Query Composition: Why Does It Have to Be So Hard?

Lucy Terry Nowell, Deborah Hix, and Eric D. Labow

TR 93-19

Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

June 10, 1993

Accepted in East-West International Conference on Human-Computer Interaction, Moscow, 1993

Query Composition: Why Does It Have to Be So Hard?

Lucy Terry Nowell, Deborah Hix, and Eric D. Labow

Department of Computer Science, Virginia Tech
Blacksburg, VA 24061-0106 USA
Email: nowell@vtcc1.cc.vt.edu hix@cs.vt.edu elabow@manu.cs.vt.edu

Abstract. Project Envision, a large research effort at Virginia Tech, focuses on developing a user centered, multimedia database from the computer science literature, with full- text searching and full-content retrieval capabilities. User interviews indicate that people have trouble composing queries. Widely available boolean retrieval systems present problems with both syntax and logic. Natural language queries for vector space retrieval systems are easier to compose, but users complain that they do not understand the matching principles used; users also complain that they have too little control over the search and fear being overwhelmed by an enormous retrieval set. We describe the Envision query window, which has as a usability goal making query composition easy while increasing user control. Results of formative usability evaluation and subsequent redesign are discussed.

1. Querying Existing Information Retrieval Systems

A *query* is simply a description of information sought, in a form which a computer can use as the basis for searching a database. Retrieval by an information system is typically evaluated by two measures: *recall* and *precision* [5, 9, 10, 11]. Recall is the percentage of relevant documents in the collection that are actually retrieved by the query. Precision is the percentage of documents retrieved which are indeed relevant. The two performance measures are near inverses, so that increasing performance on one decreases performance on the other. User satisfaction with the system depends in part on striking an appropriate balance between precision and recall.

The majority of existing electronic information systems rely on some variety of boolean searching of documents indexed by a specified range of keywords [5, 9]. These systems require users to formulate queries that are syntactically complex and involve logical operations that most users do not fully understand. Simply stated, people are not very good at composing boolean queries [7]. Thus, recall from boolean queries is usually very low, although precision is fairly high. The precision measure may be somewhat confounded by inadvertent elimination of relevant documents due to errors in boolean syntax [1, 5, 7, 9, 14].

The *indexing language*, that restricted list of terms chosen as those by which a given collection is indexed, is itself a problem for boolean systems [6, 9, 11, 12]. Studies have shown very little consistency among human indexing personnel, either across time or across individuals [6]. Additionally, the user must know or guess the keywords by which works of interest have been indexed. Boolean system recall may be significantly lessened by user unfamiliarity with the indexing language or by indexing errors: If the terms in a query are "ANDed" with one term that is not in the index, no documents will be found. A given relevant document may be omitted from the retrieval set because it is indexed by all but one of the "ANDed" terms in a query. A more thorough discussion of the "vocabulary problem" may be found in [6]. Use of search intermediaries, typically reference librarians, is one response to the complexities of boolean systems searching restricted-language indexes.

One alternative to boolean systems, to date found mainly in research systems, is the *vector space retrieval model* [9, 13]. The vector space model allows searches of full document text from natural language queries. The natural language query, or *query vector,* is simply a list of words the user expects to occur in works sought, so the user does not have to struggle with complicated syntax or boolean logic. Because a restricted indexing language is not used, users do not have to guess how the works of interest might be described. Statistical pattern matching algorithms allow the system to infer which documents are most likely to be of interest and to return them to the user in order of probable relevance.

However, we found usability problems with vector space retrieval systems as well. While studying query formulation, we participated in formative usability evaluation of MARIAN [3], a vector space retrieval system at Virginia Tech that uses natural language queries. Two concerns emerged that have shaped our current work:

• Users feared being overwhelmed by a potentially enormous retrieval set, having rapidly realized that words describing works of interest to them may also occur in a large portion of the collection. User concern about very large retrieval sets has been found in other studies, as well [7].

