Words, Concepts, or Both: Optimal Indexing Units for Automated Information Retrieval

William R. Hersh, M.D., David H. Hickam, M.D., M.P.H., T.J. Leone, M.S.
Biomedical Information Communication Center
Oregon Health Sciences University
Portland, OR, USA

ABSTRACT

What is the best way to represent the content of documents in an information retrieval system? This study compares the retrieval effectiveness of five different methods for automated (machine-assigned) indexing using three test collections. The consistently best methods are those that use indexing based on the words that occur in the available text of each document. Methods used to map text into concepts from a controlled vocabulary showed no advantage over the word-based methods. This study also looked at an approach to relevance feedback which showed benefit for both word-based and concept-based methods.

INTRODUCTION

One of the biggest controversies in the information science field concerns the optimal units for representing the content of a document. Early text retrieval systems relied mostly on human indexing, where multiword terms from controlled vocabularies were assigned by trained coders. Early automated retrieval systems tended to rely on individual words from the document itself. Later, methods for indexing based on phrases or concepts, often utilizing computational linguistic techniques, emerged.

Each method has its proponents. Blair, an advocate of human indexing, has argued, "It is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by *all* (or most) relevant documents and *only* (or primarily) by those documents" [1]. Salton has stated, however, that "accuracy and consistency are difficult to maintain" for human indexers [2], a view that is reinforced by data showing only 30-60% consistency among indexing assignments made to MEDLINE records indexed in duplicate [3]. As for whether the optimal units of indexing in automated systems are single words or phrases, Salton claims that his work over the last 30 years shows that single words lead to the best and most consistent performance in heterogeneous text retrieval environments. But Evans has claimed that "string- and keyword-based indexing cannot accommodate variation in language use, which becomes amplified in large databases" [4], arguing that the underlying conceptual content of documents must be represented for adequate automated indexing.

The SAPHIRE retrieval system represents a middle ground in the automated indexing spectrum [5]. Indexing is based on concepts, although without the need for ambiguity-resolving natural language processing techniques and their requirement for large knowledge bases. SAPHIRE uses a non-syntactic pattern-matching approach and exploits the Metathesaurus of the Unified Medical Language System (UMLS) of the National Library of Medicine [6]. When modified for use by SAPHIRE, the Metathesaurus contains over 28,000 medical concepts and 70,000 synonyms for those concepts, obviating the need for knowledge base construction. SAPHIRE's retrieval techniques are based on the standard statistical methods descending from SMART [7]. Free text queries are processed by the same concept-matching algorithm, with weighting of concepts (as opposed to individual words) used to rank retrieved documents for relevance.

The theoretical benefit of using phrases and concepts in automated indexing is that they add specificity to document representation. Documents on topics other than hypertension may contain the three words that comprise its synonym, *high*, *blood*, and *pressure*, but are used in different contexts (e.g., "Ocular *pressure* was *high*.....despite adequate *blood* levels of medication..."). By recognizing the meaning of the three words high, blood, and pressure adjacent to each other, as is done in SAPHIRE, indexing should be more specific, resulting in improved retrieval performance. The theoretical benefit of using phrases and concepts for indexing, however, has not translated into significantly improved retrieval effectiveness. The best methods developed so far use statistical co-occurrences of words to generate phrases [8]. Methods using syntactic approaches have been less successful [9, 10], although a recent approach using structured queries with a probabilistic model based on inference nets showed enhanced retrieval in the computer science domain [11].

To determine which of these diverse approaches to indexing shows the greatest promise, it is necessary to conduct scientific evaluations of retrieval systems. Salton has performed studies indicating the superiority of word-based automated indexing, although his comparison with MEDLARS is old and based upon use in non-operational settings [12]. Blair points to a study done in the legal domain showing the alleged futility of

0195-4210/92/$5.00

full-text, word-based searching, yet this approach is not compared with any others, and the actual system used features few advanced retrieval methodologies [13]. Most linguistic-based systems, such as CLARIT [14], have not yet been evaluated in broad domains such as medicine.

