Document Filtering Using Semantic Information from a Machine Readable Dictionary[1]

Elizabeth D. Liddy, Woojin Paik
School of Information Studies
Syracuse University

Edmund S. Yu
College of Engineering and Computer Science
Syracuse University

Abstract

Large scale information retrieval systems need to refine the flow of documents which will receive further fine-grain analysis to those documents with a high potential for relevance to their respective users. This paper reports on research we have conducted into the usefulness of semantic codes from a machine readable dictionary for filtering large sets of incoming documents for their broad subject appropriateness to a topic of interest. The Subject Field Coder produces a summary-level semantic representation of a text's contents by tagging each word in the document with the appropriate, disambiguated Subject Field Code (SFC). The within-document SFCs are normalized to produce a vector of the SFCs representing that document's contents. Queries are likewise represented as SFC vectors and then compared to SFC vectors of incoming documents, which are then ranked according to similarity to the query SFC vector. Only those documents whose SFC vectors exhibit a predetermined degree of similarity to the query SFC vector are passed to later system components for more refined representation and matching. The assignment of SFCs is fully automatic, efficient and has been empirically tested as a reasonable approach for ranking documents from a very large incoming flow of documents. We report details of the implementation, as well as results of an empirical testing of the Subject Field Coder on fifty queries.

1. Information Filtering

Two realities regarding the current context of information retrieval motivate the research herein reported: 1) Document collections from which individuals need to receive and/or retrieve relevant information are immense in size and only likely to increase; 2) Given the size of both the daily influx of documents and the document databases in which the daily input is then stored, a finer level of representation of both information needs and documents is necessary in order to ensure higher precision results. Although precision has always been a concern in information retrieval, the problem assumes new significance when low precision translates into thousands of non-relevant documents that each user must peruse. Improved precision can be achieved by using a more conceptual level of representation of documents and queries, so that the system provides to the user documents containing their *concepts of interest*, not just the user's keywords. However, this level of analysis is computationally expensive and not reasonable to perform on documents that are unlikely to be relevant. Therefore, preliminary filtering of documents in an information retrieval system would permit later, finer levels of text analysis to be more efficiently applied to a smaller subset of documents.

This suggests the view that information retrieval be approached as a multi-stage filtering process, with the types and optimal number of filterings dependent on both the size of the document collection and the desired granularity of filtering. We believe that intelligent filtering is needed in document detection applications, where millions of documents are received daily by an organization, while only a relatively small subset of documents is of potential interest to any individual user. Furthermore, we believe that a

[1]Support for this research was provided by DARPA Contract #91-F136100-00 under the auspices of the TIPSTER Project.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1993		**00-00-1993 to 00-00-1993**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Document Filtering Using Semantic Information from a Machine Readable Dictionary	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **School of Information Studies,Syracuse University,Syracuse,NY,13244**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**10**	

purely content-based document filter would be useful in delineating a subject-appropriate preliminary set of documents for each user on which the system would then perform finer levels of analysis and matching.

The notion of filtering as used here, is to be distinguished from one sense of the term currently in use. Belkin and Croft (1992) define information filtering broadly as "a variety of processes involving the delivery of information to people who need it" (p.29). Defined as such, the work we are herein reporting, fits the definition of filtering. However, Belkin and Croft describe a particular application of information filtering which equates to Selective Dissemination of Information (SDI). This view of filtering is at a finer grain of matching than our notion of filtering. In an SDI application, filtering is the full matching process, while we conceive of filtering as a rougher-grain, first stage, topic-area matching. In a one-stage SDI application, user-profiles may contain facets of description beyond the desired content of useful documents, whereas our preliminary filter relies solely on topic-based criteria. The goal of our filter is to efficiently and effectively skim off those documents which possess the greatest likelihood of proving relevant to a user's need, here conceived of as a natural language statement of their long-standing information requirement. Later stages of processing in the system will perform the more refined conceptual level of matching.

