INFORMATION FILTERING AND RETRIEVAL: Overview, Issues and Directions BASIS FOR A PANEL DISCUSSION *

Moderator: Nicholas DeClaris † ,
Members: Donna Harman ‡ , Christos Faloutsos § , Susan Dumais ¶ and Douglas Oard ‖

Abstract

This paper is intended to serve as a springboard for a panel discussion with audience participation on information filtering and retrieval. Medical informatics is an emerging specialty which links medicine and information technology. With the unprecedented availability of digital information, retrieval and filtering are becoming important aspects of medical informatics. An overview of medical applications for filtering and retrieval is provided, and important state-of-the-art techniques are introduced in a way that does not presuppose prior knowledge in this field. With this basis for understanding issues and research developments, the ensuing discussion will examine challenges and opportunities for workers in this field.

1 Introduction

Information filtering and retrieval are emerging as important aspects of medical informatics. Recent advances in this area are reviewed here and discussed from a focused perspective with an emphasis on medical applications which are highly relevant for medical practitioners, biomedical engineers and health science educators. This focus is provided by our selection of the approaches included in this review. Our intention is to motivate our panel discussion and encourage audience participation on the following issues:

- State of the art information filtering and retrieval techniques, with emphasis on performance evaluation for very large collections
- Emerging filtering and retrieval techniques for images and multilingual texts

Recent advances in information technology have accelerated the importance of information retrieval from large databases. The increased storage capacity has fueled the growth of databases to gigantic sizes. Simultaneously, the emerging international information infrastructure is producing dramatic increases in available bandwidth at significantly reduced cost per bit. These two factors have combined to make the problem of efficiently selecting the interesting pieces of information more important than ever. Achieving this efficiency will require both computationally efficient search engines and effective, easily managed intelligent user interfaces. Although our presentations will focus on the search engine, the discussion which follows will likely raise several interesting issues with regard to the user interface.

Retrieval and filtering are two extremes on a continuum. In information retrieval user's queries may vary significantly during a single session, while the collection of information to be searched is relatively static. In information filtering, a user's interests are relatively static, but those interests are matched against a dynamic information stream, such as newly published medical journals or conference proceedings. Common to both is the need for techniques to select relevant information. It is those techniques on which we will focus our discussion. For that reason, we shall use retrieval in it's most inclusive sense. Only in those cases where we wish to contrast some aspect of filtering and retrieval will we distinguish the two.

While medicine has scientific aspects, the delivery aspects of medicine (diagnosis, therapy and evaluation) all have strong empirical components. Traditionally, the field of information retrieval has focused on text information. Language is the ultimate form in which humans express ideas, and text is the form in which language is most easily handled by present computer systems. Large amounts of this information are stored in text form. Three basic approaches have been developed to represent texts for

*Individual sections are marked with the author's names.

†School of Medicine and College of Engineering, University of Maryland, College Park, MD 20742, declaris@glue.umd.edu

‡Computer Systems Laboratory, NIST, Gaithersburg, MD 20899, harman@magi.ncsl.nist.gov

§Department of Computer Science and Institute for Systems Research, University of Maryland, College Park, MD 20742, christos@cs.umd.edu

¶Information Science Research Group, Bellcore, 445 South Street, Morristown, NJ 07960, std@bellcore.com

‖Department of Electrical Engineering, University of Maryland, College Park, MD 20742, oard@glue.umd.edu

information retrieval: boolean representations, vector representations, and statistical representations. We will concentrate on vector representations during this panel, but the other representations have also been shown to be effective.

Images, audio, and video are gaining importance as increased processing power and storage make it practical to process these types of information. Some techniques originally developed for processing text can be applied or extended to other media as well. We will describe the application of two such techniques, relevance feedback and latent semantic indexing, to still images. Because they presently represent the most tractable information retrieval problems, text and images will receive the greatest attention in this panel.

One important requirement for our retrieval engines is that they have to be robust in the presence of noise. In text, such noise could be introduced by, for example, typographical errors, optical character recognition errors, or simply a particular users choice of words. In images, this noise could result from artifacts, calibration errors, etc. Practical solutions require robust retrieval algorithms which are able to search for approximate but acceptable solutions. The latent semantic indexing method we describe below is among the most promising techniques we know with these characteristics, not only for text but also for images (using the Karhunen-Loeve transformation) and possibly for other signals as well.