• Users wanted more control over the search than they felt MARIAN gave them. They were confused by the way MARIAN used combinations of query words and by the ordering of results produced. In particular, they missed the ability to eliminate works using a "NOT" operator.

Solutions to some of the problems with MARIAN are in the realm of the search systems used. MARIAN does not yet have the capability to process phrases or compound terms, which would lessen some user confusion. *Probabilistic retrieval,* by which the system reports only on a small number of documents with high probability of usefulness, limits retrieval set size in MARIAN; but test users were not aware of that feature. Ranking by estimated relevance should also alleviate some of the concerns about the number of documents to be examined, but only if the user understands that ranking is occurring and has faith in the ranking algorithm. In any case, the user interface needs to communicate clearly about the kinds of matching that are used and to give users as much control as possible over matching algorithms.

Users accustomed to boolean search systems have suggested combining full-text searching with boolean searching. Their hope is that they will get the benefit of not having to know an indexing language combined with the restrictive power of boolean systems. In practice, such systems are less than satisfactory. Because the complete text of a document is searched, if the user has restricted a term with a "NOT" operator in the query, a single trivial occurrence of the negated term can eliminate an otherwise valuable document from the retrieval set. This result is not anticipated by users of existing boolean systems, since they are accustomed to searches that access only a small part of the vocabulary of each work in the database. Research systems have been developed using *extended boolean searching*, which replaces strict boolean logic with fuzzy logic, softening the impact of the logical operators [4, 12]. Extended boolean search systems provide some protection against elimination or inclusion or works due to trivial occurrence of words. However, extended boolean systems still impose on users the burden of composing syntactically complex queries.

2. Project Envision

Project Envision [2], a large research effort at Virginia Tech, focuses on developing a user centered, multimedia database from the computer science literature. Envision is a step toward fulfilling the dream of digital electronic libraries, providing extensive remote access to research collections. One aim of the project is to develop, implement, and validate new database methods and systems that will help computer scientists and others in computing professions more easily become aware of, locate, manipulate, and understand a variety of objects that are important to the science of computing. Envision serves computer science researchers, teachers, and students at all levels of expertise. Envision also provides a model for developing databases of other scientific literatures.

An important goal for Project Envision is developing a user interface that is demonstrably usable, efficient, and effective in supporting computer scientists working with and exploring the database. To that end, a major portion of the resources of Project Envision has been devoted to development of the user interface, through a process described below. In this paper, we focus on design of the Envision Query Window. The Envision search results display is discussed in [8].

3. Interviews with Users

From the proposal stages through its current prototypes, Envision has been conceived as a user centered system. Users have been closely involved in development of Envision both through an interviewing process that guided decisions about system functionality, as well as through formative usability evaluation.

Over a four month period we interviewed twelve professionals in the areas of computer science and information retrieval. Interviewees were carefully chosen to broadly represent the type of user we expect for Envision. Five interviewees are active industry researchers in the areas of information retrieval and/or human-computer interaction. One is a corporate research department head, responsible for document retrieval to support the work of others. One is a faculty member specializing in information storage and retrieval at another research university. Five Virginia Tech faculty members, three from the Department of Computer Science and two from Industrial and Systems Engineering, were also interviewed.

During intensive interviews lasting from one to two hours, interviewees responded to questions focused on four topics:

1. Current information retrieval practices.
2. Current information dissemination practices.
3. Desired information retrieval and manipulation capabilities.
4. Demographic data.

Interviewees shared reliance on journals and conference attendance as major sources of information, with additional attention to conference proceedings. Talk with colleagues was ranked as equal in importance to journals as a source. Colleagues are especially helpful in providing pointers into the literature — that is, specific references to works likely to be helpful in solving a particular problem. One interviewee indicated that colleagues serve as valuable filters, since they point to the few best works in an area without providing an exhaustive list of less valuable materials. A few interviewees make use of network bulletin board services, but most do not.

