ON THE REPRESENTATION OF QUERY TERM RELATIONS BY SOFT BOOLEAN OPERATORS

Gerard Salton*
Department of Computer Science
Cornell University
Ithaca, NY 14853, USA

ABSTRACT

The language analysis component in most text retrieval systems is confined to a recognition of noun phrases of the type normally included in back-of-the-book indexes, and an identification of related terms included in a preconstructed thesaurus of quasi-synonyms. Even such a restricted language analysis is fraught with difficulties because of the well-known problems in the analysis of compound nominals, and the hazards and cost of constructing word synonym classes valid for large text samples.

In this study an extended (soft) Boolean logic is used for the formulation of information retrieval queries which is capable of representing both the use of compound noun phrases as well as the inclusion of synonym constructions in the query statements. The operations of the extended Boolean logic are described, and evaluation output is included to demonstrate the effectiveness of the extended logic compared with that of ordinary text retrieval systems.

1. Linguistic Approaches in Information Retrieval

It is possible to classify the various automatic text processing systems by the depth and type of linguistic analysis needed for their operations. Sophisticated language understanding components are believed to be essential to carry out automatic text transformations such as text abstracting and text translation. [1,14,24] Complete language understanding systems are also needed in automatic question-answering where direct responses to user queries are automatically generated by the system. [11] On the other hand, relatively less sophisticated language analysis systems may be adequate for bibliographic information retrieval, where references as opposed to direct answers are retrieved in response to user queries. [21]

In bibliographic retrieval, the content of individual documents is normally represented by sets of key words, or key phrases, and only a few specified term relationships are recognized using preconstructed dictionaries or thesauruses. Even in this relatively simplified environment one does not normally undertake a linguistic analysis of any scope. In fact, syntactic and semantic analysis have been used in bibliographic information retrieval only under special circumstances to analyze query phrases [22], to process structured text samples of a certain kind, [7,15], or finally to process texts in severely restricted topic areas. [2]

Where special conditions do not obtain, the preferred approach in information retrieval has been to use statistical or probabilistic criteria for the generation of the content identifiers assigned to documents and search queries. Obviously, not all terms are equally useful for content identification. According to the term discrimination theory, the following criteria are of importance in this connection [16]:

a) terms which occur with high frequency in the documents of a collection are not preferred for content representation because such terms are too broad to distinguish the documents from each other;

b) terms which occur with very low frequency in the collection are also not optimal, because such terms affect only a very small fraction of documents;

c) the best terms tend to be low-to-medium frequency entities which can be produced by taking single terms that exhibit the required frequency characteristics; alternatively, it is possible to obtain medium frequency entities by refining high frequency terms thereby rendering them more narrow, or by broadening low frequency terms.

In many operational information situations, the term broadening and narrowing operations are effectively carried out by using formulations in which the terms are connected by Boolean operators. The use of Boolean logic in retrieval is discussed in more detail in the remainder of this note.

* Department of Computer Science, Cornell University, Ithaca, New York 14853.
This study was supported in part by the National Science Foundation under grant IST 83-16166.

2. Extended Boolean Logic in Information Retrieval

It is customary to express information search requests by using Boolean formulas that include the operators and, or, and not. Of particular interest in a linguistic context are the and and or operators:

a) The and-operator is a device for specifying a compulsory phrase where all terms in the and-clause must be present to affect the retrieval operation. Thus a query statement such as "information and retrieval" is used to represent the compound nominals "information retrieval", or "retrieval of information". The and-operator is used as a refining device since a broad term such as "information" is made more specific when it is incorporated in an and-clause.

b) The or-operator, on the other hand, is a device for specifying a group of synonymous terms, or alternatively, a thesaurus class of terms in which all terms are treated as coequal. That is, any one term in an or-clause will cause retrieval of the corresponding document, and each term is assumed to be as good as any other term. The or-operator is a broadening device because each or-clause has a broader scope than any individual clause component.

While the logical operators and and or are used universally in retrieval environments, the assumptions of Boolean logic are not verified in normal text processing environments. Strict synonyms occur relatively rarely in query formulations or in the texts of documents, so that the normal or-clause does not reflect a practical situation. In fact, it should be possible to make distinctions between more or less important terms in an or-clause; furthermore, or-clauses should be usable to represent collections of loosely related terms instead of only strict synonyms. Analogously, it should be possible to relax the compulsory nature of the phrase components included in an and-clause, and distinctions ought to be introducable between phrase components of greater or lesser importance.