The next section begins with and overview of the state of the art in text retrieval. The results from TREC in that section offer an excellent summary of resent work in the field. In the third section, latent semantic indexing is discussed, together with its mathematical foundations and its overall advantages for retrieval. The fourth section focuses on innovations that offer particular promise to medicine. It begins with a discussion of retrieval by image content, which has the potential for the same sort of impact on image retrieval that the relational data model has had on Database Management Systems (DBMS). We then turn to the potential for multilingually searchable text information filtering and retrieval systems. As medical practice and research becomes increasingly international, access to documents in multiple languages is becoming imperative. In that section we present three approaches to this problem, including an extension of the latent semantic indexing technique.

Finally, the paper concludes with a short commentary that will be the beginning point of the panel. No doubt, during the ensuing discussion, recommendations and suggestions will come from the discussion between the panel members and the audience, which will be recorded for future presentations. A number of important engineering issues must be resolved in order to advance the state of the art in information retrieval and filtering, both in design and implementation. We trust that the biomedical engineering audience will find it a promising research and development area.

2 Text Retrieval and TREC

Donna Harman

Text retrieval is defined as the matching of some stated user query against useful parts of free text records. These records could be any type of mainly unstructured text, such as bibliographic records, newspaper articles, or paragraphs in a manual. The user queries could represent a fixed interest, i.e. a filter or profile that is used against an incoming text stream, or could represent a new interest, i.e. an ad hoc query posed against archived text.

The majority of the commercial retrieval systems are based on exact pattern matching of terms in queries with terms in text. They require queries that range from simple Boolean expressions using a few “ANDs” and “ORs” between terms to complex pattern matching expressions using proximity operators, nested expressions, etc. This type of input requires some user skill, except for very simple information needs.

There has been considerable research into partial matching systems that allow the user to input natural language statements, such as a sentence or a phrase, and retrieve a list of records ranked in order of likely relevance. These systems use statistical methods to automatically match natural language user queries against documents [1]. Evaluation of these systems on small test collections has shown that these statistical techniques are generally very effective, but large-scale commercial implementation of these methods has been slow, partially because of lack of proof that these methods scale up to real-world retrieval environments.

In 1992 a new test collection was built to address this problem. The TIPSTER collection [2] was built at the National Institute of Standards and Technology, and initially contained over 750,000 documents (records) and 100 test queries. Traditional test collection evaluation of text retrieval requires that systems retrieve sets of documents from a fixed document collection in response to test queries. The results from this retrieval are evaluated against the set of correct answers (right documents) that have been determined based on (usually) manual relevance judgments. The creation of this new collection was sponsored by the Advanced Research Projects Agency (ARPA) and was done in conjunction with a project (TIPSTER) which involved four contractors building new retrieval algorithms. These contractors were joined by 24 other research groups in a new workshop to jointly evaluate retrieval systems. This workshop, the Text REtrieval Conference (TREC), was first held in November 1992.

A second conference was subsequently held in August of 1993, and a third conference is currently underway. Many more systems are participating and the test collection has continued to grow. There has been an open exchange between research groups in universities and research groups in commercial organizations, leading to improved commercial products. The use of a larger test collection has also lead into interesting new research issues for all groups.

One of the most exciting parts of TREC is the demonstration of a wide range of different approaches to retrieval of text. Whereas some of the systems are based on traditional vector space or probabilistic approaches, there are also N-gram systems, systems using sophisticated natural language processing techniques, etc. This range is partially described by looking at the sample of the top performing systems from TREC-2 shown in table 1. For full descriptions of these systems, see the TREC-2 proceedings [3]. This session includes a talk giving more detail about one of these systems (the Latent Semantic Indexing system from Bellcore).

3 Latent Semantic Indexing

Susan Dumais

Latent Semantic Indexing (LSI) is a novel information retrieval method which can improve people's access to electronically available textual materials by 20-30% compared with popular word-matching methods [4]. A problem with word-based retrieval is that different people often use different words to describe the same idea or concept. This means that relevant information is missed when a searcher uses different words than an author. LSI helps to overcome the problem of variability in human word choice by automatically organizing textual information into a semantic structure more appropriate for information retrieval. One important consequence of using LSI to organize and access information is that users can retrieve documents that do not share any words with their query.

The particular latent semantic indexing analysis that we have tried uses singular-value decomposition, a statistical technique closely related to eigenvector decomposition and factor analysis. We take a large matrix of term to document association data and decompose it into a set of, typically 100-300, orthogonal factors from which the original matrix can be approximated by linear combination. These factors represent the important "concepts" in the original term by document matrix. We use these derived concepts, rather than the raw word overlap for retrieval.

