COMPOSITE DOCUMENT EXTENDED RETRIEVAL

An Overview

Edward A. Fox
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

1. Introduction

This paper provides a partial overview of practice, problems, proposals, and plans relating to the handling of 'composite documents' by extended information storage and retrieval systems. It aims to describe such documents, to explore various areas of application for them, to portray a number of representative test collections under development, to survey related studies as well as previous work of this author, and to examine plans for further investigation.

1.1. Background

Experimental information retrieval (IR) systems, some dating back to the sixties, have demonstrated the viability of fully automatic document storage and retrieval methodologies with small to medium size bibliographic collections [72]. Many of these experimental systems utilize the vector space model in which each important term (such as a word stem) identifies a different dimension in a space, so that matrix methods and vector operations can be defined on queries and documents. Statistical techniques have been very effective, and probabilistic enhancements have given additional improvements [84]. However, the basic vector space model is oriented towards recording the essential information in the text of a title/abstract combination rather than describing more complex document structures. It is necessary to extend the model in order to handle composite documents.

On the other hand, commonly available retrieval systems that employ Boolean logic queries and utilize inverted file storage schemes can without modification accommodate such documents, albeit with somewhat less effectiveness than is possible with more sophisticated systems. Hence, it is also of interest to consider how Boolean logic systems can be extended to give better performance, especially with composite documents, and to integrate those approaches with vector methods.

1.2. Representations

Retrieval systems work with representations of user queries as well as stored documents. Sparck Jones, discussing the application of natural language understanding techniques to better characterize users' information needs, points out that in recent years studying queries rather than documents has been the most viable research approach [81]. However, there are limits to how effective retrieval can be when one uses simple vectors to represent documents, even when one employs elaborate probabilistic models of the dependencies between query terms, along with involved mathematical developments [87].

Croft, in discussing experiments with representations, has shown that simple within document frequency information can be used to improve the performance of probabilistic query processing methods [21]. Further, in a recent survey of representation methods in information retrieval (IR), Smith and Warner point out the importance of studying documents, especially ones they refer to as complex objects [79].

1.3. Composite Documents

This author believes that it is vital to more effectively model composite documents, and that

This work was supported by the NSF under grant IST-8418877 and by Virginia's Center for Innovative Technology under grant INF-85-016.

there are at least two important aspects to consider. First, items to be indexed often possess some type of internal structure or syntax which is essential for proper interpretation. At one extreme there are entire books, with all of their complex and overlapping modes of organization. On the other hand, there are short to medium length text items such as electronic mail messages, which are a mixture of formated entries and free text. These messages can often be broken down into a 'header' and a 'body,' where the header contains a number of 'fields.'

Secondly, there is the matter that different types of data are found in such documents. There are words, dates, times, tables of numbers, abbreviations of organizations and geographic locations, proper names, and a variety of other constructs. Often, too, there are explicit references to other documents (eg., by message identifier) in the formated fields, and implicit references (eg., by author and date) in the textual portion. One thus has three basic forms of data to contend with: text, formated data, and document interrelationships, as well as a large number of specialized data types.

1.4. Applications

In recent years many researchers have realized the enormous value of utilizing IR methods in a wide variety of new application areas. Many of those situations involve composite documents.

1.4.1. Bibliographic Retrieval

In the realm of bibliographic retrieval, where both commercially available and experimental IR systems have been in use for decades, each record present is composed of various fields. Katzer et al. note rather small overlap in the relevant documents retrieved when different combinations of those fields were considered during searching [47]. That fact, which suggests that further study of the interrelationships between fields is needed, is further complicated by the considerable variation regarding **which** fields are included in typical databases. In the related domain of providing online public access catalogs for libraries, standardization has taken place, such as in the employment of MARC records, but in that case there are so many fields that occasional users have trouble remembering and making optimal use of them. As bibliographic retrieval systems and software are used more and more for free text searching of existing and new databases, the importance of effectively handling complex document structures will increase even further.

1.4.2. Office Information Systems

In offices, reports and other types of composite documents are commonplace. Many office automation applications are possible for IR methods even when simple models are employed [22]. Early experimentation with a small collection of correspondence demonstrated that document retrieval techniques could be applied to business letters and that utilizing links between consecutive letters improved system effectiveness [62]. More extensive use of document retrieval for office problems has been proposed and preliminary experiments have shown that such methods are appropriate [19,20].

In large organizations, office information systems are interconnected and provide a wide range of capabilities. One such rather sophisticated distributed system is Xerox's Grapevine, which by 1981 was servicing 1500 individuals and 500 groups. Electronic mail was already one of the principal applications; 2500 messages were sent daily, each one to an average of four recipients. In addition to providing for retrieval of messages, the designers had to deal with issues of naming entities and of accessing databases to locate proper addresses [7].

1.4.3. Computer Networks

On an even larger scale, computer networks like the ARPAnet have such large volumes of electronic mail being transmitted that standards are imperative [18]. The United States National Bureau of Standards has published a standard on message formats [61], and in 1984 global standards were drafted [60]. With such standards in place it will be worthwhile to develop mail handlers with even more effective retrieval capabilities than RdMail [52] or Hermes [58], which provide fairly complete message selection and search commands. Retrieval techniques could also be beneficially applied to computer conferencing systems [23] or bulletin boards.

In each of these situations, there is the added problem of keeping track of the syntax and content of names, addresses, or routes for message recipients. Special programs have been developed to help properly format these message components [2]. Standards have been proposed for naming and addressing [8], and the development of 'user friendly names' is an important research problem [44]. Name servers are needed to provide directory assistance for names in networks [75], and a rather interesting name server (NS) has been developed [80] for the Computer Science Network (CSNET) [15]. Both electronic messages and entries recorded in name servers can be viewed as types of composite documents.