2. DR-LINK Project

Our ongoing research into the development and implementation of an effective document filter has produced a module used within a larger document detection system, the DR-LINK System (Liddy & Myaeng, 1993). The DR-LINK Project is research being conducted under the auspices of DARPA's TIPSTER Project whose goal is the development of algorithms both for the detection of documents of interest and the extraction of selected information from these documents for a large group of users. The DR-LINK system architecture is modular in design, with six separate processing modules. These modules enhance the documents at every stage by various semantic enrichments which are used to refine the flow of documents in terms of both appropriateness to the query and pure numbers. Briefly summarized, the six modules' processing is as follows:

1) The **Subject Field Coder** uses semantic word knowledge to produce a summary-level topical vector representation of a document's contents that is matched to a vector representation of a query in order to select for further processing only those documents which have real potential of being relevant. This subset of documents is then passed to the:
2) **The Proper Noun Interpreter**, which uses a variety of knowledge bases and context-based heuristics to recognize, categorize, and standardize every proper noun in the text. The similarity between a query's proper noun requirements and each document's Proper Noun Field is computed at either the category level or by precise string matching. This similarity value is combined with the similarity value from the Subject Field Coder for a reranking of all documents in response to the query. Those documents which exceed an empirically determined cut-off criterion based on this combined similarity value, are then passed to:
3) **The Text Structurer**, which sub-divides a text into its discourse-level segments in order to focus matching on the appropriate discourse component in the documents in response to the particular requirements of an information need. For example, for queries run against the newspaper database that are seeking information about a particular possible future event (e.g. Japanese acquisition of U.S. companies), the Text Structurer matching algorithm will weight more highly those articles in which mention of the event occurs in an 'Expectation' component. When retrieved, the structured texts, with the appropriate components high-lighted, are passed to the:
4) **Relation-Concept Detector**, which raises the level at which we do matching from a key-word or key-phrase level to a more conceptual level by expanding terms in the query to all terms which have been shown to be 'substitutable' for them, and then by extracting semantic relations between concepts in both documents and queries. This component produces concept-relation-concept triples which are passed to the:

5) **Conceptual Graph Generator** which converts the triples into the Conceptual Graph (CG) formalism, a representation similar to semantic networks, but with labelled arcs (Sowa, 1984). The resultant CGs are passed to the:
6) **Conceptual Graph Matcher**, which measures the degree to which a particular query CG and candidate document CGs share a common structure, and then produces a final ranking of the documents.

Since the later modules in our system require very complex processing in order to produce conceptually enriched representations of documents and queries, preliminary filtering of the incoming flow of documents by means of the Subject Field Coder has proven to be extremely useful. For while CGs enable us to do fine-grained representation, such fine-grained representation is not necessary in order to determine, for instance, that a document on 'computer games' is not likely to be relevant to a query on 'merit pay'. Therefore, the SFCoder produces a first rough cut of those documents which have real potential for matching a query as the first of a multi-stage model of retrieval. Because the SFCoder is based on the implicit semantics of the words in the text, it has the ability to successfully eliminate non-topic relevant documents during a preliminary stage without the attendant risks of filtering approaches which are based on less semantically reliable characteristics of documents.

3. Representation Used in Filtering

Subject filtering is a difficult problem, given the richness and variety of natural language. In addition, imposition of an overly stringent subject filter in too homogenous a document collection runs the risk of excluding docu-ments which might match the query during a later, finer matching process. This is particularly true if a lexical or keyword analysis of text is the basis of the filtering. However, if based on an appropriate semantic representation combined with a reasonable cut-off criterion, a subject-based filter offers the means of siphoning off from a large heterogenous stream of documents, smaller, more appropriate sub-collections of documents for various users or user-groups, for which the system then produces more conceptual representations and performs finer-grain matching.

The success of our filtering approach is attributable to the nature of the representation scheme we use for every text (whether document or query). The representation of each text is a summary vector of the Subject Field Codes (SFCs) from Longman's Dictionary of Contemporary English (LDOCE) representing the correct sense of each word in the text that is in LDOCE and which has SFCs assigned in LDOCE. For example, Figure 1 presents a short Wall Street Journal article and a humanly readable version of the normalized SFC vector which serves as the document's semantic summary representation.