More formally, any rectangular matrix, for example the $t \times o$ matrix of terms and objects, X, can be decomposed into the product of three other matrices:

$$\underset{t\times o}{X} = \underset{t\times r}{T_0} \cdot \underset{r\times r}{S_0} \cdot \underset{r\times o}{O_0{}'},$$

such that T_0 and O_0 have orthonormal columns, S_0 is diagonal, and r is the rank of X. This is the so-called *singular value decomposition* of X.

If only the k largest singular values of S_0 are kept along with their corresponding columns in the T_0 and O_0 matrices, and the rest deleted (yielding matrices S, T and O), the resulting matrix, $\hat{X}$, is the unique matrix of rank k which is closest in the least squares sense to X:

$$\underset{t\times o}{X} \approx \underset{t\times o}{\hat{X}} = \underset{t\times k}{T} \cdot \underset{k\times k}{S} \cdot \underset{k\times o}{O'}.$$

The idea is that this matrix, by containing only the first k independent linear components of X, captures the major associational structure in the matrix. Since the number of dimensions in the reduced model (k) is much smaller than the number of unique words (t), minor differences in terminology are ignored. As a result, terms that did not actually appear in an object may still end up close to it, if that is consistent with the major patterns of association in the data. Position in the space then serves as the new kind of "semantic indexing".

The LSI analysis involves a completely automatic statistical analysis. Because no human effort is required for knowledge engineering, the LSI method is widely applicable to different domains. The LSI method can be used both for information retrieval and information filtering applications [5]. It has also been used successfully for cross-language retrieval [6], and as a way of suggesting new indexing terms for conventional retrieval methods.

Several experiments using LSI for information filtering will be summarized. An important issue in this work is how best to represent people's interests - e.g., the words in their self-described interests; information about relevant and/or irrelevant documents; etc. One series of experiments used LSI to predict which technical memos would be of interest to researchers. Knowing four relevant documents was as effective as a list of 25 words and phrases in describing peoples' interests. Combining information from several models also improved the selection of relevant materials. In more recent experiments, LSI has been used with the large, diverse TREC collection (Harman, this session) for information filtering. Again, knowledge of previously relevant items was the single best predictor of which new documents would be of interest.

4 Promising Directions of High Relevance to Medicine

Nicholas DeClaris, Christos Faloutsos, Douglas Oard

4.1 Innovations in Image Retrieval

A highly promising effort is in the direction of fast searching on large collections of medical images. For example, consider a collection of 2-dimensional X-ray images or 3-dimensional Magnetic Resonance Imaging (MRI) brain scans, along with demographic data (gender, age group, race etc.) and diagnoses. On this database it would be very useful to have a system which will permit the following types of queries (in increasing order of difficulty):

- **similarity search:** e.g., *find all brain scans that are similar to Smith's brain scan and report the respective diagnoses.*
- **sub-pattern matching:** e.g., *find all (3-dimensional) images that contain tissue with tumor-like texture.*
- **hypothesis testing:** e.g., *is there any correlation between enlarged right lobe and epilepsy?*
- **rule discovery:** e.g., *report the strongest correlations among image features, demographic characteristics and symptoms.*

Such a system would be a convenient tool in medical research, in medical teaching and as a diagnostic aid. For example, medical researchers would be able to retrieve images that look 'similar' to a given image, examine the rest of the associated data (demographic data etc.), form hypotheses and develop theories. Similarly, medical students or non-specialists could use the system to find similar images, examine the associated diagnoses, and thus obtain some help to make a diagnosis.

Such a system could be based on 'similarity search' queries. In this effort, the technical challenges are the following:

- Feature extraction functions that produce feature spaces which can be handled by known indexing methods. Feature spaces with high dimensionality are particularly difficult to deal with.
- Deriving distance functions that capture the perceived similarity between images. The Euclidean distance (pixel-by-pixel sum of squared differences) between two images is but a first step in this direction. More elaborate distance functions which use warping or 'registration' of images to align the corresponding organs in the two images will be required.
- Fast response: This is difficult to achieve because each image is large, of the order of megabytes. In a database of thousands of such images, scanning each image sequentially and comparing it to the query image would be prohibitively slow. High performance spatial access methods will be required.

Fast searching among images is possible through the use of feature-extraction functions that map an image into a point in feature space. Feature extraction functions could be, e.g., the strongest coefficients of the Discrete Fourier Transform (DFT) of the image [7], or some other transform (e.g., the Discrete Cosine transform(DCT), the wavelet transform etc.) or other features that medical doctors will indicate. Then, well-known database methods for searching, such as the so-called *R-trees*, can be used. Figure 1 illustrates the approach.