1.4.4. Knowledge Bases

A good deal of research in artificial intelligence (AI) has focused on the representation of knowledge in a variety of domains. This author believes that some of the techniques developed can be of value in studying composite documents and that some of the lessons learned from modeling such documents may impact various types of AI investigations. One example is in the area of expert systems [42]. A project is underway at Virginia Tech to develop an expert system for placement of children in foster care settings. Automatic processing of descriptions of children and homes, and techniques for ranking those homes that might be most suitable for a child, are some of the tasks which the system must perform and which will benefit from techniques applicable to composite documents.

Another example of working with a knowledge base is that of finding interesting passages in books. University students studying textbooks, researchers accessing reference volumes, and individuals in many other walks of life frequently wish to locate something they once read, or to find a passage relating to a current interest. ' Collections of books are one of the most important knowledge bases now in existence, and representing such composite documents is thus a further area of application for IR methods.

2. Test Collections

Experimental information retrieval studies require realistic test collections, so that the value of models and techniques can be demonstrated. In order to validate proposed models of composite documents and to confirm the effectiveness of extended vector and Boolean methods, four types of collections have been prepared previously or are now being prepared at Virginia Tech.

The first type is that of bibliographic records. Two collections of this type were specially prepared for testing of extended vector representations [34]. In 'ISI 1460' there are three types of information: author names, terms in titles and abstracts, and counts of cocitations [76]. This set of 1460 documents in information science was selected based on cocitation data provided by the Institute for Scientific Information and includes many of the most important journal articles and manuscripts published between 1969 and 1977. A total of 76 queries were formulated and exhaustive judgments of relevance of each document to each query were made. The second collection, 'CACM 3204,' contains author names, terms in title/abstract/keyword sections, classification codes from the CR category system [27], publication and date specifics, cocitations, bibliographic coupling counts [48], and indications of citation links between articles. The articles are from *Communications of the ACM* and include all papers published until the end of 1979. Fifty two queries were obtained and a large number of searches were carried out to approximate the ideal of having exhaustive relevance decisions.

The second type of collection is of books. The full text of *Introduction to Modern Information Retrieval* (IMIR) by Salton and McGill has been made available and 42 queries and relevance judgments were prepared in connection with pages in the first five chapters. The full text of a number of books published by the Baha'i Publishing Trust are also available. For the book *Gleanings from the Writings of Baha'u'llah* (GWB) a set of 53 queries and exhaustive relevance information have been gathered. In addition to these two books, IMIR and GWB, several computer manuals have been provided by Digital Equipment Corporation, but additional queries are needed for them to be usable in experimental studies. Structural units and data items present in these several book-based collections include: paragraphs, pages, sections, chapters, figures, tables, photos, captions, references between passages, and implicit links (such as of adjacency between paragraphs or 'anchoring' of captions to photos).

The third type of collection is of electronic mail. The principle file that will be utilized is of digests distributed to members of the AILIST mailing list maintained for users of the ARPAnet. New issues appear roughly every other day, and all files since the first one released in Spring 1983 have been recorded. Each digest has an overall header, with standard fields as well as a table of contents for the individual messages. The messages themselves have a header made up of several fields, and a body that sometimes includes address and closing comments about the recipient. Messages may be requests for information, calls for papers, seminar announcements, bibliographies, commentaries on previous messages, opinion statements, etc. Queries for this collection, along with relevance judgments, will be solicited from students at Virginia Tech who may submit profiles as well as retrospective need statements.

The fourth and final type of collection is of data from the CSNET name server database. This file contains descriptions submitted by computer researchers at institutions that have joined the computer science network. There will be more than 3000 profiles of individuals, each one with full name, institution, mail and network address, telephone, and a free text field to describe a person with whatever words or phrases are deemed useful for subsequent searchers. Queries and relevance judgments will be solicited at Virginia Tech.

These various collections cover a wide spectrum of application areas. Thus, if proposed methods for handling composite documents show promise on a number of them, then fairly convincing proof of the utility of those techniques will exist.

3. Related and Previous Work

A good deal of work has been carried out in areas relating to the study of composite documents, and some preliminary studies have been carried out by this author in recent years. The sections below focus on the most important investigations, and briefly summarize the findings.

3.1. Passage Retrieval

Searching through books is probably one of the least thoroughly studied areas under consideration. O'Connor has investigated the location of answer-passages in scientific journal articles in what is probably the most closely related work [63]. Fairly good performance was obtained in rather small scale tests of heuristic methods, which were based on linguistic considerations.

This author has undertaken several preliminary studies in this area. One experiment attempted to determine if retrieval of sentences, paragraphs, or other discourse items would be most appropriate, and the results indicated that paragraphs seemed usually to be the correct unit. That experiment demonstrated also that vector processing methods could be utilized to effectively process non-technical works such as some of the Sacred Writings of the Baha'i Faith [20].

Further work, with the the GWB collection, has lead to more definitive results. This 345 page book of religious prose has 707 paragraphs and 165 sections. For each of the 52 queries, all paragraphs were considered for possible relevance. Sections were considered relevant if they contained a relevant paragraph. Two representations were produced by automatic indexing methods, one where a vector represented a paragraph and the other where an entire section was represented. Using vector processing methods and evaluation techniques built into a modern version of the SMART system [33], the following recall-precision chart was obtained. The comparison of paragraph versus section retrieval results suggests that sections might be what an average user would like to see.