Previous studies have demonstrated the equivalence of SAPHIRE with conventional searching techniques for novice searchers [5, 15, 16]. This study compares SAPHIRE with other automated approaches to information retrieval that feature automated indexing along with natural language query input and relevance ranking of output. For this experiment, we utilized a second program that has been created for information retrieval research at the BICC called SWORD (Statistical Word-Oriented Retrieval from Databases). SWORD utilizes the approach to information retrieval pioneered by Salton in the SMART system [7]. In SWORD all words in documents (or their abstracts) not on a stop list of 250 common words are stemmed and are given a weight based on their frequency in the document and infrequency in other documents. Some modifications were made to both SAPHIRE and SWORD for additional experiments.

This study compared the following five approaches to automated indexing:

1. Concept-based (SAPHIRE)
2. Word-based (SWORD)
3. Word-based with only medical words (words that occur in the Metathesaurus)
4. Combination of word-based and concept-based
5. Concept-based with broader concept recognition

All of the above approaches have several features in common. Indexing is performed by breaking out the individual indexing units (words or concepts) and assigning them a weight based on their frequency in the document and infrequency in other documents. Retrieval occurs by the entry of a natural language query, from which the appropriate words or concepts are broken out. Each document is given a score, based on the sum of the indexing item weights for all items that occur in both the query and document. In essence, the retrieval process is a logical OR of all search terms with weighting of the documents by methods that aim to cluster the documents most similar (and presumably most relevant) to the query near the top of the retrieval list.

Where the approaches differ is in their indexing units. SAPHIRE, as has been described elsewhere, uses a concept-matching algorithm based on a vocabulary from the Metathesaurus [17]. SWORD indexes based on each individual word that occurs in a document, discarding those in a stop list of common words and stemming the others. The third method utilizes the word-based approach but only includes medically important words, where important is defined as occurring in the Metathesaurus vocabulary. The fourth method uses a combination of SAPHIRE and SWORD indexing. The rationale for this approach is that it gives weight to words in documents that cannot be mapped into concepts from SAPHIRE's vocabulary.

The final method substitutes a different concept-matching algorithm in SAPHIRE. A criticism of the original SAPHIRE approach, which requires concepts in text being indexed to appear in the exact word order that they occur in the vocabulary, is that it is too restrictive for matching concepts expressed in text. It has also been criticized for its inability to match concepts partially, as occurs when some but not all of the words of a concept are present. Text analysis has shown that the words in a concept can not only be spread across words in sentences, but in adjacent sentences as well, reflecting the diversity of human writing [18]. The substitute concept-matching algorithm relaxes the word order requirement, allowing words in a concept to appear adjacent to each other (with stop words ignored), in any order. For example, the concept *Hypertension* would be extracted from the phrase "pressure of the blood is high" in the newer algorithm but not the original. This is because the stop words *of*, *the*, and *is* would be removed, with the resulting words *pressure*, *blood*, and *high* mapping into the term *High Blood Pressure*, which is a synonym of *Hypertension*. An additional feature of the substitute algorithm is that it allows partial matching of concepts. When more than one word of a concept but not all are present, a concept is matched. When a smaller concept is subsumed by a larger one, both are matched. For example, the phrase "heart failure" will map into the concept *Congestive Heart Failure* in the newer algorithm but not the original.

This broader concept recognition is not, however, without disadvantages. First, there is greater potential for mismatched concepts. For example, the phrase "ocular pressure was high" will map into the concept *Hypertension*, since the phrase "ocular pressure" does not occur in the Metathesaurus vocabulary and the words *high* and *pressure* are adjacent. (The concept *Ocular Hypertension* does occur, although there is no synonym form that contains the word *pressure*.) Second, the computational complexity of the new algorithm is much higher. In fact, the algorithm runs two to four times slower than SAPHIRE's original one.

In this study we also investigated the use of relevance feedback. The purpose of relevance feedback is to use retrieved documents that are deemed relevant by the user to find additional similar documents. While several elaborate methods for this process have been described [19, 20], SAPHIRE and SWORD use a simpler method whereby the title and text of a relevant document become a new query. This procedure was tested using the original SAPHIRE and SWORD programs.

