INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

ON THE RELATIONSHIP BETWEEN THEORETICAL RETRIEVAL MODELS

Gerard SALTON

Department of Computer Science, Cornell University, Ithaca, NY 14853

Abstract

In recent years, a number of theoretical information retrieval models have been developed, including the vector space model, the Boolean logic model, and the probabilistic retrieval model. Each of these models exhibits advantages and disadvantages both from the conceptual and the practical viewpoints. some of the strengths and weaknesses of these models are described and various relationships between them are examined. None of the available models is fully adequate in representing the characteristics of retrieval systems and operations.

1. THE VECTOR SPACE MODEL

In the vector space model, the assumption is made that the stored records (documents) and the information requests are represented by sets of assigned keywords or index terms; [1-4]. This implies that queries and documents can be modelled by term vectors of the form

$$D_i = (a_{i_1}, a_{i_2}, \ldots a_{i_t})$$

$$Q_j = (q_{j_1}, q_{j_2}, \ldots q_{j_t})$$

where t is the number of distinct index terms available in the system, and a_{i_k} and q_{j_k} represent the values of term k in document D_i or query Q_j, respectively. Typically a_{i_k} or (q_{j_k}) might be set equal to 1 when term k appears in document D_i (or in query Q_j), and to 0 if the term is absent from the vector.

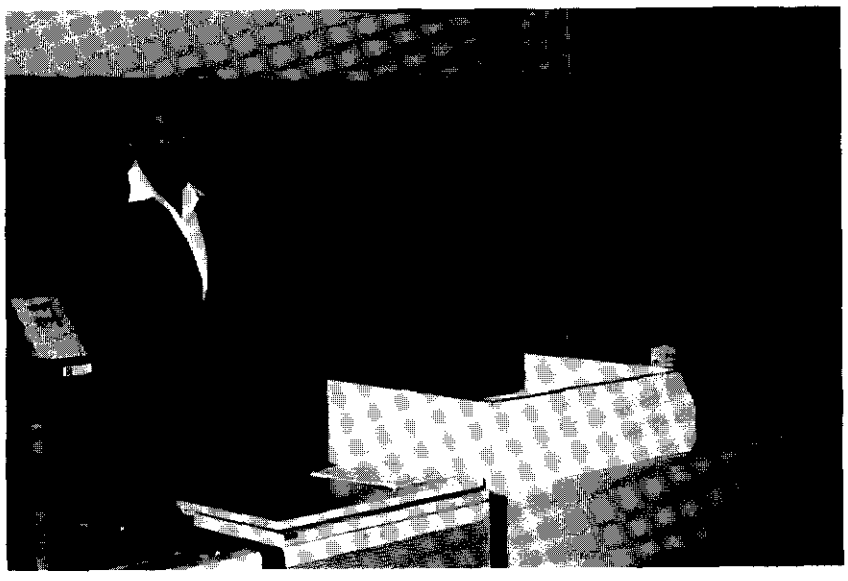
Alternatively, the vector coefficients could take on numerical values, the size of each coefficient depending on the importance of the term in the respective document or query.

The vector space model is known to be advantageous for a variety of reasons:

a)The similarity between term vectors is easily computed, based on the similarities between the term assignments to the corresponding vectors. Similarity coefficients can then be generated between queries and documents for information retrieval, or between different document vectors for document clustering purposes.

b)When the documents are arranged in decreasing order of query document similarity, a ranking of the documents becomes available, and documents can be retrieved in decreasing order of query-document similarity. A document ranking feature improves the interaction between users and system during the retrieval process.

c)In the vector system, the document vectors are easily modified either by addition of new terms and removal of old terms, or by suitable alterations in the term weights. This vector modification process is especially useful in query vector



G. Salton

manipulations of the kind used for relevance feedback, and in dynamic collection management. [4]

The main disadvantages of the vector processing system relate to the fact that the representation of document or query content by vectors of terms is relatively crude, and that the model does not account for many of the parameters needed to control the system. A trade-off thus exists between the flexible term weighting and vector modification possibilities on the one hand, and the system limitations imposed by the lack of term relationships, and the ad hoc choice of vector similarity and term weighting coefficients.
The use of term weights is known to furnish a high order of retrieval effectiveness. Three term weighting components are of main interest:

a)A factor that measures the importance of a term in a complete collection of items known as the inverse document frequency (idf) factor; the idf factor varies inversely with the number of documents n to which a term is assigned, and might be computed as log N:n when N is the collection size.
b)A factor that measures the importance of a term in an individual document known as the term frequency (tf) factor; the tf factor could be represented by the number of occurrences of a term in a document.
c)A normalization factor that reduces the possible term weights to a particular range.