A U. S. magistrate in Florida ordered Carlos Lehder Rivas, described as among the world's leading cocaine traffickers, held without bond on 11 drug-smuggling counts. Lehder, who was captured last week in Colombia and immediately extradited to the U.S., pleaded innocent to the charges in federal court in Jacksonville.

LAW	.2667	SOCIOLOGY	.1333
BUSINESS	.1333	ECONOMICS	.0667
DRUGS	.1333	MILITARY	.0667
POLITICAL SCIENCE	.1333	OCCUPATIONS	.0667

Fig. 1: Sample Wall Street Journal document and its SFC representation

---

As can be seen by reading either the original text or the SFC vector values, the text's main topic is law, while the topics of business, drugs, political science and sociology are equally, but less significantly mentioned. The vector suggests a passing reference to the fields of economics, military, and occupations.

The system would consider this document relevant to a query whose SFC representation was distributed proportionately among the same SFCs slots on the vector. The important aspect of this representation is that a document does **not** need to include any of the same words that are included in a query in order for a high similarity to be found between the query and a document, since the matching is based on similarity of SFC vectors, not according to the particular words used.

Therefore, it can be seen that the SFC representation, which is one level of abstraction above the actual words in a text, implicitly handles both the synonymy (multiple words having the same meaning) and polysemy (one word having multiple meanings) problems which have plagued the use of natural language in information retrieval systems. This level of abstraction is an essential feature of the representation since it has been shown (Furnas et al, 1987) that users' information requests frequently share little vocabulary overlap with the documents which actually contain relevant information.

4. Longman's Dictionary of Contemporary English

Our text representation is based on the machine-readable version of Longman's Dictionary of Contemporary English (LDOCE), a British-produced learner's dictionary. The first edition of LDOCE has been used in a number of investigations into natural language processing applications (Boguraev & Briscoe, 1989). We are using the second edition (1987) which contains 35,899 headwords and 53,838 senses. The machine-readable tape of LDOCE contains several fields of information not visible in the hard-copy version which are extremely useful in natural language processing tasks. Some of these are relevant for syntactic processing, while others contain semantic information, which indicate the class of entities to which a noun belongs (e.g. animate, abstract) or the semantic constraints for the arguments of a verb or an adjective, and the SFCs, which are the basis of our text representation for document filtering.

The SFCs comprise a classification scheme of 124 major fields, based on an earlier classification scheme of Merriam-Webster. SFCs are manually assigned to words in LDOCE by the Longman lexicographers. There are two types of problems with the SFCs which we have resolved in order to use them computationally. First, a particular word may function as more than one part of speech and secondly, if a word has more than one sense, each of these senses may be tagged in the lexicon with different SFCs. Therefore, in order for SFCs to provide a reasonable representation of texts, a system must ascertain both the grammatical function and sense of a word in the text, so that the appropriate SFC for each orthographic form can be chosen. We have incorporated in our system means for choosing amongst each word's syntactic categories and senses found in LDOCE, thereby enabling the system to assign just one SFC to each word in a given text.

In related research, Walker and Amsler used the Subject Field Codes to determine the appropriate subject domains for a set of texts (1986). However, they used the most frequent SFC to characterize a document's content, whereas we represent a document by a vector of frequencies of SFCs for words in that text. Slator (1991) has taken the original 124 SFCs and added an additional layer of seven pragmatic classes to the original two-level hierarchy. He has found the reconstructed hierarchy useful when attempting to disambiguate multiple senses and SFCs attached to words. His metric for preferring one sense over another relies on values within an individual text, whereas we add corpus correlation values as a further stage in the disambiguation process. Krovetz (1991) has been exploring the effect of combining the evidence from SFCs with evidence from other fields in LDOCE for selection of a correct word sense. His goal is to represent documents by their appropriate senses rather than just the orthographic forms of words, for use in an information retrieval system.