When they seek publications relevant to a particular topic, most of our interviewees have used electronic information systems of some kind. These include computerized library catalogs, CD-ROM systems, and online search services. However, our interviewees found the existing systems difficult to use for a variety of reasons. Inadequate access to any electronic information system is one common problem. Interviewees also complained about the difficulty of structuring queries, the number of diverse user interfaces, inadequacy of feedback about unsuccessful searches, and the amount of knowledge required before systems are really usable. Our interviewees generally dislike any requirement or need to consult an intermediary, or search system expert, to access the literature.

The feature our interviewees most want in an information retrieval system is access from their offices or workstations. Most specifically requested or implied the need for full text retrieval. Other features commonly requested include access to multiple forms of information (abstract, resume, brief description, full text, bibliographic entry) about each document retrieved, print capability, user annotation facilities, and the ability to establish and work within a personal subset of the database. A usable interface was mentioned often as a needed feature and complaints about user interfaces of existing electronic information retrieval systems were frequently cited reasons for not using those systems.

User task analysis followed our interviews and was based in part on interview results. Interface design work to date has focused on the major tasks of targeted search and retrieval, and use of search results. Refinement of task analyses is proceeding iteratively with user interface design, prototyping, and usability testing.

4. Envision Query Window Design

Responding to interviewees' concern that an information retrieval system must be accessible from their offices, we have based our design on the premise that the Envision user interface will run as a client process on the user's desktop computer, communicating with the Envision retrieval system server via network. The first implementation platforms for the user interface client are X-Windows/Motif and the Macintosh.[1] Other platforms will follow as resources permit and demand requires. The design will be adapted to comply with standards applicable to the various platforms. The interface designs provide for flexible use of varying configurations of monitors, both in size and number of displays. The minimal configuration supported uses a thirteen inch gray-scale display.

Envision uses a vector space retrieval system, with all the benefits and potential difficulties described previously. The Envision Query Window design gives users the benefits of natural language query formulation expected in a vector space retrieval system (that is, no complex syntax or use of logical operators is required, nor is knowledge of an artificial indexing language), while also providing means to restrict searches in ways requested by users in our work with MARIAN. The Query Window has two categories of use:

- New queries are created and searches performed from this window.
- It provides access to previously completed, or "old," queries and the results of the related searches. Old queries may simply be viewed or they may be revised and used for another search. Results of searches from old queries may also be redisplayed.

The Query Window offers the user three ways to create new queries:

- By entering document descriptors in the new query fields for authors' names, title words, content terms, and words found in other parts of a document as specified by a pop-up menu labeled "Special Query."
- By editing earlier queries.
- By combining results of previously completed searches, using set operations.

4.1 New Query Fields

The Query Window, shown in Figure 1, features four new query fields for Authors, Words in Title, Content Terms, and Special Query, and a Query History field. The Special Query field has a pop-up menu control that allows users to specify searches of abstracts, chapter titles, figure headings, and tables of contents, as well as to enter a complete bibliographic citation as a single block.

When creating a new query or editing an old one, the user may make changes in addition to or instead of simply editing the text in the fields. Other options include changing the match types (explained further below) used for each field, changing the relationship among fields, and changing filters that restrict search results.

As shown in Figure 1, match type options are given to the right of the query fields, with text and related radio buttons. Users have control over whether terms

[1] Macintosh is a registered trademark of Apple Computer, Inc.

within a given field are "ANDed" ("Match all...") or "ORed" ("Match any..."). Some field-specific options apply, as for Authors, where users may specify "Match full names exactly" or "Match family names;" in the latter case the family name is matched exactly and the rest of a name entry is matched as closely as possible but perhaps not exactly. An option to "Match approximately" enables Envision to apply statistical matching to compensate for misspellings, name abbreviations in the query or database, and other minor entry errors and anomalies.

For Words in Title, users may indicate whether ordering of the query words must be matched. For entries in the Content Terms field, users may specify whether the full text of documents is searched or only the title and subject descriptors.