A typical term weighting measure w_d for term d in document D might thus be specified as $(tf_d * \log \frac{N}{n_d}) / [\sum_{k=1}^{t} (tf_{d_k} * \log \frac{N}{n_{d_k}})^2]^{1/2}$ where the sum in the denominator ranges over all terms in a vector. A typical similarity measure between a query and document could be computed as the inner product between the corresponding normalized term vectors, that is

$$sim\ (D,Q) = \sum_{k=1}^{t} w_{d_k} \cdot w_{q_k} .$$

The foregoing query-document term weighting and comparison methods have been used in the design of iterative retrieval systems where the results of previous retrieval operations are used to generate improved queries through addition of terms obtained from the relevant items previously retrieved, and deletion of terms obtained from the corresponding nonrelevant items. A typical relevance feedback step produces a new query formulation Q' from an original query Q as follows

$$Q' = Q + \alpha \sum_{\text{Relevant}} D_i - \beta \sum_{\text{Non-Relevant}} D_j .$$

Here α and β are suitable constants and the summations range over a set of previously retrieved relevant and non-relevant items, respectively. The feedback equation, like the earlier term weighting formulas, is known to be effective in practice, but it is not obtained by any accepted formal considerations.

Even though the retrieval results obtainable with a simple vector processing model are hard to equal by using other apparently more sophisticated models, proposals have nevertheless been made for extending the basic vector model by incorporating term dependency information obtained, for example, from existing thesauruses, or volunteered by the system users. [5,6] Whether such modified vector models can produce reliable improvements in retrieval effectiveness remains to be seen.

2. THE BOOLEAN LOGIC MODEL

In the Boolean model, the documents are represented by term vectors as before, but the queries consist of Boolean statements, that is, terms interrelated by the operators and, or, and not. The immediate advantage is the structured query form which implicitly includes the specification of term synonyms in or-clauses, and of phrases in and-clauses. Thus in the query ((information or document) and retrieval), the terms "information" and "document" are treated as synonymous, and the user need is expressed by the phrases "information retrieval" and "document retrieval".

The disadvantages of the basic Boolean model are well-known: there are no provisions for term weighting in either documents or queries, and hence no ranked document output is obtained. Instead all items that match the query specification are retrieved in more or less random order. Because the output is not ranked in decreasing order of presumed goodness, it is much more difficult to perform sensible query reformulations following the initial search operations, and the complete search process starting with the initial structured query formulations is much more complex than the corresponding operations in the vector system.

The existing retrieval evaluations for the operations of the conventional Boolean *system demonstrate that the conventional Boolean logic is not well adapted to the* retrieval task. In particular, the conventional Boolean logic is too strict and produces counter intuitive results:

a) In response to and-queries (that is queries containing and-clauses), a document containing all query terms but one is treated just as badly as a document containing no query terms at all.
b) In response to or-queries, a document containing all the query terms is not treated any better than an item containing only one query term.

To respond to the limitations of the conventional Boolean system, many proposals have been made for improving the operations of the Boolean operations. The following extensions are of most interest in this connection:

a)The so-called quorum-level search system where only and-clauses are used in the formulations, and methods are provided for a suitable relaxing of the query formulations in order to retrieve a desired number of documents. [7]

b)The fuzzy set retrieval model which adds document term weights to the basic Boolean model, but preserves the normal Boolean formulations. When the term weights attached to the documents are limited to 0 and 1, the fuzzy set model reduces to the conventional Boolean model ([8-10],[28]).

c)An extended Boolean model based on the computation of generalized distance measures between documents and queries, where term weights are assignable to both documents and queries, and strictness parameters (p-values) are attached to the Boolean operators to control the strictness of interpretations of the operators. [11-13] A query in the extended system might be formulated as (((A,a) or ^{p_1} (B,b)) and ^{p_2} (C,c)) where a, b, and c are the weights of terms A, B, and C respectively, and p_1 and p_2 are the p-values. The extended Boolean system provides a general retrieval model which includes the vector processing and the conventional models as special cases. It is known that the extended model provides vastly improved retrieval operations compared with the normal Boolean system.

d)A generalized Boolean model which transforms the Boolean query space into a vector space, using the Boolean minterms as the basis of the vector space. Like the extended Boolean systems, the generalized model represents a bridge between the vector space and the Boolean model, and it also accommodates document term weights. In addition, the generalized model takes into account dependencies between the terms. [14]

In summary, the Boolean system appears to bring flexibility to the retrieval environment by adding some term relationships as part of the query formulations. However, a strict interpretation of the Boolean operators does not fit the needs of the retrieval application. In practice, the query-document comparisons are better based on global, approximate matches between sets of weighted terms. The proposed extensions to the Boolean system relate the Boolean operations to the vector processing model, and allow term weighting as well as some use of term relationships. However, these models vastly complicate the normal Boolean processing, and they have not so far found their way into practical applications.