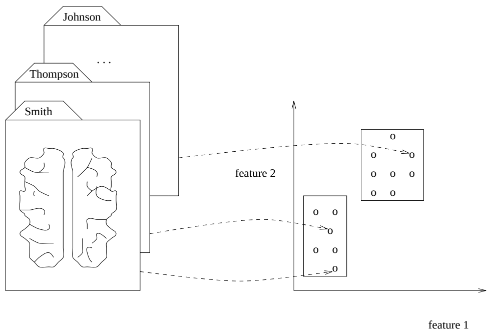

Figure 1: Illustration of our approach: Images are mapped to points in feature space

We have some experience with feature extraction as a result of our work on the IBM-Almaden Query By Image Content (QBIC) project. There, the goal was to retrieve color images by content, for example, *find all images that have similar colors with this photograph of a sunset above the ocean.* The system handles queries on color, shape and texture [8]. For color, the average amount of red, green and blue is used; for shapes the area, perimeter and twenty moments of the (manually outlined) shape are used. For texture, we used coarseness, directionality and contrast. Thus, every image and shape was mapped into some point in the feature space. In [9] it was shown that these features indeed capture the similarity among images, and that the indexing approach with R-trees outperforms sequential scanning, as expected. It was also shown that reducing the dimensionality of the feature space using the Karhunen-Loeve transform (a technique similar to LSI) improved retrieval performance.

On the database end, improvements over the R-trees have been designed. R-trees manage multidimensional points (and rectangles), by grouping nearby points to form 'parent nodes'; parent nodes are represented by their Minimum Bounding Rectangle (MBR), and they are grouped recursively, to form grand-parent nodes, etc. R-trees achieve performance speed-ups on queries, because large parts of the tree can be discarded if the MBRs do not intersect the query region. Figure 2 illustrates some 2-dimensional points (black squares) organized into R-tree nodes: 'A', 'B' and 'C' are parent nodes and 'D' is a grand-parent node. In the same figure, the dotted circle 'Q' indicates a query around the target point 'q0' (white square); notice that the query need only examine the contents of

'B', because the two other parent nodes do not even intersect the query region.

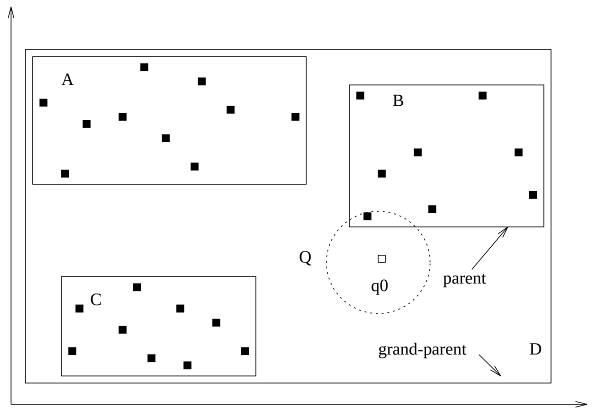

Figure 2: Data organized in an R-tree

We have designed methods that use fractals (specifically the Hilbert curve) to cluster points into parent nodes [10], and have shown that they outperform traditional R-trees. We have also been studying methods to handle the 'dimensionality curse' of R-trees and most other spatial access methods (quadtrees etc.). Although these techniques work well for 2-dimensional and 3-dimensional spaces, in medical images the number of features (i.e. dimensions) may be in the tens or hundreds. The performance of R-trees and other similar methods suffer when applied to problems with high dimensionalities, usually exploding the time complexity exponentially. We have designed the so-called *TV-tree* (Telescopic-Vector tree) [11] to solve this problem. To construct a TV-tree we choose a few of the most important dimensions first, using additional dimensions only if it is necessary.

4.2 Multilingual Text Retrieval

Unlike images, texts are only accessible to users who are able to understand the language in which the text is expressed. Translation is required before users who are not fluent in that language can make use of the information. Despite this limitation, language-independent text representation and selection techniques make retrieval of texts possible without regard to their source language. This is what H. Nevil called a "multilingually searchable system" at the Second European Conference on Information Systems and Networks. In this way we can dedicate scarce and expensive translation resources to texts that are likely to satisfy the user's information requirements. We have proposed three approaches to the construction of a multilingually searchable system: text translation, term vector translation, and latent semantic coindexing [12].