In summary, the uncertain (fuzzy) nature of the term relationships which obtain in the natural language are not reflected by the rules of ordinary Boolean logic. [25] Instead a relaxed type of logic is needed which is capable of broadening or narrowing the term units, while also providing for distinctions in term importance and for the specification of fuzzy or soft term relationships. Such an extended logical system was introduced recently with the following main properties: [17-18]

a) The extended logic system distinguishes among more or less important terms in both queries and documents by using weights, or importance indicators attached to the terms. Thus instead of terms A and B, the system processes terms (A,a) and (B,b) respectively, where a and b designate the weights of terms A and B.

b) The extended system simulates the linguistic characteristics of more or less strict synonyms, by attaching a p-value to each or-operator that specifies the degree of strictness of the corresponding operator. The higher the p-value attached to an operator, the closer is the interpretation of that operator in accordance with the rules of ordinary Boolean logic. On the other hand, the smaller the p-value, the more relaxed is the interpretation of the or-operator.

c) The extended system also simulates the linguistic characteristics of more or less strict phrase attachment, by using a p-value for each and-operator. The higher the p-value, the more similar the corresponding operator will be to the compulsory Boolean and. Correspondingly, the smaller the p-value, the more relaxed is the interpretation of the and operator.

d) The extended system (unlike the ordinary Boolean system) provides ranked output of the stored documents in presumed decreasing order of importance of a given item with respect to a query. In addition, the extended system provides much better retrieval output, than systems based on conventional Boolean logic. Experimentally, improvements of 100 to 200 percent in retrieval effectiveness have been noted for the extended logic over the conventional Boolean system. [17,18]

It is not possible in the present context to furnish the details of the operation of the extended logic system. The following results are, however, relatively easy to prove: [17]

a) When p-values equal to infinity are used, the extended system produces results identical to that of the conventional Boolean logic systems;

b) When the p-values are reduced from infinity, the distinctions between phrase components (and) and synonym specification (or) become more and more blurred;

c) When p reaches its lower limit of 1, the distinction between and and or operators is completely lost, and the system reduces the queries (A and B) and (A or B) to a system with terms (A,B), without any relationship specification between terms A and B.

Using linguistic analogues, the following examples illustrate the operations of the extended logic system. The p-value attached to operators is shown in each case as an exponent:

i) (A $\underline{\text{and}}^{\infty}$ B) interpreted as ALL OF (A,B) (strict phrase)

ii) (A $\underline{\text{and}}^{3}$ B) interpreted as MOST OF (A,B) (fuzzy phrase)

iii) (A $\underline{\text{and}}^{1}$ B) interpreted as SET (A,B) (more matching terms are worth more than fewer matching terms)

iv) (A $\underline{\text{or}}^{1}$ B) identical to (A $\underline{\text{and}}^{1}$ B) interpreted as SET (A,B)

v) (A $\underline{\text{or}}^{3}$ B) interpreted as SOME OF (A,B) (fuzzy synonym)

vi) (A $\underline{\text{or}}^{\infty}$ B) interpreted as ONE OF (A,B) (strict synonym)

3. Experimental Results

The operations of the extended logic system are illustrated by using a collection of 3204 computer science articles (titles and abstracts) originally published in the Communications of the ACM (the CACM collection), and a collection of 1460 articles in library science obtained from the Institute for Scientific Information (the CISI collection). Table 1 shows average performance figures for 7 selected queries used with CACM, and 4 selected queries for CISI. The performance in Table 1 is stated in terms of the search precision at various recall points averaged over the set of search requests in use. [19]

The data of Table 1 indicate that the conventional Boolean searches (p = ∞, Boolean) produce by far the worst performance for both collections. Performance improvements between 100 and 200 percent are obtained by relaxing the interpretation of the Boolean operators (that is, by using lower p-values). A distinction must be made between taking into account only single term matches (p-values are equal to 1), and giving extra weight to term phrase matches (A and B and ...), and to synonym set matches (A or B or ...), when p-values higher than 1 must be used. The results of Table 1 show that for the CACM queries the best overall policy is a complete softening of the Boolean operators down to p = 1. Evidently not many of the quasi-Boolean phrases included in the CACM queries were also present in the document abstracts. For the ISI queries, on the other hand, 154 percent improvement is produced when p = 1; when the phrase combinations are given extra weight, the improvement in performance jumps to 164 percent for p = 2, and to 182 percent when and- and or-operators are given different values (p and = 2.5 and p or = 1.5, respectively).