5. Subject Field Coding of Texts

In the Subject Field Coder, the following stages of processing are done in order to generate a SFC vector representation of each text:

In **Stage 1** processing, we run the documents and query through POST, a probabilistic part of speech

tagger (Meeter et al, 1991) in order to restrict candidate SFCs of a word to those of just the appropriate syntactic category of each word as determined by POST.

Stage 2 processing consists of retrieving the SFCs of each word's correct part of speech from the lexical database. The SFC retrieval process utilizes the Kelly & Stone (1975) stemming algorithm to reduce morphological variants to their simple stem as found in LDOCE. Kelly & Stone's approach is one of weak stemming and produces correct simple stems for look-up in LDOCE, rather than stripping suffixes.

Having selected candidate SFCs for each word's correct syntactic category, we begin sense disambiguation at **Stage 3**, using sentence-level context-heuristics to determine a single word's correct SFC. We begin with context-heuristics because empirical results have shown that local context is used by humans for sense disambiguation (Choueka & Lusignan, 1985) and context-heuristics have been experimentally tested in Walker & Amsler's (1986) and Slator's work (1991) with promising results. The input to Stage 3 is a word, its part-of-speech tag, and the SFCs for each sense of that grammatical category. For some words, no disambiguation may be necessary at this stage because the SFCs for the part-of-speech of the input word may all be GENERAL or there may be no SFCs provided by LDOCE. However, for the majority of words in each sentence there are multiple SFCs, so the input would be as seen in Figure 2.

State	n	POLITICAL SCIENCE[4], ORDERS
companies	n	BUSINESS, MUSIC, THEATER
employ	v	LABOR, BUSINESS
about	adv	-
one	adj	-
billion	adj	NUMBERS
people.	n	SOCIOLOGY, POLITICAL SCIENCE[2], ANTHROPOLOGY

Fig 2: Subject Field Codes & Frequencies (in Superscript) for Part-of-Speech Tagged Words

To select a single SFC for each word in a sentence, Stage 3 uses an ordered set of context-heuristics. First, the SFCs attached to all words in a sentence are evaluated to determine at the sentence level: 1) whether any words have only one SFC assigned to all senses of that word, and; 2) the SFCs which are most frequently assigned across all words in the sentence. Each sentence may have more than one unique SFC, as there may be more than one word whose senses have all been assigned a single SFC. In Figure 2, NUMBERS is a unique SFC, being the only SFC assigned to the word 'billion' and POLITICAL SCIENCE is the highly frequent SFC for this sentence, being assigned to 6 senses in total. The unique SFCs and the highly frequent SFCs have proven to be good local determinants of the subject domain of the sentence. We have established the criterion that if no SFC has a frequency equal to or greater than three, a frequency-based SFC for that particular sentence is not selected. Empirical results show that SFCs having a within-sentence frequency less than three do not accurately represent the domain of the sentence.

Stage 4 evaluates the remaining words in the sentence, and for some words chooses a single SFC based on the locally-important SFCs determined in Stage 3. The system scans the SFCs of each remaining word to determine whether the SFCs which have been identified as unique or highly frequent occur amongst the multiple SFCs which LDOCE lexicographers have assigned to that word. In Figure 2, for example, POLITICAL SCIENCE would be selected as the appropriate SFC for 'people' and 'state' because POLITICAL SCIENCE was determined in Stage 3 to be the most frequent SFC value for the sentence.

For the ambiguous words which have no SFC in common with the unique or highly frequent SFCs for that sentence, **Stage 5** incorporates two global knowledge sources to complete the sense disambiguation task. The primary source is a 122 x 122 correlation matrix computed from the SFC frequencies of the 442,059 words that occurred in a sample of 977 Wall Street Journal (WSJ) articles. The matrix reflects stable

estimates of SFCs which co-occur within documents for this text type. The second source is the order in which the senses of a word are listed in LDOCE. Ordering of senses in LDOCE is determined by Longman's lexicographers based on frequency of use in the English language.