Recall	Average Precision Using:		%Improvement
Level	Paragraphs	Sections	Sect. vs. Parag.
0.0	.3724	.4333	16.4
0.1	.3515	.3947	12.3
0.2	.3181	.3434	8.0
0.3	.2607	.2936	12.6
0.4	.2440	.2412	-1.1
0.5	.2165	.2220	2.5
0.6	.1489	.1834	23.2
0.7	.1307	.1560	19.4
0.8	.1104	.1374	24.5
0.9	.1054	.1305	23.8
1.0	.1022	.1269	24.2

Clearly, IR methods do retrieve some portion of relevant items from GWB, but precision is lower than is often the case for technical literature. Further study with this and other book collections is currently underway. It is expected that when more comprehensive use is made of the structure of the book, and of other data forms such as cocitations, that performance will significantly improve.

3.2. Database Management

Since certain components of documents have a particular format and can be assigned a specific data type, utilization of popular database models for IR has been considered. The FIRST system at Xerox employed a network type database to supplement vector storage software [24]. Crawford suggested using the relational model [17]. More recently, extensions to the INGRES relational system were proposed so as to facilitate document processing [82].

Various proposals have been advanced to integrate information retrieval and database models [6]. A fairly recent survey of this area appears in [55]; many ideas but few cohesive theories and practical systems have appeared to date.

3.3. Document Models

Typically in connection with office information system projects, however, there has been real progress toward implementing systems that are based on more comprehensive models of documents. CALIBAN aims at providing a two dimensional representation of documents using bit mapped graphics [36]. Documents can be modeled as trees, possibly extended with connections, or as networks, and weighted queries can be processed in connection with searching and ranking for retrieval [4].

The Officeaid project aims at integrating document handling, using the MISTRESS relational system, and handles fields within documents as well as various document versions such as galleys [53]. The OTTER system is based on the assumption that queries in an office automation system might refer to a variety of types of data, such as: terms, objects (i.e., triples specifying a type, a comparison, and a desired value), collections of previous identified documents, or concepts formulated earlier (stored as prior queries) [71].

Kimura and Shaw deal with abstract document objects, extending simple tree models to handle references, and defining several kinds of sequence on children of nodes [49]. Standards efforts have lead to the Office Document Architecture (ODA) which specifically deals with 'composite objects,' though retrieval applications are not emphasized [43]. Publishers, sponsoring the Electronic Manuscript Project, are very interested in reducing the proliferation of incompatible document formats, and so it is plausible that a certain degree of agreement will result in describing such documents [45]. But in addition to having models for documents, it is necessary to be able to analyze documents according to those models and to represent the contents in a suitable form in order for retrieval to effectively take place.

3.4. Natural Language Analysis

A variety of knowledge representation schemes have been proposed for recording the results of natural language analysis. Many of those can be viewed as variants of semantic networks [67]. A group at Olivetti employs semantic memory, built up in semi-automatic fashion, to record concepts about office documents so that subsequent query processing can perform simple inferencing [3]. Frames [57] may be viewed as a type of semantic net; high level representations using them might be particularly suitable for documents. The FRUMP system, for example, tries to 'skim' newspaper stories, and can achieve reasonable speed since it presumes stories fit into a relatively small number of 'scripts' [26]. Choosing what frame to consider is a difficult problem in general [12], demanding sophisticated context recognition [13].

In order to construct one of these representations it is necessary to analyze each document. Most work in natural language analysis has been based on syntactic recognition [86]; processing one sentence may require several seconds. Further difficulties result from the occurrence of grammatical and spelling errors, which are very common in unpublished communications [11]. Flexible parsing methods, often based on pattern matching, are of value in these situations [41]. Word expert parsers [77] seem particularly suitable; the TOPIC system employs one to condense information from article abstracts into frames [39]. Thus, gradual progress is being made toward having realistic language comprehension capabilities which are fast, flexible, and robust [66].

The recent trend toward using logic for solving a variety of problems [51], especially in connection with natural language analysis [65], has lead to systems for analyzing and representing stories [16]. Based on models of story structure [70] these efforts have lead to efforts for representing more complex text items [74]. Prolog has already been shown to simplify modeling of database systems [54] and has been recommended for file search functions [28]. At Virginia Tech, a version of Prolog called HC has undergone substantial development [69] and has been used to construct one type of pattern parser [68]. This effort by Roach et al. has also yielded a general purpose expert system shell. These tools will be of use in future work of this author on analyzing and accessing composite documents and on applying IR methods to items in knowledge bases of expert systems.

3.5. Multiple Concept Types

In order to employ the vector space model with composite documents, certain extensions have been proposed and tested by this author. The vector space model usually views documents as being made up of say T terms, so that a T dimensional vector space is obtained. The idea behind multiple concepts types as proposed by this investigator is to have separate subvectors, one for each type of concept [32]. For the CACM test collection of 3204 documents there were seven concept types, as mentioned earlier [34]. Dictionaries could be modeled by relations of form

(concept-type, concept-no., concept-value)

and document vectors as

(document-no., concept-type, concept-no.,
concept-frequency-in-doc.).

Using this method, content terms such as word stems, direct or indirect links of various sorts between documents, and facts or other forms of data can be stored according to a uniform scheme. In addition to the problems of analyzing and representing composite documents, it is also necessary to devise suitable matching and similarity functions for the various types of data.