METHODS

In order to test the performance of each method, we used methods standard to the information science field. Performance is assessed by two parameters, recall and precision:

$$\text{Recall} = \frac{\text{num docs retrieved and relevant}}{\text{num docs relevant in database}} \quad (1)$$

$$\text{Precision} = \frac{\text{num docs retrieved and relevant}}{\text{num docs relevant in retrieval set}} \quad (2)$$

In order to calculate recall and precision, one must have judgments of which documents a test collection of documents and queries are relevant to each query. From previous studies, we had three test collections:

1. 200 abstracts and 12 queries from the AIDSLINE database [5]
2. 2,344 abstracts and 75 queries from the MEDLINE database [15]
3. 1,992 documents and 10 queries from the Yearbook Series [16]

The first two collections have the title and abstract for indexing each document, while the third has the full text from each Yearbook used, consisting of the title, abstract, and expert commentary for each article. Each query was a free text expression from the user who originated it. Relevance judgments were made by physicians.

Because each of the systems tested in this study uses relevance ranking, each query generates a ranked list of matching documents. The standard evaluation approach is to generate a recall-precision table consisting of the corresponding precision at fixed values of recall. This allows comparison of different methods by comparing precision levels at each point. (This assumes equal importance between recall and precision; there are other evaluation metrics that designate the importance of one over the other.) As an example of this process, consider a query that contains ten relevant documents. When the first relevant document on the retrieval list is found, the 10% recall level has been reached, and the precision at that point is the precision level for 10% recall. This continues as far down the list as there are relevant documents. When values for the fixed levels of recall are missing (e.g., a query with five relevant documents whose first relevant document will give a value for precision at the 20% recall level), the missing values are interpolated between those which are present. In general, the value for precision will decrease as recall increases. For an entire test collection, the values of precision for each fixed level of recall are averaged. For relevance feedback, the text from the top-ranking relevant document was designated as the new query. This generated new recall-precision values, which were evaluated as above.

For statistical analysis to compare performance among methods for each test collection, all recall-precision graphs were converted to Receiver Operating Characteristic (ROC) curves by translating the precision values to false-positive rates. Areas under the ROC curves summarize the performance of each indexing method for capturing documents relevant to a query. A separate ROC curve was constructed for each query-method pair, and the ROC areas were calculated using the trapezoidal technique [21]. Analysis of variance was used to compare ROC areas among methods. Each indexing method was included in the analysis as a repeated measure. The identity of the document collection was treated as a between-subjects factor. Post-hoc comparisons between methods were performed using paired t-tests with the Bonferroni correlation.

RESULTS

Each of the word-based indexing methods produced significantly ($P < .0001$) better retrieval performance than the concept-based methods. There was no significant difference among the two word-based methods (SWORD, medical words only) and the combination of SWORD and SAPHIRE. There was also no significant difference between either concept-based approach. The results from each indexing approach for each test collection are shown in Table 1. Precision at recall levels of 20%, 50%, and 80%, representing low-, medium-, and high-recall searches are listed, along with the mean area under the ROC curve for each query. Larger ROC areas denote better retrieval performance. Although the ROC areas tended to be higher in the Yearbook test collection for all methods, this trend was not statistically significant in the analysis of variance.

The relevance feedback was associated with significant ($P < .0001$) improvement in ROC areas for both SAPHIRE and SWORD. There is improvement in average ROC area for all three test collections, but it did not occur uniformly along the recall-precision continuum for the AIDSLINE and Yearbook collections. For the MEDLINE collection, however, the improvement was substantial for both programs.

DISCUSSION

The results of this study suggest that present concept-based indexing methods provide no apparent advantages over word-based methods. While the method of combining SAPHIRE and SWORD performs about as well as SWORD and the medical-words method, it never exceeds the word-only approaches. The pure concept-based approaches, original SAPHIRE and SAPHIRE with broader concept recognition, perform inferiorly.