The approach we call text translation involves transforming each term in the source language into a unique form useful for retrieval. Multilingual thesauri have shown good results in multilingually searchable information retrieval systems [13]. A multilingual thesaurus is a mapping from each term in the text's source language to one or more members of a set of language independent concepts. This mapping is usually manually constructed and is designed to optimize performance in a specific application domain. When the mapping is one-to-many, some technique is required to determine the sense in which each term is being used. Restricting the application domain minimizes the need for this term sense disambiguation, though. Queries can be similarly mapped and then relevance determinations can be made in the manually constructed concept space.

With this thesaurus-based text translation technique, errors in term sense selection can adversely affect performance. Our term vector translation approach eliminates the requirement for term sense disambiguation by mapping each term in the source language to a distribution on terms in a preselected target language. This requires a multilingual lexicon which specifies a distribution on the possible translations for every term. Methods for automatically collecting similar statistics from a large parallel bilingual corpus have been developed for statistical machine translation [14]. For each term in a source language term frequency vector, the frequency of that term and the distribution on its translations are used to compute the expected frequency of terms in the target language. The complete target language term frequency vector is formed by summing the expected frequency of each term in the target language over every term in the source language.

In cases where one word has several possible translations we expect that some "spreading" of the distribution represented by the term frequency vector will occur as each term is mapped to a distribution on terms in the target language, reducing the structural information available for a document selection method to exploit. While this could adversely affect retrieval performance, by avoiding some of the problems which can be caused by incorrect term sense selection in the text translation approach we may achieve an offsetting performance improvement.

Term vector translation requires that every term frequency vector be mapped into a single language. That requirement is undesirable if the texts are evenly distributed among several languages. Furthermore, the conditional distributions we construct for term vector translation exploit relatively little of the term cooccurrence information that is present in a parallel multilingual training corpus. We believe that we can eliminate those limitations with an approach we call latent semantic coindexing.

In latent semantic coindexing we first perform a singular value decomposition on a training collection of multilingual documents in which each document contains several versions of a single text, one for each language. By eliminating small singular values which correspond to term usage variations we expect that cross-linguistic term usage variations will be suppressed as well. Term vectors can then be transformed into concept vectors using the approach developed for LSI, regardless of their source language. We expect that similar articles in different languages would be transformed directly into similar concept vectors.

Latent semantic coindexing suffers from the same limitations as latent semantic indexing when used for routing texts. If suboptimal alignment of the concept space develops as topic shifts occur, regeneration of a suitable parallel multilingual corpus and expensive recomputation of the singular value decomposition will be required. Similar limitations exist for the other two techniques that we have described for building multilingually searchable systems, however, and the demonstrated performance of LSI leads us to believe that the performance improvement may justify the computational costs.

5 Future Directions

Medicine traditionally has relied on specialized information and knowledge exchanged directly between people. The strength of this approach is that practitioners learn from the original sources. A serious weakness is that it limits the availability of information. As the volume of information continues to increase, expert panels and consensus-based forums have attempted to sift through the increasing number of medical publications in search of the information most relevant to practitioners in their fields. It has been widely recognized that this process is both slow and expensive, and practical automated approaches are required.

Text filtering and retrieval offer a potential to transcend these limitations. Many issues still remain to be addressed, however. One broad area is work on better algorithms to handle the many types of full text documents becoming available in electronic form, including complete books and technical manuals. A second area is the construction of new tools to support end users. Many current tools are designed for use by intermediaries, and are not very useful for untrained searchers.

Learning algorithms capable of modeling changes in interests over time are needed, and the use of non-content sources of information (e.g., which other people have interests similar to mine) to find relevant materials appears to be a promising area for further research. Further development of multilingually searchable systems will provide a basis for truly multilingual systems in which every documents can be presented in the user's preferred language regardless of the language in which it was expressed.

Historically, most information retrieval research has focused on text. As it becomes more practical to manipulate images, audio and video with information technology, the importance to medicine of retrieval techniques for these media will increase markedly. We have described how some of the techniques originally developed for text retrieval can be applied or extended to these media. Further development of specialized methods will be required to achieve improved performance.

In conclusion, the very practice of medicine requires that we automate information filtering and retrieval for the medical practitioner. The increasing penetration of information technology is beginning to offer the potential to do this effectively. Applying these techniques to medical applications will require a broad, interdisciplinary effort. Both the public and private sectors are investing in this effort. Each brings a unique set of resources, and both will benefit from the resulting capabilities. Merging medical and engineering expertise, biomedical engineers are well positioned to contribute to, and to exploit advances in, this dynamic field.