These phenomena are further illustrated in the output of Tables 2 and 3. The comparison between query CACM Q5 and Document 756 is outlined in Table 2. No abstract was available for document 756; hence only the title words could be used in the query-document comparison. As the example shows, only the term "editing" was present in both document title and query. This explains why the single term match (p = 1) produces the best output rank of 5 for this document. Obviously, the sample document is not retrievable by the pure Boolean search (p = ∞) as demonstrated by the simulated retrieval rank of 1667 out of 3204 CACM documents.

Table 3 shows an example where matching phrases make a substantial difference in the retrieval results. The matched phrases in Document 1410 are given a double underline in Table 3, whereas matched single terms have a single underline. The output of Table 3 shows that when the single terms alone are considered, document 1410 is retrieved with a rank of 53 in response to query ISI Q33. When the phrase matches are given extra weight (p = 2, or p and = 5, p or = 2), the retrieval rank improves to 2 and 7, respectively.

These results demonstrate that the conventional Boolean logic does not adequately reflect the tentative and uncertain nature of the relations between terms in the language. When a relaxed interpretation of Boolean logic is used, the correspondence with the fuzzy nature of linguistic relations is much greater and dramatic improvements in term matching and hence retrieval effectiveness are obtained.

4. Relationship of Extended Boolean Model with Other Retrieval Developments

The extended Boolean system is based on the use of certain term relationships--notably term phrases and synonymous constructions. These relations are, however, interpreted flexibly, reflecting the uncertain nature of term relations in the language. In the extended system, soft Boolean queries are easy to formulate, and methods exist for a completely automatic formulation of the soft queries, given only some basic information about user needs. [20] Analogously, initial queries may be automatically reformulated, following an initial search operation, based on information obtained from the user about the relevance of previously retrieved documents. [18]

The current development may then be related to other retrieval models that incorporate term relations, and to systems with advanced user interfaces. Term relations of a statistical, or probabilistic nature are included in the probabilistic retrieval model; more general linguistic relations are used in systems that include a natural language analyzer. In the probabilistic retrieval system, the documents are ranked in decreasing order of the probabilistic expression $P(\underline{x}|rel)/P(\underline{x}|nonrel)$ where $P(\underline{x}|rel)$ and $P(\underline{x}|nonrel)$ represent the occurrence probabilities of an item $\underline{x}$ in the relevant and non-relevant document subsets, respectively. [23] The

Type of Query-Document Comparisons	CACM Collection 7 selected queries (5,6,9,12,15,21,40)	CISI Collection 4 selected queries 4,7,18,33
p = ∞, strict Boolean interpretation	.2020	.1465
p = ∞, weighted document terms (fuzzy set interpretation)	.2170 (+7.5%)	.1978 (+35.0%)
p = 1, only single terms taken into account, weighted terms	.4812 (+138.2%)	.3733 (+154.8%)
p = 2, some and and or combinations taken into account, weighted terms	.3779 (+87.1%)	.3879 (+164.8%)
p (and) = 2.5 anded phrases p (or) = 1.5 count more than ored combinations	.4164 (+106.2%)	.4136 (+182.4%)
p (and) = 5.0 anded phrases p (or) = 2.0 much more strict than ored combinations	.3758 (+86.1%)	.3966 (+170.7%)

Average Search Precision at Three Recall Points (0.25, 0.50, 0.75) for Two Collections

Table 1

CACM Q5 Query Statement (natural language)

Design and implementation of editing interfaces, window-managers, command interpreters, etc. The essential issues are human interface design, with views on improvements to user efficiency, effectiveness and satisfaction

Boolean Form (partial statement)

(editing) and [(human and satisfaction) or (user and satisfaction) or (human and efficiency) or (....)]

Document 756 A Computer Program for Editing the News
(no abstract, one single term match with query)

Retrieval Ranks for Document 756

p = oo Boolean	Rank	1667
p = 1	Rank	5
p = 2	Rank	10
p and = 5, p or = 2	Rank	13

Illustration for Single Term Match of Item Rejected by Conventional Search.

Table 2

ISI Q33 Query Statement (natural language)

Retrieval systems providing the automated transmission of information to the user from a distance

Boolean Form (partial statement)

[(distance or transmission) and (retrieval or informaton)]

or (telefacsimile and system) or ...