Why did SAPHIRE's concept-recognition not perform retrieval as well as free words in text? The most likely answer is the inherent ambiguity of text, where writers express concepts in ways more numerous than can be captured in controlled vocabularies such as the Metathesaurus. To this end, ability to identify syntactic and semantic features in text may be required. We recently showed that natural language processing methods lead to better discrimination among documents from the *New England Journal of Medicine* in the AIDS domain [22]. It is also possible that the abstract alone is insufficient to profile adequately the content of a document, and the full text or a more informative (e.g., structured) abstract may be required.

Why did the substitute concept-matching algorithm fail to perform better than SAPHIRE's original algorithm? Inspection of the indexing logs showed that a large number of inappropriate concepts were matched. For example, the words *leukemia* and *cell* adjacent to each other led to designation of every variety of T-Cell Leukemia, B-Cell Leukemia, and others. While one can speculate on how to engineer one's way out of these and other particular failures, it is clear that inexact concept matching often leads to inappropriate assignment of indexing terms.

The results of the relevance feedback method shows marked improvement in performance with the MEDLINE test collection for SAPHIRE and SWORD. While there is overall improvement for both programs with the other collections as measured by the increased average ROC area, the improvement is not uniform. The one significant difference between the MEDLINE collection and the other two is the overall poorer performance with all methods. Perhaps the relevance feedback used by SAPHIRE and SWORD works best with queries that initially achieve poorer results. In any case, more analysis is needed to determine the role for this and other feedback measures.

There are limitations to this study, the most important of which is the non-interactive, batch method of query entry. Before concluding that word-based indexing methods are superior, they must be assessed in operational retrieval settings. Previous studies have shown SAPHIRE's equivalence to MEDLINE-style Boolean searching [5] and word-based Boolean searching for novice searchers [15]; studies like these have not yet been done comparing a system like SWORD in actual use.

Another limitation to this study is the relatively small size of the test collections. Blair has argued:

> Information retrieval systems do not scale up. That is, retrieval strategies that work well on small systems do not necessarily work well on larger systems (primary because of output overload). This means that studies of retrieval effectiveness must be done on full-sized retrieval systems if the results are to be indicative of how a large, operation system would perform [1].

In a conventional Boolean retrieval system, for example, a search strategy that would return 10 documents in a 1,000 document collection would return 10,000 documents in a million document collection, which no searcher could possibly manage. However, Salton notes that most statistical approaches to retrieval use relevance ranking, which orders the retrieval set by similarity to the query, thus suggesting to the user which documents are most likely to be relevant [23].

The past two years of evaluation of SAPHIRE have generated equivocal results about the potential of this system. The techniques of automated indexing, natural language query, and relevance ranking clearly perform at least as well as traditional techniques. Our goal now is to scale these techniques up to large test collections and operational retrieval settings. To this end, we have been collecting data on the usage of a workstation featuring multiple retrieval applications installed in the work area of the General Medicine Clinic at Oregon Health Sciences University. This system is used by both faculty and housestaff. With the queries and references retrieved, we are also building a test collection that will contain approximately 200,000 documents.

Acknowledgments

This study was supported in part by Grant LM05307 from the National Library of Medicine.

References

1. Blair DC. "Language and Representation in Information Retrieval." 1990, Elsevier, New York.
2. Salton G. "Introduction to Modern Information Retrieval." 1983, McGraw-Hill, New York.
3. Funk ME, Reid CA, Indexing consistency in MEDLINE. Bull Med Lib Assoc. 1983; 71: 176-183.
4. Evans DA, Ginther-Webster K, Hart M, Lefferts RG, Monarch IA, Automatic Indexing Using Selective NLP and First-Order Thesauri. RIAO 91. 1991; 624-644.
5. Hersh WR, Hickam DH, A comparison of retrieval effectiveness for three methods of indexing medical literature. Am J Med Sci. 1992; 303: 292-300.
6. Lindberg DAB, Humphreys BL, The UMLS knowledge sources: Tools for building better user interfaces. SCAMC 14. 1990; 121-125.
7. Salton G, Developments in automatic text retrieval. Science. 1991; 253: 974-980.
8. Fagan J, The effectiveness of a non-syntactic approach to automatic phrase indexing for document retrieval. J Amer Soc Info Sci. 1989; 40: 115-132.
9. Salton G, Buckley C, Smith M, On the application of syntactic methodologies in automatic text analysis. Info Proc Mgmt. 1990; 26: 73-92.