The relationship among the four fields is also user-controlled via radio buttons, one of which must be selected for any query. These are shown with related text in the space below the new query fields. The group is labeled "Match Between Fields: Author, Title, Content, & Special Query." The default setting is "Match all four fields above," which is similar in effect to "ANDing" across the four fields. The alternative is "Match any fields above," which is akin to applying an inclusive "OR" across the fields.

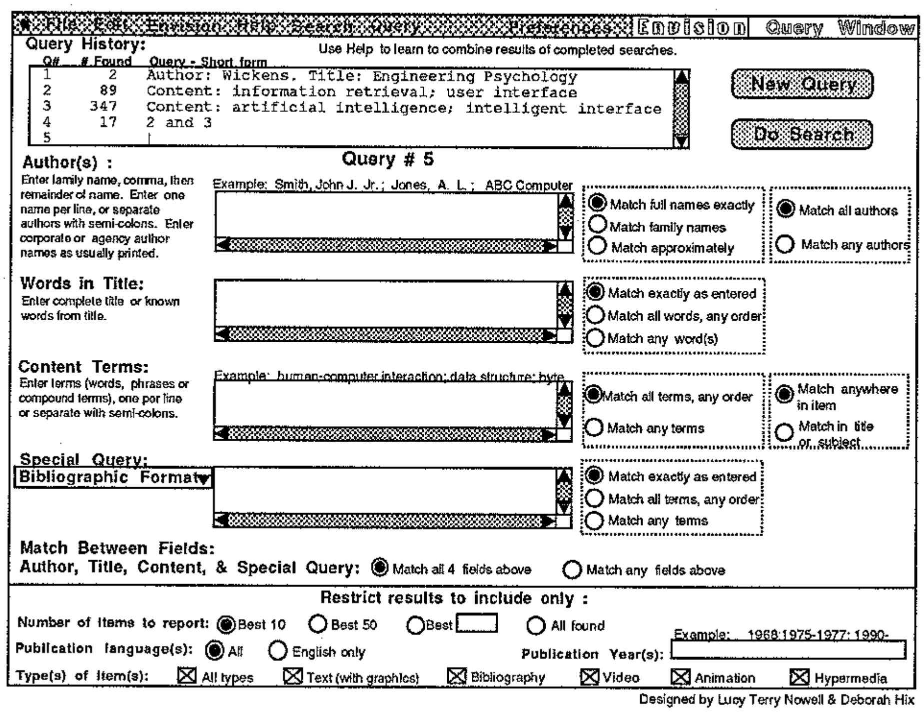

Figure 1. Envision Query Window

4.2 User-Controlled Filters

Filters subject to user control are accessible at the very bottom of the Query Window, below the label "Restrict Results to Include Only." (See Figure 1.) Included are filter controls for publication language, document type, year of publication, and the number of items to be retrieved.

Radio buttons allow selection of publication languages. For the first version of Envision, the choices are limited to "English" and "All." Check-boxes are used for selection of document types, allowing more than one type to be specified. The options are "All Types," "Text with graphics," "Bibliography," "Video," "Animation," and "Hypermedia."

The filter for Publication Years is set via text-entry field in which the user enters a single year, a series of years, and/or a range of years. If no year or date range is specified, all years of publication are assumed to be of interest.

Users control the number of items retrieved via radio buttons at the bottom of the window. The default number of items to retrieve is ten, where the ten most relevant items will be returned. Other options are "Best 50"; "Best __," where the user enters the number of items desired; and "All found."

4.3 Query History

As queries are stored or related searches performed, the user establishes a history which is accessible through the Query History field at the top of the window. (See Figure 1.) In the Query History, a one-line summary form of each query is displayed in order by query number, along with the number of items retrieved by the related search. The Query History provides access to the results of previous searches, means to redisplay the full content of previous queries for possible revision, and a mechanism for combining the results of completed searches.