3.6. Matching and Similarity Functions

A great deal of attention has been given to the matter of devising proper similarity functions for information retrieval systems [84]. One study compared a large number of these and grouped them into various classes that seemed to behave similarly [56]. However, in all of these cases the presumption is that matching items are being considered (eg., identical word stems in queries and documents).

Composite documents, however, contain various data items such as time or date indicators, or various types of strings, where inexact matching of items is necessary. The RESEDA system, for example, utilizes a number of representations of temporal data, all based on time intervals along with modifying facts [89]. Complex inference rules are necessary to identify entries matching a temporal description, which might cover a range of time, be stated in uncertain terms, or be specified as relative to another event.

Matching of strings is also a difficult problem [40]. Early efforts focused on computing string similarities [1]; an important example is that of handling names. In 1962 Davidson considered retrieving names from an airline reservation database when given possibly misspelled query information [25]. Further experiments were done by Greenfield [38] and Joseph & Wong [46]. Since inexact matching of names is important for many types of composite documents, this author supervised yet another study.

Using 3081 last names of authors from collections of articles in computer and information science as a base, 350 names were chosen at random and divided into 10 sets of 35. Ten subjects were chosen, and a set assigned to each. The names in a set were read aloud, slowly, with one repetition allowed (on request), and the subject listed as many versions as could be thought of for spelling a name. Using this data, an evaluation of name matching algorithms was conducted. For a total of 744 spelling versions, a search was made in the author file using each algorithm of interest to determine if the original author name would be retrieved. For a given retrieved set, recall and precision were computed and then overall averages were calculated.

The methods considered were that of Davidson (D62), Davidson's algorithm revised to conflate letters 'n' and 'm' (DNM), Soundex [38] phonetic scheme (SOU), and several versions of the Levenstein distance function [40] adapted for retrieval. The Levenstein function computes distances between two strings based on number of changes required to transform one to the other. The various test versions (L02,L03,L10) indicate how many retrieved items were examined and thus specify cutoff levels (eg., 2, 3, or 10).

Name of	Averages over 744 searches of:		
Method	No. Ret.	Recall	Precision
D62	2.1	.74	.35
DNM	2.2	.74	.33
SOU	3.8	.81	.21
L02	2.0	.76	.38
L03	3.0	.80	.27
L10	10.0	.87	.09

Results indicate that the DNM scheme is actually slightly worse than D62, and that the Davidson approach yields lower recall but higher precision than Soundex, since retrieved sets for the two are around two and four names, respectively. The Levenstein runs are included for clarification and comparison only since they are computationally so demanding but do indicate that fairly high levels of recall are possible if enough names are considered. Further work should aim at combining the speed possible with schemes such as Soundex and the ranking capability of the Levenstein distance function.

Clearly, a number of similarity functions are needed to handle the various types of data present in composite documents. Even for words, there are varying degrees of similarity involved with matches of synonyms or other lexically related entries [35]. With suitably defined abstract data types, improved retrieval and conceptual clarity should result [29]. However, for retrieval and ranking of composite documents it is also necessary to combine the similarity values of individual data items into an overall query-document similarity.

3.7. Retrieval of Composite Documents

Very little work has been done in the area of developing comprehensive schemes for retrieving composite documents. Most of the work has been experimental. Katzer et al. noted that Boolean logic searches using different portions of a document representation gave retrieved sets with small overlap, but did not develop a method for integrating such searches [47]. Bichteler & Eaton compared two cases: having both cocitation and bibliographic coupling counts available, or just utilizing one count, and found the former better but did not generalize their idea [5]. Fuhr & Knorz used Boolean and probabilistic methods for combining different types of data in order to improve automatic classification methods but had limited success [37].

A theoretical study including a sophisticated model, which divides vector information into characteristic and term spaces, seems to have good potential but needs refinement and experimental validation [50]. The characteristic space allows recording of variations in types of data. The topologically motivated discussion of the several spaces does aim at providing a more comprehensive document representation.

This author has studied the value of the above mentioned multiple concept type approach to vector representation, through experimentation with several test collections. Clustering of such documents (to group similar documents close to each other) using the algorithm of Williamson performed reasonably well. Feedback was chosen, however, as the method most suitable for determining whether extended vectors give better retrieval performance than do vectors made up only of terms. Regression techniques were used to determine the best way to combine together the separate similarities computed for each subvector. Relevance feedback tests on two collections showed improvements in effectiveness of 10-30% due to regression based weighting of subvectors [31,32]. Further testing of this model on other collections of composite documents will indicate whether the representation scheme gives similar results on different types of data, in a predictive role as well.

3.8. Soft Boolean Evaluation

Considering retrieval of composite documents using Boolean logic queries, one is forced to deal with inexact matching and similarity functions that provide continuous rather than binary values. Proponents of fuzzy set theory have elaborated upon the early ideas enunciated by Zadeh [88] regarding continuous rather than Boolean membership functions. Numerous proposals have been made for applying fuzzy logic to document retrieval systems. Early work focused on mathematical criteria and basic definitions for weighted Boolean expressions [9]. Smeaton has encouraged construction of fuzzy sets of search terms to aid in feedback operations [78].

Paice, in a recent discussion of what he refers to as 'soft evaluation of Boolean search queries,' argues for the use of operators that are less 'strict' than *OR* and *AND* [64]. He indicates that performance is better than with standard Boolean queries, but due to limitations in his testing apparatus is unable to select which definition of the two operators is more effective.