Document 1410 Telefacsimile in Libraries

(/ single term match)
(// phrase match)

The use of telefacsimile systems to provide rapid transfer of information has great appeal. Because of a growing interest in the applicability of this technology to libraries, a grant was provided to the Institute of Library Research to conduct an experiment in telefacsimile equipment in a working library situation. The feasibility of telefacsimile for interlibrary use was explored. Information is provided on the performance, cost, and utility of telefacsimile systems for libraries

Retrieval Ranks for Doc 1410

p = ∞ Boolean	Rank	29
p = 1	Rank	53
p = 2	Rank	2
pand = 5, por = 2	Rank	7

Illustration for Phrase Matching Process

Table 3

required occurrence probabilities of the various documents depend on the occurrence probabilities in the respective document subsets of the individual terms x_i, x_j, x_k, etc. When term relationships are to be used, the occurrence probabilities must also be available for term pairs--for example, $P(x_{ij}|rel)$, and $P(x_{ij}|nonrel)$; for term triples $P(x_{ijk}|rel)$, $P(x_{ijk}|nonrel)$, and so on, for higher order term combinations.

Unfortunately, the experiences accumulated with the probabilistic retrieval model show that enough information is rarely available in practical situations to render possible an accurate estimation of the needed probabilities. In practice, it then becomes necessary to avoid the use of term dependencies by assuming that all terms occur independently. The probabilistic model is then effectively equivalent to a vector processing system that does not include any term relations. [3]

When linguistic analysis methods are used to analyze query and document content, it is in theory possible to provide a precise representation of query and document content by including a great variety of term relations in the search and retrieval operations. In particular, complex indexing units such as noun and prepositional phrases might then be assigned to the information items for content representation. Unfortunately, a complete treatment of noun phrases by automatic means remains elusive in view of the multiplicity of different term relations that are expressible by noun and prepositional phrases. An automatic recognition of semantically equivalent noun phrases of the kind needed for the construction of classification schedules is also exceedingly difficult.

For practical purposes, the use of term relations that is theoretically possible in the probabilistic and language-based retrieval models is

thus of questionable help in general retrieval situations where topic areas and linguistic complexities are not severely restricted. The Boolean model which includes only a general phrase (denoted by the Boolean and) and a general synonym relation (denoted by the Boolean or) may not therefore represent an intolerable simplification when measured against the realistically possible, alternative methodologies.

Considering now the user-system interfaces that have been designed for use in information retrieval, the following types of development may be distinguished.

a) The use of minicomputer-based file accessing methods providing simple access to specific data bases, or to specific file catalogs. Such systems are often menu-driven and offer a conversational style, permitting the user to consult a given term classification or thesaurus, and to browse through the document corresponding to a given query formulation. [4,6]

b) The construction of large, sophisticated systems designed to provide unified interface methods to a variety of data bases implemented on a single retrieval facility, or to data bases available on a multiplicity of different retrieval systems. [12,13] A common command language may then be provided by the interface system, in addition to tutorial and help provisions, or even diagnostic procedures able to detect, and possibly to correct questionable search strategies.

c) The use of interface methods based on fancy graphic displays that make it possible to exhibit vocabulary schedules, command sequences, and messages that may be helpful during the course of the search operations. [5,10]

d) The simulation of automatic "search experts" that are able to translate arbitrary queries in natural language by using stored knowledge bases for query analysis and search purposes. Such automatic experts may perform the work normally assigned to human search intermediaries, in the sense that a conversational dialog system ascertains user requirements and chooses search strategies corresponding to particular user needs. [8,9]

In each case the automatic interface system is designed to help the user to access a possibly unfamiliar retrieval system and to pick a useful search strategy. The operational retrieval system that actually performs the searches is normally not modified by the interface system. The extended Boolean system described in this note differs from these other developments because the conventional search system is actually modified by replacing a complete Boolean match by a fuzzy query-document comparison system. Furthermore, the burden placed on the user during the query construction process is kept as small as possible.

The minicomputer-based facilities and the fancy graphic display systems may be used in conjunction with the extended Boolean processing, since the two types of developments are somewhat independent of each other. The same is true of the systems that provide common interfaces to multiple data bases. The retrieval expert capable of interacting with the user in natural language may not be usable in practical situations for some years to come, unless severe restrictions are imposed on the topic areas under consideration, and the freedom of formulating the search requests. An interface system of more limited scope may be more effective under current circumstances than the automated "expert" of the future.