Previous queries are accessed by selecting the Query History's text summary line for the query. Selection is accomplished either by clicking on the text line, causing the line to be highlighted, or, when the Query History is already active, by using the up and down arrow keys to move the highlight/selection bar to it. Once an old query is selected, the full text of it is displayed in the new query fields of the Query Window and the various buttons show the settings established for that query. The original query number of the old query appears above the new query fields.

When the summary line for a previous query (and completed search) is selected / highlighted and its field values displayed, the button which otherwise reads "Do Search" shows a different label and function: "Show Results." Clicking on this button, making a "Show Results" menu choice, and double-clicking the Query History line will all have the same effect: reopening or bringing to the front the results windows associated with the selected query.

5. Formative Usability Evaluation of the Query Window Design

Prior to rapid prototyping, the Envision Query Window design was modified several times as a result of critique sessions both with prospective users and with human-computer interface design experts. Sessions with the Human-Computer Interaction Research Group at Virginia Tech were particularly productive. A prototype was then created for the Macintosh using SuperCard[2].

5.1 Participants

Usability testing was conducted with four users. Of the four participants, two were experienced Macintosh users, one was primarily a text-interface user with significant experience with graphical user interfaces, and the fourth was a text-interface user with minimal graphical user interface experience. Viewing participants by their areas of expertise, the four included a graduate student in computer science who is also a programmer for a library, an undergraduate in computer science, a computer science faculty member, and a reference librarian.

5.2 Method

Participants unfamiliar with the Macintosh were given brief training in use of a graphical user interface. Participants received no training in use of the Envision user interface itself. They were given time to familiarize themselves with the layout of the window before usability testing began.

During formative usability evaluation of the Query Window, we used eight benchmark tasks, each of which had a number of subtasks. The tasks were described in language not found on the Query Window itself. They required participants to construct a wide variety of queries, some using a single field while others used multiple fields; a variety of author name formats were requested, as were numerous combinations of single words, compound and hyphenated terms, and phrases. The full range of matching options and filter restrictions supported by the interface was utilized in properly completing the tasks. A sample task description is provided in Figure 2.

Prior to usability testing, we established 63 objective usability goals, measured by time for completion of each subtask, number of errors, and number of uses of Help. For the sample task shown in Figure 2, our usability specifications and observed results are shown in Figure 3.

[2] SuperCard is a trademark of Aldus Corporation.

a. Suppose you change your mind before you perform the last search. Erase all of the information in the request.
b. Request works by AT&T, Bell Labs, or Jock Mackinlay.
c. Limit the results to works from 1990 to 1992.
d. Have the works include the words "intelligent" and "layout" somewhere in the title.
e. Request that works meet all requirements from lines "b" through "d" above.
f. Request that 25 works be found.
g. Have Envision perform the search.

Figure 2. Sample task with seven subtasks

Usability Attribute	**Measured By Task**	**Value Measured**	**Worst Accept. Level**	**Planned Level**	**Mean Observed Result**
Initial use: Clear fields	a	Questions	0	0	0
		Errors	1	0	.25
		Time	10	5	6.3 sec.
Learning curve: multiple authors & match-type	b	Errors	0	0	0
		Time	40	20	43.8 sec.
Initial use: Date range	c	Questions	0	0	0
		Errors	0	0	0
		Time	10	5	6 sec.
Learning curve Title field & match-type control	d	Time	30	15	31.5 sec.
Inter-field relationship control	e	Time	4	2	12.8 sec.
Initial use: Best n retrieved	f	Questions	1	0	0
		Errors	1	0	0
		Time	12	6	9 sec.
Learning curve: Do search	g	Time	4	2	1.5 sec.

Figure 3. Usability specifications and observed results for sample task in Figure 2

We had 35 subjective usability goals, measured by a questionnaire adapted from QUIS, the Questionnaire for User Interaction Satisfaction (Copyright © 1988-1991 Human-Computer Interaction Laboratory, University of Maryland at College Park). In the questionnaire, users were asked their opinions of the screen layout, adequacy of instructions provided on the screen, usability of the various features, and their overall impressions of the user interface and system functionality implied by it. The subjective questions were answered on a scale of -3 for the worst response to +3 for the best. Figure 4 shows sample subjective question, with mean participant responses in [ ]. Open-ended questions invited comments on specific strengths and weaknesses of the design and recommended improvements.

