give a characterization of convergent or divergent products of the form (10) by quoting a Theorem in Apostol (1974), p. 209.

Theorem 1:

Since all a_i satisfy $0 < a_i < 1$ we have that the product $\prod_{i=1}^{\infty} (1 - a_i)$ converges if and only if the series $\sum_{i=1}^{\infty} a_i$ converges. This represents the case (8), hence where we are not able to come close to a complete coverage of the sought documents (no matter how many databases that are used). Complete coverage (as in (6)) is hence possible using the next Theorem which follows immediately from Theorem 1.

Theorem 2:

We have that the product $\prod_{i=1}^{\infty} (1 - a_i)$ diverges (hence where (7) or (6) is valid) if and only if the series $\sum_{i=1}^{\infty} a_i$ diverges.

A divergent series $\sum_{i=1}^{\infty} a_i$ means in practice that the fractions a_i must be "large enough" so that each database has "enough" coverage of the sought documents in order to make a complete coverage (6) possible. A convergent series $\sum_{i=1}^{\infty} a_i$ means in practice that the fractions a_i are too small, preventing complete coverage. We give an example of each case.

Example 1:

Let $a_i = \frac{1}{i+1}$, $i = 1, 2, \ldots$. Hence $\sum_{i=1}^{n} a_i$ diverges and, according to Theorem 2, $\prod_{i=1}^{\infty}(1-a_i) = 0$ (i.e. diverges), so (6) and (7) are valid and complete coverage of the sought documents is (in the limit) possible. We can verify this directly. We have, for every $n = 1, 2, \ldots$

$$\prod_{i=1}^{n}(1-a_i) = \prod_{i=1}^{n}\left(1-\frac{1}{i+1}\right) = \prod_{i=1}^{n}\left(\frac{i}{i+1}\right) = \frac{1}{n+1}$$

So

$$\prod_{i=1}^{\infty}(1-a_i) = 0$$

and hence

$$\lim_{n\to\infty} f(a_1, \ldots, a_n) = \lim_{n\to\infty}\left(1-\frac{1}{n+1}\right) = 1$$

Since $f(a_1, \ldots, a_n) = \frac{n}{n+1}$ we can illustrate "how fast" we approximate the 100% coverage.

Take e.g. $n = 10$ databases, then we can cover $\frac{10}{11} > 0.9$, hence more than 90% of the sought deocuments.

Example2:

Let $a_i = \frac{1}{(i+1)^2}$, $i = 1, 2, \ldots$. Now we have that $\sum_{i=1}^{\infty} a_i$ is convergent and hence the product $\prod_{i=1}^{\infty}(1-a_i)$ is convergent (i.e is $\neq 0$). This means that (8) is valid and that we cannot

approximate complete coverage of the sought documents. We can here, concretely, calculate what fraction of the sought documents can be covered. We have, for every $n=1,2,...$

$$\prod_{i=1}^{n}(1-a_i)=\prod_{i=1}^{n}\left(1-\frac{1}{(i+1)^2}\right)$$
$$=\prod_{i=1}^{n}\left(\frac{i^2+2i}{(i+1)^2}\right)$$
$$=\frac{3}{4}.\frac{8}{9}.\frac{15}{16}.\frac{24}{25}.\frac{35}{36}.\frac{48}{49}...\frac{n^2+2n}{(n+1)^2}$$
$$=\frac{3}{2}.\frac{4}{3}.\frac{5}{4}.\frac{6}{5}.\frac{7}{6}.\frac{8}{7}...\frac{n+1}{n}.\frac{n^2+2n}{(n+1)^2}$$
$$=\frac{1}{2}\frac{n+2}{n+1}$$

and hence $\prod_{i=1}^{\infty}(1-a_i)=\frac{1}{2}$. This also implies that

$$\lim_{n\to\infty} f(a_1,...,a_n)=\frac{1}{2}$$

so that we certainly do not cover at least 50% of the sought documents (no matter how many databases we will use). This is due to the small coverage $a_i=\frac{1}{(i+1)^2}$ of the sought documents of each database $i=1,2,...$. This example shows the interest in the general model (3) above the limited model (1) where always

$$\lim_{n\to\infty} f(a_1,...,a_n)=\lim_{n\to\infty} f(a,...,a)=\lim_{n\to\infty}\left[1-(1-a)^n\right]=1$$

Note that in both examples $f(a_1,...,a_n)$ is concavely increasing since the sequence $(a_i)_{i=1,2,...}$ decreases and by Proposition 1.