Acknowledgments

The work of N. DeClaris and D. Oard has been supported in part by NIH grant 1S10RR06460-01 (Medical Informatics Network) and LOGOS Corporation. The work of C. Faloutsos has been partially supported by NSF awards IRI-9205273 and IRI-8958546, with matching funds from Empress Software, Inc. and Thinking Machines, Inc.

References

[1] N.J. Belkin and W.B. Croft. Retrieval techniques. In M. Williams, editor, *Annual Review of Information Science and Technology*, pages 109–145. Elsevier Science Publishers, New York, NY, 1987.

[2] D. K. Harman, editor. *The First Text Retrieval Conference (TREC-1)*, Gaithersburg, MD, March 1993. NIST. Special Publication 500-207.

[3] D. K. Harman, editor. *The Second Text Retrieval Conference (TREC-2)*, Gaithersburg, MD, March 1994. NIST. Special Publication 500-215.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, September 1990.

[5] Peter W. Foltz. Using latent semantic indexing for information filtering. In Frederick H. Lochovsky and

Robert B. Allen, editors, *Conference on Office Information Systems*, pages 40–47. ACM, April 1990.

[6] Thomas K. Landauer and Michael L. Littman. A statistical method for language-independent representation of the topical content of text segments. In *Proceedings of the Eleventh International Conference: Expert Systems and Their Applications*, volume 8, pages 77–85. Avignon France, May 1991.

[7] Rakesh Agrawal, Christos Faloutsos, and Arun Swami. Efficient similarity search in sequence databases. In *Foundations of Data Organization and Algorithms: 4th International Conference, FODO '93*, Evanston, Illinois, October 1993.

[8] Wayne Niblack, Ron Barber, Will Equitz, Myron Flickner, Eduardo Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. The QBIC project: Querying images by content using color, texture and shape. *SPIE 1993 Intl. Symposium on Electronic Imaging: Science and Technology, Conf. 1908, Storage and Retrieval for Image and Video Databases*, February 1993.

[9] Christos Faloutsos, William Equitz, Myron Flickner, Wayne Niblack, Dragutin Petkovic, and Ron Barber. Efficient and effective querying by image content. *Journal of Intelligent Information Systems.* To appear.

[10] Ibrahim Kamel and Christos Faloutsos. On packing R-trees. *Second International Conference on Information and Knowledge Management (CIKM)*, November 1993.

[11] King-Ip Lin, H.V. Jagadish, and Christos Faloutsos. The TV-tree - an index structure for high-dimensional data. *Very Large Databases (VLDB) Journal*, 1993. To appear.

[12] Douglas W. Oard, Nicholas DeClaris, Bonnie J. Dorr, and Christos Faloutsos. On automatic filtering of multilingual texts. In *Conference Proceedings, 1994 IEEE International Conference on Systems, Man and Cybernetics.* IEEE, October 1994. To appear.

[13] A. Steven Pollitt and Geoff Ellis. Multilingual access to document databases. In *21st Annual Conference Canadian Society for Information Science*, pages 128–140, July 1993.

[14] Peter F. Brown, Steven A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

Organization	Retrieval Technique
U Mass at Amherst	Uses probabilistic term weighting and a probabilistic inference net to combine various topic and document features.
Cornell	The basic SMART system using a vector space model.
Univ. of Dortmund, Germany	Uses polynomial regression on the training data to find weights for various pre-set term features.
UC Berkeley	Uses logistic regression analysis to learn optimal weighting for various term frequency measures.
CLARIT Corp.	Expands each topic with noun phrases found in a thesaurus that is automatically generated for each topic.
Bellcore	Uses latent semantic indexing to create smaller vectors than the more traditional vector-space models.
Siemens Corporate Research	Uses the Cornell SMART system, but with the topics manually expanded using WordNet.
Virginia Tech	Combines the results from SMART vector-space queries with the results from manually-constructed soft Boolean P-Norm queries.
ConQuest Software	Uses a very large general-purpose semantic net to aid in constructing better queries from the topics, along with sophisticated morphological analysis of the topics.
Verity Corp.	Uses an expert system working off specially-constructed knowledge bases to improve performance.
TRW Corp.	Uses an adaptation of their Fast Data Finder pattern matching system to allow term weighting.
NYU	Uses intensive natural language processing techniques including a full part of the documents to locate syntactic phrases.

Table 1: Top-performing TREC-2 participants