The RUBRIC system has employed a rule-based AI type approach to explore basic issues of uncertainty in indexing, and related problems of query formulation when inexact matching will take place [83]. The concern with fuzzy set theory and other logics, and with norms and ranking of documents by similarity to queries, makes RUBRIC especially relevant to work with Boolean searching of composite documents. It is limited in terms of the types of data handled and the range of applications considered, however. Furthermore, the experimental design chosen excludes the possibility of reaching firm conclusions even on the test collection reported. The success of RUBRIC to date suggests that related work from the perspective of document retrieval should be vigorously pursued.

One scheme that has been extensively studied, and which should be of particular value for handling composite documents, is the 'p-norm' method.

3.9. P-Norm Formalism

Building upon fuzzy set theory and early attempts to apply it to IR, Wu introduced the p-norm notation, with good theoretical properties and potential value for understanding and extending both Boolean logic and vector query based search systems [73]. Its theory was further explained, and practical usage demonstrated, by this researcher [32].

The basic idea is to use distance measures to define the Boolean operations *AND* and *OR*. Consider, for example, terms A and B that are used to index a document. Assume that fuzzy set membership functions are used for indexing, so that if term A occurs only a few times in the document, while term B occurs many times, then the degree of indexing by terms A and B might be 0.3 and 0.8, respectively. Evaluating 'A *OR* B' should measure how far that document is from having neither term present, while 'A *AND* B' should yield a higher value if a document is closer to having both terms fully present (i.e., high membership function values).

When the L_p family of norms is chosen, it turns out that for one extreme case, namely $p=1$, vector inner product similarity values are obtained, while for the other extreme case, namely $p=\infty$, strict Boolean (i.e., fuzzy set theory) similarity results. The *AND* and *OR* operators can therefore be parameterized by the p-value. Further extensions are possible whereby terms in a Boolean expression can be assigned relative weights. Thus,

$$<A,2>\ AND^{1.5}\ <B,1>$$

indicates that the conjunction of terms A and B is desired, where *AND* is given a p-norm interpretations with $p=1.5$, and term A should be viewed as being twice as important as term B.

Using standard document retrieval evaluation measures including recall and precision [84], the effectiveness of standard Boolean queries, Boolean queries with a p-norm interpretation, and vector queries were considered. P-norm interpretation was consistently better than the standard Boolean one, and when good Boolean queries were initially present the p-norm interpretation results were better than simple vector similarity methods.

Since p-norm queries have many parameters, and since users have trouble enough constructing good Boolean queries, a method for automatically forming Boolean or p-norm queries from natural language statements was sought. By using frequency information about terms, it was possible to automatically produce p-norm queries which performed reasonably well in most of the collections. To improve upon these results, feedback techniques were adapted. A straightforward scheme for automatically improving Boolean queries by considering user supplied relevance data was devised and shown to be effective. When p-norm interpretation of these feedback queries was employed, very large improvements in the effectiveness of retrieval were observed [32].

It is expected that p-norm or similar approaches to 'soft' Boolean evaluation can be effectively incorporated into composite document retrieval systems.

4. Planned Work

This author is involved in three principal efforts relating to the development of improved information storage and retrieval methods for handling composite documents. The first is a small scale effort, aimed at developing an expert system for placement of children in foster care settings. Based on the software available as part of the Virginia Tech Prolog project, this system will aim at representing homes and children using frames and other knowledge structures. Similarity functions will be defined for low level data items, and overall matching values will be determined using a rule based approach. The results of matching such composite structures should facilitate a ranking of homes in terms of expected suitability for a particular child.

A second effort deals specifically with composite document extended retrieval. Work on soft Boolean evaluation schemes including the p-norm formulations, retrieval of composite documents, inexact matching and similarity functions, natural language analysis, document modeling, and message retrieval will be emphasized. The principal test vehicle for experiments in document analysis and retrieval will be the AILIST message collection, though other collections described earlier will be used as well.

The third effort, being carried out in collaboration with Matthew Koll, aims at further developing composite document models and at extending them to be applicable to needs in the public and private sectors as well. Attention is given to the unfulfilled retrieval needs of industrial and other organizations [85]. Statistical and rule based approaches will be compared. A variety of test collections will be utilized. In particular, the CACM, CSNET name server, and book passage collections will aid initial investigations. Later, corporate documentation files can be modeled and used for tests. Library and bibliographic systems will also be studied employing the intelligent front end approach [59]. Algorithms for obtaining ranked retrieval results in such situations will have to be specially adapted to allow incorporation of soft Boolean evaluation techniques [10].

It is hoped that better modeling of composite documents will result from these studies, and that both vector and Boolean query approaches will provide more effective retrieval performance. Development of these methodologies will be attempted so that results will also be applicable to needs in the public and private sectors.

5. Acknowledgements

Funding for previous and planned work have come from the Virginia Tech College of Arts and Sciences, from the IST division of NSF, and from Virginia's CIT/IIT. Collections have been provided by Henry Small of ISI, Robert Dattola of Xerox, Richard Edmiston of CSNET CIC, Kenneth Laws of SRI International, Lee Nelson (with permission of the Baha'i Publishing Trust), and Gerard Salton (with permission of McGraw-Hill). A number of Virginia Tech students have carried out or are now involved in research relating to this project, including Allen Maxwell Jr., John Imholz, Sheila Winett, Robert France, and David Joslin. Thanks also go to the SMART project team at Cornell for collaboration in software development and collection preparation, and to other individuals who contributed queries and relevance judgments. Special thanks go to my wife who helped with a number of collections.

6. References

[1] Alberga, C.N., String Similarity and Misspellings. *Commun. ACM*, 10(6), June 1967, 302-313.