Remark:

Since all a_i satisfy $0 < a_i < 1$ we have that convergence of $\sum_{i=1}^{\infty} a_i$ also means absolute convergence. This also means that the series $\sum_{i=1}^{\infty} a_i$ converges unconditionally, i.e. it converges in any order of the databases *i*. More exactly, let π denote any permutation of the natural numbers, i.e. a function whose domain is the natural numbers and whose range is the natural numbers and which is a bijection. Then convergence of $\sum_{i=1}^{\infty} a_i$ implies convergence of $\sum_{i=1}^{\infty} a_{\pi(i)}$ (see e.g. Apostol (1974), Theorem 8.32, p. 196) and hence, by Theorem 1, the product $\prod_{i=1}^{\infty}\left(1-a_{\pi(i)}\right)$ converges (and is equal to $\prod_{i=1}^{\infty}\left(1-a_i\right)$). Similarly, if $\sum_{i=1}^{\infty} a_i$ diverges, then $\sum_{i=1}^{\infty} a_{\pi(i)}$, diverges and hence, by Theorem 2, the product $\prod_{i=1}^{\infty}\left(1-a_{\pi(i)}\right)$ diverges (i.e. its value equals 0). This means that the coverage of sought documents, in the limit, is not influenced by the order in which we use the databases. Of course, for every finite $n = 1, 2, \ldots$, the values of $f\left(a_1, \ldots, a_n\right)$ are determined by the used order of the databases.

Note:

Considering an infinite number of databases is, of course, only a theoretical issue. Yet our results on complete/incomplete coverage (Theorems 1 and 2) yield insight in the finite case where there are *n* databases (*n*: natural number and high).

Conclusions and suggestions for further research

In this paper we studied the topical coverage of multiple databases. We showed that, when a_i ($0 < a_i < 1$) denotes the fraction of the sought documents (on a certain topic) of the i^{th} database, we cover a fraction

$$f(a_1,...,a_n) = 1 - \prod_{i=1}^{n}(1-a_i)$$

of the sought documents on a certain topic. We showed that in most practical cases, this function is concavely increasing in *n*.

We also showed that the limiting case (for *n* going to ∞) does not always yields a complete coverage $\left(\lim_{n\to\infty} f(a_1,...,a_n) = 1\right)$ of the sought documents. This is only so, if and only if the series $\sum_{i=1}^{\infty} a_i$ diverges.

Examples of complete coverage $\left(\lim_{n\to\infty} f(a_1,...,a_n) = 1\right)$ and incomplete coverage $\left(\lim_{n\to\infty} f(a_1,...,a_n) < 1\right)$ are given and we also showed that this is independent of the order in which we use the databases.

We underline that these generalizations of the simple model (1) (originating from the theory of finding a fraction of usability problems of a certain service) are also meaningful to this application. Indeed it is much likely that different interviewed users give a different number of usability problems and hence that model (1) is not applicable but that model (3) and its applications must be used. Further research on this application (which is outside the field of information retrieval) would be interesting.

We would also welcome other new field of application of this theory. In this context we could think of applications in the area of shopping in more than one supermarket or in the diffusion

of information in several documents (e.g. reviews, books, …) on a certain topic. Further applications could be seen by the reader.

References

T.M. Apostol (1974). Mathematical Analysis. Second Edition. Addison-Wesley Publishing Company, Reading, Massachusetts, USA.

W.W. Hood and C.S. Wilson (2001). The scatter of documents over databases in different subject domains: How many databases are needed? Journal of the American Society for Information Science and Technology 52(14), 1242-1254.

http://www.useit.com/alertbox/20000319.html (retrieved on January 5, 2012). Jakob Nielsen's Alertbox, March 19, 2000: Why you only need to test with 5 users.

J. Nielsen and T.K. Landauer (1993). A mathematical model of the finding of usability problems. Proceedings of ACM INTERCHI '93 Conference (Amsterdam, The Netherlands, 24-29 April 1993), 206-213.