The correlation matrix was produced by running SAS on the SFC output of the 977 WSJ articles processed through Stage 2. At that stage, each document is represented by a vector of SFCs of all senses of the correct part-of-speech of each word as determined by POST. The observation unit is the document and the variables being correlated are the 122 SFCs. The scores for the variables are the within-document frequencies of each SFC. There are 255,709 scores across the 977 articles on which the matrix is computed. The resulting values in the 122 x 122 matrix are the Pearson product moment correlation coefficients between SFCs and range from a +1 to a -1, with 0 indicating no relationship between the SFCs. For example, NET GAMES and COURT GAMES have the highest correlation, with ECONOMICS and BUSINESS having the next highest correlation.

The matrix is then used in Stage 5, where each of the remaining ambiguous words is resolved a word at a time, by accessing the matrix via the unique and highly frequent SFCs determined for a sentence in Stage 3. The system evaluates the correlation coefficients between the unique/highly frequent SFCs of the sentence and the multiple SFCs assigned to the word being disambiguated to determine which of the multiple SFCs has the highest correlation with the unique and/or highly frequent SFCs. The system then selects that SFC as the unambiguous representation of the sense of the word.

We have developed heuristics for three cases for selecting a single SFC for a word using the correlation matrix. The three cases function better than handling all instances as a single case because of the special treatment needed for words with the less-substantive GENERAL (XX) code. When XX is amongst the SFCs, we take order of the SFCs into consideration, reflecting the fact that the first SFC listed is more likely to be correct, since the most widely used sense is listed first in LDOCE. So, to overcome this likelihood, a more substantive SFC listed later in the entry must have a much higher correlation with a sentence-determined SFC in order to be selected over the GENERAL (XX) code.

The system implementation of the disambiguation procedures was tested on a sample of 1638 words from WSJ having SFCs in LDOCE. The system selected a single SFC for each word, which was compared to the sense-selections made by an independent judge who was instructed to read the sentences and the definitions of the senses of each word and then to select that sense which was most correct. Overall, the SFC disambiguator selected the correct SFC 89% of the time (Liddy & Paik, 1992).

Stage 6 processing produces a vector of SFCs and their frequencies for each document and for the query. At this point the non-substantive GENERAL SFCs are removed from the vector sums, since these contribute nothing to a text's subject content representation.

In **Stage 7**, the vectors of each text are normalized using Sager's (1976) term weighting formula in order to control for the effect of document length. The choice of Sager's (1976) term weighting formula and Sager and Lockemann's (1976) similarity measure were based on an extensive study done by McGill, Koll & Noreault (1979) which empirically evaluated twenty one term-weighting formulae and twenty four similarity measures. Using the coefficient of ranking effectiveness (CRE) measure, each term weighting scheme was tested in combination with each similarity measure to determine which combination was best for either controlled representations or free text representations. Since we consider SFCodes to be a form of a controlled vocabulary (all free-text terms are reduced to 122 codes), we chose Sager's (1976) term weighting scheme and Sager & Lockemann's (1976) similarity measure since they were shown to be the best formulae for use with controlled vocabulary representation (McGill et al, 1979).

At **Stage 8**, the document vectors are compared to the query vector using Sager & Lockemann's (1976) similarity measure and a ranked listing of the documents in decreasing order of similarity is produced. Having created this ranked list of documents for each query, the system must determine how many of these

documents should reasonably be passed on to the next module. The method used is an adaptive cut-off criterion that predicts for each query and at each recall level, how many documents from the ranked list should be forwarded. The cut-off criterion uses a multiple regression formula which was developed on training data consisting of the odd-numbered Topic Statements (queries) from 1 to 50, used in both TIPSTER-Phase I and TREC-1. The training corpora consisted of Wall Street Journal articles from 1986-1992, a total of 173,255 separate documents. The regression formula used in Stage 8 is:

$$SPSV_i = e^{0.9909 - (0.6112 * RL) + (0.5455 * STDSVi) - 5}$$

where: $SPSV_i$ is the Standardized Predicted Similarity Value
RL is the designated Recall Level
$STDSV_i$ is the Standardized Top-Ranked Document Similarity Value, logarithmically transformed
i is the Topic Statement whose cut-off criterion is being predicted