[2] Allman, E., SENDMAIL - An Internetwork Mail Router. In *UNIX Programmer's Manual*, Berkeley Release 4.2, 1983.

[3] Barbi, E., Calvo, F., Perale, C., Sirovich, F. and Turini, F., A Conceptual Approach to Document Retrieval. *Proc. of the Second ACM-SIGOA Conference on Office Information Systems-June 25-27, 1984*, 219-226.

[4] Bartschi, M. and Frei, H.P., Adapting a Data Organization to the Structure of Stored Information. In *Research and Development in Information Retrieval, Proc., Berlin, May 18-20, 1982*, ed. by Gerard Salton and Hans-Jochen Schneider, Springer-Verlag, Berlin, 1983, 62-79.

[5] Bichteler, J. and Eaton III, E.A., The Combined Use of Bibliographic Coupling and Cocitation for Document Retrieval. *J. Am. Soc. Inf. Sci.*, 31(4), July 1980.

[6] Biller, H., On the Architecture of a System Integrating Data Base Management and Information Retrieval. In *Research and Development in Information Retrieval, Proc., Berlin, May 18-20, 1982*, ed. by Gerard Salton and Hans-Jochen Schneider, Springer-Verlag, Berlin, 1983, 80-97.

[7] Birrell, A.D., Levin, R., Needham, R.M. and Schroeder, M.D., Grapevine: An Exercise in Distributed Computing. *Commun. ACM*, 25(4), April 1982, 260-274.

[8] Bolt Beranek, and Newman, Inc., Naming and Addressing in Computer Based Message Systems. Draft Report No. ICST/CBOS-82-4, Dept. of Commerce, National Bureau of Standards, Aug. 1982.

[9] Bookstein, A., Fuzzy Requests: An Approach to Weighted Boolean Searches. *J. Am. Soc. Inf. Sci.*, 31(4), July 1980, 240-247.

[10] Bovey, J.D. and Robertson, S.E., An Algorithm for Weighted Searching on a Boolean System. *Inf. Tech.: Res. Dev. Applications*, 3(2), April 1984, 84-87.

[11] Carbonell, J.G. and Hayes, P.J., Coping with Extragrammaticality. *Proc. of Coling84, 2-6 July 1984*, Stanford Univ., 437-443.

[12] Charniak, E., With Spoon in Hand this Must Be the Eating Frame. In *TINLAP-2: Theoretical Issues in Natural Language Processing-2*, ed. by David L. Walz, ACM, 1978, 187-193.

[13] Charniak, E., Context Recognition in Language Comprehension. In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 1982, 435-454.

[14] Chupin, J.C. and Joloboff, V., A Data Model for Office Systems. In *Office Information. Systems*, ed. by N. Naffah, North-Holland, Amsterdam, 1982, 39-56.

[15] Comer, D., The Computer Science Research Network CSNET: A History and Status Report. *Commun. ACM*, 26(10), Oct. 1983, 747-753.

[16] Correira, A., Computing Story Trees. *Amer. J. Comp. Ling.*, 6(3-4), 1980, 135-149.

[17] Crawford, R.G., The Relational Model in Information Retrieval. *J. Am. Soc. Inf. Sci.*, 32(1), 1981, 51-64.

[18] Crocker, D.H., Standard for the Format of ARPA Internet Text Messages. RFC 822, ARPANET Networking Group, Aug. 1982.

[19] Croft, W.B., Experiments with Automatic Text Filing and Retrieval in the Office Environment. *ACM SIGIR Forum*, 16(4), Spring 1982, 2-9.

[20] Croft, W.B. and Pezarro, M.T., Text Retrieval Techniques for the Automated Office. In *Office Information Systems*, ed. by N. Naffah, North-Holland, Amsterdam, 1982, 565-576.

[21] Croft, W.B., Experiments with Representation in a Document Retrieval System. *Inf. Tech.: Res. Dev. Applications*, 2(1), Jan. 1983, 1-22.

[22] Croft, W.B., Applications for Information Retrieval Techniques in the Office. *ACM SIGIR Forum and Proc. 6th Annual Int. ACM SIGIR Conf. on R&D in IR*, 17(4), Summer 1983, 18-23.

[23] Daney, C., The VMSHARE Computer Conferencing Facility. In *Computer Message Systems*, ed. by Ronald P. Uhlig, North-Holland, Amsterdam, 1982, 115-127.

[24] Dattola, R.T., FIRST: Flexible Information Retrieval System for Text. *J. Am. Soc. Inf. Sci.*, 30(1), 1979, 9-14.

[25] Davidson, L., Retrieval of Misspelled Names in an Airlines Passenger Record System. *Commun. ACM*, 5(5), May 1962, 169-171.

[26] DeJong, G., An Overview of the FRUMP System. In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 1982, 149-176.

[27] Denning, D.E., et al., The Proposed New Computing Reviews Classification Scheme: A Report of the Computing Reviews Category Revision Committee. *Commun. ACM*, 24(7), July 1981, 419-433.

[28] Eastman, C.M., File Searching Problems in Logic Programming Systems. Technical Report 83-CSE-8, Dept. of Comp. Sci. and Eng., Southern Methodist Univ., Feb. 1983.

[29] Fox, E.A., Implementing SMART for Minicomputers Via Relational Processing with Abstract Data Types. *Joint Proc. of SIGSMALL Symposium on Small Systems and SIGMOD Workshop on Small Data Base Systems, ACM SIGSMALL Newsletter*, 7(2), Oct. 1981, 119-129.