---

1. Overall reactions to the system: terrible ... wonderful
   -3 -2 -1 0 1 2 3 NA

   [Mean = 1.75]

1a. difficult ... easy
   -3 -2 -1 0 1 2 3 NA

   [Mean = 1.75]

1b. inadequate power ... adequate power
   -3 -2 -1 0 1 2 3 NA

   [Mean = 2.25]

2. User ability to limit the results of the search by language, years, document types is: confusing ... clear
   -3 -2 -1 0 1 2 3 NA

   [Mean = 2.5]

2a. inadequate ... adequate
   -3 -2 -1 0 1 2 3 NA

   [Mean = 2]

Figure 4. Sample subjective questions from the questionnaire, shown with mean participant response in [ ].

5.3 Analysis

We monitored 31 measures of time for task completion, 16 counts of errors, and 16 counts of questions/uses of help. Planned levels were set equal to the performance of a member of the design team who was not involved in prototype development. Worse acceptable levels were established conservatively at double the planned levels. Across all participants and tasks, mean times equaled or bettered our planned usability goals for 10 subtasks and our worst acceptable case for an additional 11 subtasks; 10 mean times did not meet our worst acceptable case. Since this user interface prototype was specifically for Macintosh users, while one of our participants was known to be unfamiliar with any graphical user interface, we also looked at the results excluding that user. Excluding that person's times, the means met or bettered planned case on 13 subtasks and met or bettered worst acceptable case on another 14 subtasks; on only 4 tasks did the Macintosh-familiar users fail to meet our usability goal for worst acceptable case. All counts of errors and uses of help met or bettered our usability goals for worse acceptable case. For the sample task shown in Figure 2, mean participant time did not meet our worst acceptable level for subtasks b, d, and e; all counts of errors and questions were within the acceptable range.

For 22 subjective questions, the mean user response exceeded +2; for another 10 questions, the means were greater than or equal to +1 and less than +2; while for 3 questions, the mean response fell between 0 and +1. There were no negative mean responses. Once again we looked at the results for only those participants who were familiar with the Macintosh prior to testing the Envision prototype. For those users, 31 questions showed mean responses greater than or equal to +2, with an additional 4 means lying between +1 and +2; no means fell below +1.

Participants particularly liked the control over searching provided by the interface, including the restrictions they could place, and the large amount of information displayed.

5.4 Discussion

Numerous minor design changes have been made to the Query Window as a result of usability evaluation. These changes, reflected in Figure 1, are described below. However, because overall results of the usability evaluation were strongly affirmative of the basic design, it is unchanged.

As we anticipated, an important source of difficulty for our participants, reflected in extra time required for task completion, was determining the meaning of labels for the various types of matches. This problem is evident in the sample subtasks b and d, for which participant time exceeded the acceptable level. We knew the labels used for the prototype were ambiguous, but we had been unable to reach consensus on appropriate wording. One part of the formative usability evaluation process was interviewing participants about what they considered a "best match" for different kinds of queries and how they thought results should be ranked. Their answers guided us in revising the wording for the match-type labels and corresponding search system behavior. For the Words in Title field, word ordering was more important to participants than expected. A minor change in response to participant feedback was use of a larger font for some example and prompt text. One control button (for "Clear Fields") was eliminated because it was used inappropriately by most of our participants, with consequences that reduced usability of the Query History. The "Clear Fields" function was moved to a menu, to prevent confusion with "New Query." The field labeled "Special Query," originally labeled "Other," was renamed based on input from participants. Minor changes to the layout were also made, such as increasing spacing between the query fields and the "Match Between Fields" buttons. This last change reflects difficulty in sample subtask e, as participants took longer than expected to locate the control buttons and therefore exceeded our maximum acceptable time.