3. THE PROBABILISTIC RETRIEVAL MODEL

The probabilistic models were first introduced in the early 1960's and represent an attempt to put the retrieval operations on a sound theoretical basis (For a review, see f.i. [28]). The basic premise is that a document should be retrieved if its probability of relevance to the user's needs exceeds the probability of nonrelevance. The probabilistic approach thus introduces the notion of relevance and nonrelevance of a document which is absent from the vector and Boolean models. This renders necessary the distinction of term characteristics in the relevant and nonrelevant portions of a collection. [15-17]

The main attraction of the probabilistic models is that in principle a large number of phenomena about terms and their occurrence characteristics may be taken into account, including for example term coocurrences for any subset of terms; term relationship indications derived, for example, from existing semantic nets or other constructs used in artificial intelligence approaches; historical knowledge about how well certain terms may have done previously in retrieving relevant information in response to similar information needs; information about term meaning and term relationships derived from dictionaries and thesauruses; and any prior knowledge about the occurrence distribution of terms in certain parts of the collection. Because the probabilistic model can accommodate all this inteligence about documents and queries, it offers the promise of vastly greater effectiveness than the basic vector and Boolean models.

The difficulty with the probabilistic approach is that for the most part accurate information about term dependencies and term characterizations is unavailable, and that the distinction between relevant and nonrelevant information is to no avail

when correct relevance information is in fact not accessible. In practice, it has then become necessary to rely on much simplified pobabilistic models that do not in fact provide more information than the conceptually simpler vector space models. The best known of the simple probabilistic models is based on the assumption that the terms are independently assigned to both the relevant and the nonrelevant documents of a collection, and that binary indexing is used. In this case, the term dependency information and the term weights are both disregarded. Under the term independence assumption the probability of relevance (nonrelevance) of a document becomes the product of the probabilities of relevance (nonrelevance) of the individual terms. That is,

$$\Pr(x \mid \text{rel}) = \prod_{i=1}^{t} \Pr(x_i \mid \text{rel}) \text{ and } \Pr(x \mid \text{nonrel}) = \prod_{i=1}^{t} \Pr(x_i \mid \text{nonrel}) .$$

The foregoing expressions produce an optimal weight for query (but not for document) terms, defined as follows [18,19]

$$w_{q_i} = \log \frac{\Pr(x_i=1 \mid \text{rel}) \Pr(x_i=0 \mid \text{nonrel})}{\Pr(x_i=0 \mid \text{rel}) \Pr(x_i=1 \mid \text{nonrel})} .$$

Under appropriate assumptions, this weighting form is effectively equivalent to an inverse document frequency (idf) weight. [20,21].

The simple probabilistic approach thus leads to optimum query term weight assignments that exclude the term frequency and document length normalization factors routinely used in the vector processing system. This may explain the fact that the simple probabilistic system has not been found especially effective in practice. Furthermore, even this simple model becomes unusable when accurate information about the occurrence characteristics of the individual terms x_i in the relevant and nonrelevant documents of a collection is unavailable.

Some attempts have been made to include term frequency (tf) factors in the probabilistic mold by interpreting the importance factor of a term in a given document as a probability that is estimated by the frequency of occurrence of the term in the individual documents. [22] In these circumstances, the probabilistic model approximates a vector processing model using (tf X idf) weights, that does however offer the eventual possibility of taken into account any term dependence information that might become available. In practice, a useful method for estimating the characteristics of dependent term groups in the relavant and nonrelevant document portions has not so far been found. This restricts the probabilistic approach to models from which much of the needed information is effectively excluded:

a)Because the probabilistic system is not based on existing initial query formulations, the opportunity of independent weighting of query and document terms that exists in the vector system is lost in the probabilistic environment.

b)In the normal relevance feedback approach, the initial query terms are considered to be crucially important. Since initial query terms are not available in the probabilistic system, a probabilistic relevance feedback operation may produce inferior results.

c)The probabilistic approach can incorporate unspecified term dependencies; no distinction is made, however, between different types of dependencies of the kind implicitly specified in the Boolean model (where term synonyms are expressed by or-operators, and term phrases by and-operators). In practice, a completely parallel treatment of very different classes of term dependencies may not produce useful retrieval results.

d)Some objective measurements that are routinely used in a vector system, such as the number of terms attached to a document, or the sum of the weights of the document terms, are excluded from the existing probabilistic approaches.