[30] Fox, E.A., Automatic Document and Passage Retrieval Methods: Aids to Searching the Baha'i' Writings. *Proc. Annual Meeting Assoc. for Baha'i Studies*, April 1981.

[31] Fox, E.A., Combining Information in an Extended Automatic Information Retrieval System for Agriculture. *Infrastructure of an Information Society (Proc. 1st Int. Information Conf. Egypt, 13-16 Dec. 1982)*, 1983.

[32] Fox, E.A., Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. Dissertation, Cornell Univ., Aug. 1983.

[33] Fox, E.A., Some Considerations for Implementing the SMART Information Retrieval System under UNIX. TR 83-560, Cornell Univ., Dept. of Comp. Sci., Sept. 1983.

[34] Fox, E.A., Characterization of Two New Experimental Collections in Computer and Information Science Containing Textual and Bibliographic Concepts. TR 83-561, Cornell Univ., Dept. of Comp. Sci., Sept. 1983.

[35] Fox, E.A., Improved Retrieval Using a Relational Thesaurus for Automatic Expansion of Boolean Logic Queries. *Proc. Workshop on Relational Models of the Lexicon, Stanford Univ.*, June 29, 1984.

[36] Frei, H.P. and Jauslin, J.-F., Two-Dimensional Representation of Information Retrieval Services. In *Representation and Exchange of Knowledge as a Basis of Information Processes*, ed. by Hans J. Dietschmann, North-Holland, New York, 1984, 383-396.

[37] Fuhr, N. and Knorz, G., Retrieval Test Evaluation of a Rule Based Automatic Indexing. DV II 84-2, Technische Hochschule Darmstadt, 1984.

[38] Greenfield, R.H., An Experiment to Measure the Performance of Phonetic Key Compression Retrieval Schemes. *Meth. Inform. Med.*, 16, 1977, 230-233.

[39] Hahn, U. and Reimer, U., Heuristic Text Parsing in 'Topic': Methodological Issues in a Knowledge-based Text Condensation System. In *Representation and Exchange of Knowledge as a Basis of Information Processes*, ed. by Hans J. Dietschmann, North-Holland, New York, 1984, 143-163.

[40] Hall, P.A.V. and Dowling, G.R., Approximate String Matching. *ACM Comp. Surveys*, 12(4) 1980, 381-402.

[41] Hayes, P. and Mouradian, G.V., Flexible Parsing. *Amer. J. Comp. Ling.*, 7(4), Oct.-Dec. 1981, 232-242.

[42] Hayes-Roth, F., Waterman, D.A. and Lenat, D.B., eds. *Building Expert Systems.* Addison-Wesley, Reading, MA, 1983.

[43] Horak, W. and Kronert, G., An Object-Oriented Office Document Architecture Model for Processing and Interchange of Documents. *Proc. of the Second ACM-SIGOA Conference on Office Information Systems-June 25-27, 1984*, 152-160.

[44] IFIP WG 6.5. A User-friendly Naming Convention for Use in Communication Networks. Working Paper, Version 3, IFIP WG 6.5, March 1984.

[45] Jennings, M., The Electronic Manuscript Project. *Bulletin of the Am. Soc. Inf. Sci.*, 10(3), Feb. 1984, 11-13.

[46] Joseph, D.M. and Wong, R.L., Correction of Misspellings and Typographic Errors in a Free-Text Medical English Information Storage and Retrieval System. *Meth. Inform. Med.*, 18, 1979, 238-234.

[47] Katzer, J., et al. *A Study of the Overlap Among Document Representations.* Syracuse Univ. School of Inform. Studies, 1982.

[48] Kessler, M.M., Bibliographic Coupling Between Scientific Papers. *Amer. Doc.*, 14(1), Jan. 1963, 10-25.

[49] Kimura, G.D., The Structure of Abstract Document Objects. *Proc. of the Second ACM-SIGOA Conference on Office Information Systems-June 25-27, 1984*, 161-169.

[50] Korfhage, R.R. and Chavarria-Garza, H., Retrieval Improvement by the Interaction of Queries and User Profiles. *Proc. of COMPSAC '82, Sixth International Conference on Computer Software & Applications*, Nov. 1982, 470-475.

[51] Kowalski, R., *Logic for Problem Solving.* North Holland, New York, 1980.

[52] Lamb, D.A., *RdMail Message Management System: User's Guide and Reference. 7th Ed.*. Carnegie-Mellon Univ. Comp. Sci. Dept., Pittsburgh, PA, Aug. 1982.

[53] Lee, A., Woo, C.C. and Lochovsky, F.H., Officeaid: An Integrated Document Management System. *Proc. of the Second ACM-SIGOA Conference on Office Information Systems-June 25-27, 1984*, 170-180.

[54] Li, D., *A PROLOG Database System.* Research Studies Press Ltd., John Wiley & Sons, New York, 1984.

[55] Macleod, I.A. and Crawford, R.G., Document Retrieval as a Database Application. *Inf. Tech.: Res. Dev. Applications*, 2(1), Jan. 1983, 43-60.

[56] McGill, M.J., Koll, M. and Noreault, T., *An Evaluation of Factors Affecting Document Ranking By Information Retrieval Systems.* Syracuse Univ. School of Inform. Studies, 1979.

[57] Minsky, M., A Framework for Representing Knowledge. In *The Psychology of Computer Vision*, ed. by P. Winston, McGraw-Hill, New York, 1975.

[58] Mooers, C.D., The Hermes Guide. Report No. 4995, BBN, Inc., Aug. 1982.