6. Future Work

Participants evaluating MARIAN and the Envision search results display have told us they want access to the query when they look at the search results. They want to be able to revise or refine the query, in addition to examining it in relation to the results. Work is in progress to develop a smaller version of the Query Window for simultaneous display with the results windows. Users can toggle between the full-size window, with the complete set of features described above, and the smaller window, which will provide only basic functionality. With the Query Window we also plan to provide browser access to the list of authors whose works are in the Envision database and to the Computing Review categories list, which is Envision's closest approximation to a list of subject headings.

7. Acknowledgements

This work is funded by the National Science Foundation under Grant IRI-9116991, Dr. Maria Zemankova, Monitor. The project is also supported by the ACM and by Virginia Tech. The authors gratefully acknowledge contributions of Dr. Edward A. Fox, Dr. Lenwood S. Heath, Dennis J. Brueni, William C. Wake, Robert K. France, Kevin A. Mayo, and the Human-Computer Interaction Research Group at Virginia Tech, as well as our interviewees and usability evaluation participants.

References

1. N. Belkin, B. Croft: Retrieval Techniques. In: M. Williams (ed.): Annual Review of Information Science and Technology. New York: Elsevier Science Publishers 1987, pp.109-145

2. D.J. Brueni, E.A. Fox, L.S. Heath, D. Hix, L.T. Nowell, W.C. Wake: What If There Were Desktop Access to the Computer Science Literature? In: S.C. Kwasny (ed.): Proceedings of the 21st Annual ACM Computer Science Conference: CSC '93, Indianapolis, IN, USA, February, 1993, pp. 15-22

3. E.A. Fox, R.K. France, E. Sahle, A. Daoud, B.E. Cline: Development of a Modern OPAC: From REVTOLC to MARIAN. To appear in: Proceedings of the 16th International Conference on R & D in Information Retrieval, SIGIR '93, Pittsburgh, PA, USA, June 1993

4. E. Fox, S. Betrabet, S.; M. Koushik, W. Lee: Extended Boolean Models. In: W.B. Frakes, R. Baeza-Yates (eds.): Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice-Hall 1992, pp. 393-418

5. W.B. Frakes: Introduction to Information Storage and Retrieval Systems. In: W.B. Frakes, R. Baeza-Yates (eds.): Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice-Hall 1992, pp. 1-12

6. G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais: "The Vocabulary Problem in Human-System Communication. In: Communications of the ACM 30(11), 964-971 (Nov. 1987)

7. R.R. Larson: Evaluation of Advanced Retrieval Techniques in an Experimental Online Catalog. In: Journal of the American Society for Information Science 43(1), 34-53 (1992)

8. L.T. Nowell, D. Hix: Visualizing Search Results: User Interface Development for the Project Envision Database of Computer Science Literature. To appear in: Proceedings of HCI International '93, 5th International Conference on Human Computer Interaction jointly with 9th Symposium on Human Interface (Japan), Orlando, Florida, USA, August 8-13, 1993.

9. G. Salton: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. New York: Addison-Wesley 1989

10. G. Salton: Another Look at Automatic Text-Retrieval Systems. In: Communications of the ACM 29(7), 648-656 (July 1986)

11. G. Salton, M.J. McGill: Introduction to Modern Information Retrieval. New York: McGraw-Hill Book Company 1983

12. G. Salton, E.A. Fox, H. Wu: Extended Boolean Information Retrieval. In: Communications of the ACM 26(11), 1022-1036 (Nov. 1983)

13. G. Salton, A. Wong, C.S. Yang: A Vector Space Model for Automatic Indexing. In: Communications of the ACM 18(11), 613-620 (Nov. 1975)

14. S. Wartik: Boolean Operations. In W.B. Frakes, R. Baeza-Yates (eds.): Information Retrieval: Data Structures and Algorithms. Englewood Cliffs, NJ: Prentice-Hall, 1992, pp. 264-292