10. Dillon M, Gray AS, FASIT: A fully automatic syntactically based indexing system. J Amer Soc Info Sci. 1983; 34: 99-108.
11. Croft WB, Turtle HR, Lewis DD, The use of phrases and structured queries in information retrieval. Proc 14th Ann ACM/SIGIR Conference on Information Retrieval. 1991; 32-45.
12. Salton G, A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). J Amer Soc Info Sci. 1972; 23: 75-84.
13. Blair DC, Maron ME, An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM. 1985; 28: 289-299.
14. Evans DA, Hersh WR, Monarch IA, Lefferts RG, Handerson SK, Automatic indexing of abstracts via natural language processing using a simple thesaurus. Medical Decision Making. 1991; 11(supp): S108-S115.
15. Hersh WR, Hickam DH, Haynes RB, McKibbon KA, Evaluation of SAPHIRE: A Concept-Based Approach to Information Retrieval. SCAMC 15. 1991; 808-812.
16. Hersh WR, Hickam DA, A comparison of two methods for indexing and retrieval from a full-text medical database. Proc 55th Annual Meeting of Amer Soc Info Sci. 1992, in press.
17. Hersh WR, Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. Medical Decision Making. 1991; 11(supp): S120-S124.
18. Warner AJ, Wenzel PH, A linguistic analysis and categorization of nominal expressions. Proc 54th Annual Meeting of Amer Soc Info Sci. 1992; 186-195.
19. Cousins SB, Silverstein JC, Frisse ME, Query networks for medical information retrieval - Assigning probabilistic relationships. SCAMC 14. 1990; 800-804.
20. Salton G, Buckley C, Improving retrieval performance by relevance feedback. J Amer Soc Info Sci. 1990; 41: 288-297.
21. Hanley JA, McNeil BJ, A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983; 148: 839-843.
22. Hersh WR, Evans DA, Monarch IA, Lefferts RG, Handerson SK, Gorman PN, Indexing effectiveness of linguistic and non-linguistic approaches to automated indexing. MEDINFO 92. 1992; in press.
23. Salton G, Another look at automatic text-retrieval systems. Communications of the ACM. 1986; 29: 648-656.

Table 1 -- Performance of five methods for indexing medical literature

Mean Precision	20% recall	50% recall	80% recall	Area under ROC
AIDSLINE				
SAPHIRE	75	57	16	0.813
SWORD	81	64	28	0.819
Medical words only	79	53	26	0.811
SAPHIRE/SWORD	89	74	30	0.849
Broader SAPHIRE	62	40	18	0.735
MEDLINE				
SAPHIRE	54	43	29	0.768
SWORD	62	52	38	0.889
Medical words only	61	51	36	0.885
SAPHIRE/SWORD	62	51	39	0.881
Broader SAPHIRE	50	40	27	0.775
Yearbook				
SAPHIRE	72	50	30	0.899
SWORD	78	54	33	0.960
Medical words only	79	59	41	0.970
SAPHIRE/SWORD	76	59	36	0.968
Broader SAPHIRE	65	46	22	0.831

Table 2 -- Performance values for SAPHIRE and SWORD using relevance feedback (change from results without feedback shown in parentheses)

Mean Prec.	20% recall	50% recall	80% recall	Area under ROC
AIDSLINE				
SAPHIRE	73 (-2)	49 (-8)	31 (+15)	0.839 (+0.03)
SWORD	71 (-10)	58 (-6)	34 (+6)	0.871 (+0.05)
MEDLINE				
SAPHIRE	74 (+20)	53 (+10)	36 (+7)	0.893 (+0.1)
SWORD	86 (+24)	68 (+16)	51 (+13)	0.965 (+0.08)
Yearbook				
SAPHIRE	79 (+7)	43 (-7)	23 (-7)	0.964 (+0.07)
SWORD	74 (-4)	49 (-5)	36 (+3)	0.967 (+0.01)