RL and $STDSV_i$ significantly predicted $SPSV_i$ on the training queries ($R = .826$, $F = 265.42$, $df = 2,247$, $p < .0005$). Using this standardized value ($SPSV_i$), a linear transformation is used to produce the value which will be used as the cut-off criterion:

$$PSV_i = (SPSV_i * s.d._i) + mean_i$$

where: PSV_i is the Predicted Similarity Value
$s.d._i$ standard deviation
$mean_i$ mean

The PSV_i is used by the system to establish the cut-off criterion for each recall level for each query. The averaged results of the testing of the PSV_i using the held-out, even-numbered Topic Statements are provided in Table 1.

A. Recall level	B. Actual % of DB searched	C. % of DB searched based on PSV	D. Recall level based on PSV
0.10	1.27	0.50	0.20
0.20	2.67	0.98	0.28
0.30	4.42	2.51	0.39
0.40	6.55	5.24	0.50
0.50	8.46	8.90	0.61
0.60	10.97	13.62	0.69
0.70	13.78	19.36	0.75
0.80	17.36	25.52	0.81
0.90	23.84	32.39	0.87
1.00	52.82	39.65	0.92

Table 1: Performance of the PSV_i on 25 Topic Statements and 173,255 documents

Column A lists the recall levels used in information retrieval testing. Column B shows for each of these recall levels, what per cent of the database was **actually** searched to achieve that level of recall. These percentages are based on post hoc information and are known because the relevance assessments made by the trained analysts for these queries and documents were made available after TREC-1. Column C displays what percent of the ranked list would need to be searched to achieve that row's recall level, when the system uses the PSV as the cut-off criterion. Column D shows what the actual recall performance would be when the system uses the PSV for that recall level as the cut-off criterion.

This means that on average, if the user was striving for 70% recall, 19.36 % of the 173,255 documents would be passed on to the system's next module when the PSV_i is used as the cut-off criterion. In actuality, the PSV predicts slightly better than this, and the user would retrieve 75% of the relevant documents. And if the user were really interested in total recall, use of the PSV would require that 39.65% of the ranked documents be forwarded and these documents would in fact contain 92% of the relevant documents.

6. Testing and Results

Having produced a ranked listing of documents based on the similarity of their SFC vectors to a query vector, the most illustrative evaluation of performance would be the results provided in Table 1. We believe that these are quite reasonable filtering results. Earlier testings of the SFCoder have revealed that the most important factor in improving its performance would be recognition that a query contains a requirement that a particular proper noun be in a document in order for the document to be relevant. Therefore, we have incorporated a second level of lexical-semantic processing as an extension of the SFCoder. That is, the Proper Noun Interpreter (Paik et al; in press) includes algorithms for computing the similarity between a query's proper noun requirements and each document's Proper Noun Field. The proper noun similarity value is then combined with the similarity value produced by the SFCoder for a re-ranking in relation to the query. In the 18th month TIPSTER evaluation of our system, this re-ranking of documents based on the SFC values plus the Proper Noun values improved significantly the filtering power of the system. We have not yet adapted the PSV for predicting the cut-off criterion on the combined similarity values, but we will be doing so in the next few weeks.

7. Conclusions

As a preliminary full implementation and testing of the SFCoder as a means for semantically representing the content of texts for the purpose of delimiting a document set with a high likelihood of containing all those relevant to an individual query, we find these results promising. In a large operational system, the ability to filter out 61% of the incoming flux of millions of documents if the SFCoder alone is used, or 72% of the documents if the SFCoder + Proper Noun Interpreter is used, will have a significant impact on the system's performance.

In addition, we have also been experimenting with the SFC vectors as a means for automatically clustering documents in an established database (Liddy, Paik & Woelfel, 1992). To do this, the document vectors are clustered using Ward's agglomerative clustering algorithm (Ward, 1963) to form classes in a document database. For ad hoc retrieval, query SFC vectors are matched to the SFC vector of each cluster-centroid in the database. Clusters whose centroid vectors exhibit a predetermined similarity to the query SFC vector are either presented to the user as a semantically cohesive cluster on which to begin preliminary browsing or, passed on to other system components for further processing. A qualitative analysis of the clusters produced in this manner revealed that the use of SFCs combined with Ward's clustering algorithm resulted in meaningful groupings of documents that were similar across concepts not directly encoded in SFCs. Browsers find that documents seem to fit naturally int the cluster to which they are assigned by the system.