[59] Morrissey, J., An Intelligent Terminal for Implementing Relevance Feedback on Large Operational Retrieval Systems. In *Research and Development in Information Retrieval, Proc., Berlin, May 18-20, 1982*, ed. by Gerard Salton and Hans-Jochen Schneider, Springer-Verlag, Berlin, 1983, 38-50.

[60] Myer, T.H., Standards for Global Messaging: A Progress Report. *J. Telecommunication Networks*, 2(4), Winter 1983.

[61] National Bureau of Standards. Message Format for Computer-Based Message Systems. Federal Inf. Proc. Standards Pub. [FIPS PUB) 98, NTIS, March 1983.

[62] Nodtvedt, E., Information Retrieval in the Business Environment. TR 80-447, Cornell Univ., Dept. of Comp. Sci., 1980.

[63] O'Connor, J., Answer-Passage Retrieval by Text Searching. *J. Am. Soc. Inf. Sci.*, 31(4), 1980, 227-239.

[64] Paice, C.D., Soft Evaluation of Boolean Search Queries in Information Retrieval. *Inf. Tech.: Res. Dev. Applications*, 3(1), 1984, 33-42.

[65] Pereira, F., Logic for Natural Language Analysis. Technical Note 275, SRI International, Jan. 1983.

[66] Riesbeckh, C.K., Realistic Language Comprehension. In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 1982, 37-54.

[67] Ritchie, G.D. and Hanna, F.K., Semantic Networks - a General Definition and a Survey. *Inf. Tech.: Res. Dev. Applications*, 2(4), Oct. 1983, 187-231.

[68] Roach, J. and Savarese, J., Designing a Natural Language Interface for a Graphics Editor. Virginia Poly. Inst. and State Univ., Dept of Comp. Sci., 1982.

[69] Roach, J. and Fowler, G., The HC Manual: Virginia Tech Prolog. Technical Manual, Virginia Poly. Inst. and State Univ., Dept of Comp. Sci., 1983.

[70] Rumelhart, D.E., Notes on a Schema for Stories. In *Representation and Understanding*, ed. by D. G. Bobrow and A. Collins, Academic Press, New York, 1975, 211-236.

[71] Sacco, G.M., OTTER - An Information Retrieval System for Office Automation. *Proc. of the Second ACM-SIGOA Conference on Office Information Systems-June 25-27, 1984*, 104-112.

[72] Salton, G., The SMART System 1961-1976: Experiments in Dynamic Document Processing. In *Encyclopedia of Library and Information Science*, 1980, 1-36.

[73] Salton, G., Fox, E.A. and Wu, H., Extended Boolean Information Retrieval. *Commun. ACM*, 26(11), Nov. 1983, 1022-1036.

[74] Simmons, R.F., *Computations from the English.* Prentice Hall, Englewood Cliffs NJ, 1984.

[75] Sirbu, Jr., M.A. and Sutherland, J.B., Naming and Directory Issues in Message Transfer Systems. *Proc. IFIP 6.5 Working Conf., Nottingham, England*, May 1984.

[76] Small, H.G., Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *J. Am. Soc. Inf. Sci.*, 24(4), July-Aug. 1973, 265-269.

[77] Small, S. and Rieger, C., Parsing and Comprehending with Word Experts (A Theory' and its Realization). In *Strategies for Natural Language Processing*, ed. by Wendy G. Lehnert and Martin H. Ringle, Lawrence Erlbaum Assoc., Hillsdale NJ, 1982, 89-148.

[78] Smeaton, A. F., Relevance Feedback and a Fuzzy Set of Search Terms in an Information Retrieval System. *Inf. Tech.: Res. Dev. Applications*, 3(1), Jan. 1984, 15-24.

[79] Smith, L.C. and Warner, A.J., A Taxonomy of Representations in Information Retrieval System Design. In *Representation and Exchange of Knowledge as a Basis of Information Processes*, ed. by Hans J. Dietschmann, North-Holland, New York, 1984, 31-49.

[80] Solomon, M., Landweber, L.H. and Neuhengen, D., The CSNET Name Server. In *Computer Networks 6*, North-Holland, 1982.

[81] Spark Jones, K. and Tait, J.I., Automatic Search Term Variant Generation. *J. Doc.*, 40(1), March 1984, 50-66.

[82] Stonebraker, M., et al. Document Processing in a Relational Database Systems. *ACM Trans. on Office Information Systems*, 1(2), April 1983.

[83] Tong, R.M., et al. A Rule-Based Approach to Information Retrieval: Some Results and Comments. *Proc. AAAI-83*, 1983.

[84] Van Rijsbergen, C.J., *Information Retrieval: Second Edition*. Butterworths, London, 1979.

[85] Vickers, P.H., Common Problems of Documentary Information Transfer, Storage and Retrieval in Industrial Organizations. *J. Doc.*, 39(4), Dec. 1983, 217-229.

[86] Winograd, T., *Language as a Cognitive Process. Volume I: Syntax*. Addison-Wesley, Reading, MA, 1983.

[87] Yu, C.T., Buckley, C., Lam, K. and Salton, G., A Generalized Term Dependence Model in Information Retrieval. *Inf. Tech.: Res. and Dev.*, 2(4), Oct. 1983.

[88] Zadeh, L.A., Fuzzy Sets. *Information and Control*, 8, 1965, 338-353.

[89] Zarri, G.P., An Outline of the Representation and Use of Temporal Data in the RESEDA System. *Inf. Tech.: Res. Dev. Applications*, 2(2/3), July 1983, 89-108.