REFERENCES

[1] T.R. Addis, Machine Understanding of Natural Language, Int. Journal of Man-Machine Studies, Vol. 9, 1977, 207-222.

[2] L.M. Bernstein and R.E. Williamson, Testing a National Language Retrieval System for a Full-Text Knowledge Base, JASIS, 35:4, July 1984, 235-247.

[3] A. Bookstein, Explanation and Generalization of Vector Models in Information Retrieval, Lecture Notes in Computer Science, Vol. 146, Springer-Verlag, Berlin, 1983.

[4] E.G. Fayen and M. Cochran, A New User Interface for the Dartmouth On-Line Catalog, Proc. 1982 National On-Line Meeting, Learned Information Inc., Medford, NJ, March 1982, 87-97.

[5] H.P. Frei and J.F. Jauslin, Graphical Presentation of Information and Services: A User Oriented Interface, Information Technology: Research and Development, Vol. 2, 1983, 23-42.

[6] C.M. Goldstein and W.H. Ford, The User Cordial Interface, On-Line Review, 2:3, 1978, 269-275.

[7] R. Grishman and L. Hirschman, Question Answering from Medical Data Bases, Artificial Intelligence, Vol. 11, 1978, 25-43.

[8] G. Guida and C. Tasso, An Expert Intermediary System for Interactive Document Retrieval, Automatica, 19:6, 1983, 759-766.

[9] L.R. Harris, Natural Language Data Base Query, Report TR 77-2, Computer Science Department, Dartmouth College, Hanover, NH, February 1977.

[10] G.E. Heidorn, K. Jensen, L.A. Miller, R.J. Byrd and M.S. Chodorow, The Epistle Text Critiquing System, IBM Systems Journal, 21:3, 1982, 305-326.

[11] W. Lehnert, The Process of Question-Answering, (Ph.D. Dissertation), Research Report No. 88, Computer Science Department, Yale University, New Haven, CT, May 1977.

[12] R.S. Marcus, An Experimental Comparison of the Effectiveness of Computers and Humans as Search Intermediaries, Journal of the ASIS, 34:6, 1983, 381-404.

[13] C.T. Meadow, T.T. Hewett and E.S. Aversa, A Computer Intermediary for Interactive Data Base Searching, Part I: Design, Part II: Evaluation, Journal of the ASIS, 33:5, 1982, 325-332 and 33:6, 1982, 357-364.

[14] N. Sager, Computational Linguistics, in Natural Language in Information Science, D.E. Walker, H. Karlgren and M. Kay, editors, FID Publication 551, Skriptor, Stockholm, 1977, 75-100.

[15] N. Sager, Sublanguage Grammars in Science Information Processing, Journal of the ASIS, January-February 1975, 10-16.

[16] G. Salton, C.S. Yang, and C.T. Yu, A Theory of Term Importance in Automatic Text Analysis and Information Retrieval, Journal of the ASIS, 26:1, January-February 1975, 33-44.

[17] G. Salton, E.A. Fox and H. Wu, Extended Boolean Information Retrieval, Communications of the ACM, 26:11, November 1983, 1022-1036.

[18] G. Salton, E.A. Fox, and E. Voorhees, Advanced Feedback Methods in Information Retrieval, Technical Report 83-570, Department of Computer Science, Cornell University, Ithaca, NY, August 1983.

[19] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill Book Company, New York, 1983.

[20] G. Salton, C. Buckley and E.A. Fox, Automatic Query Formulations in Information Retrieval, Journal of the ASIS, 34:4, July 1983, 262-280.

[21] K. Sparck Jones and M. Kay, Linguistics and Information Science: A Postscript, in Natural Language in Information Science, D.E. Walker, H. Karlgren and M. Kay, editors, FID Publication 551, Skriptor, Stockholm, 1977, 183-192.

[22] K. Sparck Jones and J.I. Tait, Automatic Search Term Variant Generation, Journal of Documentation, 40:1, March 1984, 50-66.

[23] C.J. van Rijsbergen, Information Retrieval, Second Edition, Butterworths, London, 1979.

[24] D.E. Walker, The Organization and Use of Information: Contributions of System for a Full-Text Knowledge Base, JASIS, 35:4, July 1984, 235-247. Information Science, Computational Linguistics and Artificial Intelligence, Journal of the ASIS, 32:5, September 1981, 347-363.

[25] L.A. Zadeh, Making Computers Think Like People, IEEE Spectrum, 21:8, August 1984, 26-32.