Beyond its uses within the DR-LINK System, the Subject Field Coder has general applicability as a pre-filter

for a wide range of other systems. The only adjustment required would be a recomputation of the correlation matrix based on each new corpus. The recomputation is necessary due to the fact that different corpora represent different domains and the tendencies of SFCs to correlate with other SFCs will vary somewhat from domain to domain. We have used the SFC filter on various corpora and have quickly recomputed a matrix for each.

Reiterating the opening argument of this paper, we believe that the current situation in information retrieval could be effectively dealt with by considering document retrieval as a multi-stage process in which the first modules of a system filter out those texts with no real likelihood of matching a user's need. The filtering approach offers promise particularly to those systems which perform a more conceptual style of representation which is very computationally expensive if applied to all documents regardless of the likelihood that they might be relevant.

Acknowledgments

We wish to thank Longman Group, Ltd. for making the machine readable version of LDOCE, 2nd Edition available to us and BBN for making POST available for our use on the DR-LINK Project.

References

Belkin, N.J. & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35 (12): 29-38.

Boguraev, B. & Briscoe, T. (1989). Computational lexicography for natural language processing. London: Longman.

Choueka, Y. & Lusignan, S. (1985). Disambiguation by short contexts. Computers and the Humanities, pp. 147-157.

Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1987). The vocabulary problem in human-system communication. Communications of the ACM, 30 (11):964-71.

Kelly, E. F. & Stone, P. J. (1975). Computer recognition of English word senses. Amsterdam: North Holland Publishing Co.

Krovetz, R. (1991). Lexical acquisition and information retrieval. In Zernik, U. (Ed.). Lexical acquisition: exploiting on-line resources to build a lexicon. Hillsdale, NJ: Lawrence Earlbaum.

Liddy, E.D. & Myaeng, S. H. (1993). DR-LINK's linguistic-conceptual approach to document detection. Proceedings of the First Text Retrieval Conference. NIST.

Liddy, E.D. & Paik, W. (1992). Statistically-guided word sense disambiguation. In Proceedings of AAAI Fall '92 Symposium on Probabilistic Approaches to Natural Language. Boston.

Liddy, E.D., Paik, W. & Woelfel, J. (1992). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. Proceedings of 3rd ASIS Classification Research Workshop.

McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems. Final report to National Science Foundation. Syracuse, NY: Syracuse University.

Meteer, M., Schwartz, R. & Weischedel, R. (1991). POST: Using probabilities in language processing. Proceedings of the Twelfth International Conference on Artificial Intelligence. Sydney, Australia.

Paik, W., Liddy, E.D., Yu, E.S. & McKenna, M. (In press). Extracting and classifying proper nouns in documents. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March, 1993.

Sager, W.K.H. & Lockemann, P.C. (1976). Classification of ranking algorithms. International Forum on Information and Documentation. 1(4):2-25, 1976.

Slator, B. (1991). Using context for sense preference. In Zernik, U. (Ed.). Lexical acquisition: exploiting on-line resources to build a lexicon. Hillsdale, NJ: Lawrence Earlbaum.

Sowa, J. (1984). Conceptual Structures: Information Processing in Mind and Machine. Reading, MA: Addison-Wesley.

Walker, D. E. & Amsler, R. A. (1986). The use of machine-readable dictionaries in sublanguage analysis. In Grishman, R. & Kittredge, R. (Eds). Analyzing language in restricted domains: Sublanguage description and processing. Hillsdale, NJ: Lawrence Earlbaum.

Ward, J. (1963). Hierarchical grouping to optimize an objection function. Journal of the American Statistical Association. 58, p. 237-254.