It seems clear that there is little chance that the probabilistic approach will outperform the vector processing method so long as the term independence and binary indexing restrictions are maintained. More sophisticated probabilistic models must be designed, and methods must be found for estimating the probabilistic term occurrence characteristics in the relevant and nonrelevant items for higher-order groups of terms. Various possibilities in this direction have been investigated, including the use of postulated initial term distributions across the document sets, the collection of prior historical information about term usefulness, and the use of judgments about the relevance of documents with respect to queries obtained from the users during the course of the retrieval operations. Large sets of user queries, retrieved documents, and query-document relevance information will have to be captured if the needed probabilistic parameters are to be accurately estimated.

4. THE LANGUAGE PROCESSING APPROACH

In the vector processing system, document content is assumed to be representable by sets of (possibly weighted) single terms. In the other retrieval models certain relationships between terms can be taken into account, including in particular synonym and phrases represented by or- and and-query clauses respectively. Attempts have also been made to build thesauruses, or vocabulary schedules, either manually or automatically, and to incorporate additional vocabulary control measures in the indexing and retrieval processes.

Unfortunately, the construction of thesauruses and other vocabulary normalization aids is an art, and there is no guarantee that a thesaurus tailored to a particular collection can be usefully adapted to another collection. More generally, the available evidence indicates that using thesauruses and phrase generation systems as part of the document or query content analysis will not produce reliable improvements in retrieval effectiveness for many document collections. The problem seems to be that a broad interpretation of the notion of term relationship (such as the generation of phrases by using coocurrence characteristics of terms in the documents of a collection) produces some good term groups, but also many poor ones; on the other hand, a stricter interpretation of term relationships (such as the inclusion of syntactic criteria in a phrase construction process) generates fewer erroneous word groups but also fewer correct ones. Overall, little progress has been made with simple vocabulary normalization tools such as thesauruses and phrase construction methods used in a retrieval setting. [23]

In recent years, attempts have been made to conduct more thorough text analysis investigations using so-called knowledge-based approaches. In that case, deep representations of particular subject areas are used in the form of semantic nets, frames, or scripts, and attempts are made to include many types of relationship between objects. [24,25] Using the knowledge-bases approach, it becomes necessary to formulate rules for manipulating the knowledge structure, and for relating the words included in individual document and query texts to the entities included in the knowledge base.

There is substantial evidence that in well circumscribed circumstances and strictly limited subject domains, useful knowledge structures can in fact be built. As a result, rule-based expert systems using complex knowledge representations have been built to provide solutions to certain puzzles, or advice about certain financial problems, or help in diagnosing illnesses. However, a complete theory of knowledge representation does not exist at the present time, and in areas that are not severely restricted, it is unclear what type of knowledge representation is needed, what entities must be included in the structure, and what relationships between entities and common sense knowledge elements must be considered.

At the moment, there is no evidence that the knowledge-based approaches are viable for the analysis of ordinary document collections. In addition, fundamental objections have been voiced to the idea of building deep knowledge structures in advance, purporting to represent particular subject areas and divorced from the background, interests, and experience of the intended user populations. If the

non-rationalist tradition of philosophy is to be believed, understanding and knowledge are not gained by formal operations performed on well-defined objects, identified by well-defined properties. Instead understanding requires the background and participation of each individual and must be filtered by people's experiences. [26,27] In that case, the current knowledge-based approaches must lead to a dead end, because the knowledge base would then have to be tailored specifically to each particular user.

In summary, one concludes with an apparently counterintuitive observation: the more simple-minded approaches to text processing and language understanding produce the fewest mistakes and exhibit the best performance. This accounts for the fact that the single term indexing theories based mainly on objective term occurrence characteristics are still more powerful in retrieval than apparently more sophisticated procedures that are less well understood and harder to manage in practical environments.

REFERENCES

[1] Salton, G., Yang, C.S. and Wong, A., A Vector Space Model for Automatic Indexing, Communications of the ACM (1975) 18 (11), pp. 613-620

[2] Salton, G., Yang, C.S. and Yu, C.T., A Theory of Term Importance in Automatic Indexing, Journal of the ASIS (1975) 26 (1) pp. 33-44

[3] Salton, G., A Theory of Indexing, Regional Conference Series in Applied Mathematics (February 1975) 18, SIAM, Philadelphia, P.A.

[4] Salton, G., and McGill, M.J., Introduction to Modern Information Retrieval, McGraw Hill Book Co., (New York 1983).

[5] Raghavan, V.V., and Wong, S.K.M., A Critical Analysis of Vector Space Model for Information Retrieval, JASIS 37 (2) (1986) pp. 279-287.

[6] Wong, S.K.M., Ziarko, W., and Wong, P.C.N., Generalized Vector Space Model In Information Retrieval, Proc. 8th Annual Int. ACM-SIGIR Conference, Montreal (1985) ACM, NY, pp. 18-25.

[7] Cleverdon, C.W., Optimizing convenient On-Line Access to Bibliographic Databases, Information Services and Use, 4 (1984) pp. 37-47.

[8] Radecki, T., Mathematical Model of Information Retrieval Based on the Concept of a Fuzzy Thesaurus, Information Processing And Management 12 (5) (1976) pp. 313-318.

[9] Bookstein, A., Fuzzy Requests: An Approach to Weighted Boolean Searches, Journal of the ASIS, 31 (4) (1980) pp. 240-247.

[10] Buell, D.A. and Kraft, D.H., A Model for Weighted Retrieval System, Journal of the ASIS, 32 (3) (1981), pp. 211-216.

[11] Salton, G., Fox, E.A. and Wu, H., Extended Boolean Information Retrieval, Communications of the ACM, 26 (11) (1983) pp. 1022-1031.

[12] Salton, G., and Voorhees, E., Automatic Assignment of Soft Boolean Operators, Proc. 8th Annual ACM-SIGIR Conference on Research and Development in Information Retrieval, Montreal, ACM, NY, (1985) pp. 54-69.

[13] Salton, G., Fox, E.A. and Voorhees, E., Advanced Feedback Methods in Information Retrieval, Journal of the ASIS, 36 (3) (1985) pp. 200-210.

[14] Wong, S.K.M., Ziarko, W., Raghavan, V.V., and Wong, P.C.N., On Extending the Vector Space Model for Boolean Query Processing, Proc. 9th Annual International ACM-SIGIR Conference, Pisa, Italy, ACM, NY (1986) pp. 175-185.

[15] Bookstein, A., and Swanson, D.R., A Decision Theoretic Foundation for Indexing, Journal of the ASIS, 26 (1) (1975) pp. 45-50.

[16] van Rijsbergen, C.J., A Theoretical Basis for the Use of Cooccurrence Data in Information Retrieval, Journal of Documentation, 33 (2) (1977) pp. 106-119.

[17] Maron, M.E., and Kuhns, J.C., On Relevance, Probabilistic Indexing and Information Retrieval, Journal of the ACM, 7 (3) (1960) pp. 216-244.
[18] Robertson, S.E., and Sparck Jones, K., Relevance Weighting of Search Terms, Journal of the ASIS, 27 (3) (1976) pp. 129-146.
[19] Yu, C.T. and Salton, G., Precision Weighting-An Effective Automatic Indexing Method, Journal of the ACM, 23 (1) (1976) pp. 76-88.
[20] Croft, W.B., and Harper, D.J., Using Probabilistic Models of Document Retrieval Without Relevance Information, Journal of Documentation, 35 (4) (1979) pp. 285-295.
[21] Wu, H., and Salton, G., A comparison of Search Term Weighting, Term Relevance versus Inverse Document Frequency, SIGIR Forum, 16 (1), (Summer 1981) pp. 30-39.
[22] Croft, W.B., Document Representation in Probabilistic Models of Information Retrieval, Journal of the ASIS, 32 (6) (1981) pp. 451-459.
[23] Salton, G., On the Use of Term Associations in Automatic Information Retrieval, Proc. Coling-86, University of Bonn, Bonn, Germany, (September 1986) pp. 380-386.
[24] Schank, R.C. and Abelson, R.P;, Scripts Plans Goals and Understanding, Laurence Erlbaum Associates, Hillsdale, NJ, (1977).
[25] Minsky, M., A Framework for Representing Knowledge, in The Psychology of Computer Vision, P. Winston, editor, McGraw Hill Book Co., NY (1975).
[26] Winograd, T., and Flores, F., Understanding Computers and Cognition, Ablex publishing Corporation, Norwood, NJ, (1986).
[27] Dreyfus, H.L., and Dreyfus, S.E., Mind Over Machine, The Free Press, NY (1986).
[28] Bookstein, A., Probability and fuzzy-set applications to information retrieval, Annual Review of Information Science and Technology, 20 (1985) pp. 117-151.