On Automatic Filtering of Multilingual Texts *

Douglas W. Oard and Nicholas DeClaris
Department of Electrical Engineering
and
Bonnie J. Dorr and Christos Faloutsos
Department of Computer Science
University of Maryland, College Park, MD 20742

Abstract

An emerging requirement to sift through the increasing flood of text information has led to the rapid development of information filtering technology in the past five years. This study introduces novel approaches for filtering texts regardless of their source language. We begin with a brief description of related developments in text filtering and multilingual information retrieval. We then present three alternative approaches to selecting texts from a multilingual information stream which represent a logical evolution from existing techniques in related disciplines. Finally, a practical automated performance evaluation technique is proposed.

1 Introduction

Automatic filtering of information from text sources has become increasingly important as the volume of electronically accessible texts has exploded in recent years. Among these sources of electronically accessible texts are news stories, journal articles and electronic discussion forums. Since it would be impractical for any individual to examine every available source to determine whether interesting information was present, some form of information filtering is required. Information filtering systems are designed to sift through large quantities of dynamically generated texts and display only those which may be relevant to a user's interests.

As advanced technology continues to reduce the expense of international communications, the value of information for business, government and personal use will become less sensitive to its location. As a result, it has become increasingly necessary to enable appropriate handling of multilingual information in text processing systems. Although there is increasing interest in filtering multimedia information, the focus of our research is on text filtering because of the quantity of text being generated, the fact graphic information is often accompanied by text captions, and because a large body of text manipulation techniques has been developed for information retrieval applications.

*This work has been supported, in part, by NIH grant 1S10RR06460-01 (Medical Informatics Network), NSF awards IRI-9357731, IRI-9205273, and IRI-8958546 and Logos Corporation, Empress Software, Inc., and Thinking Machines, Inc.

2 Background

The earliest commonly cited reference in which the information filtering problem was described is the ACM President's Letter by Peter Denning from the Communications of the ACM of March 1982. Denning's objective was to broaden the discussion which had traditionally focused on generation of information to include reception of information as well. In the paper he describes the need to filter information arriving by electronic mail in order to separate urgent messages from routine ones and restrict the display of routine messages in a way that matches the personal mental bandwidth of the user. Among his approaches to achieving that goal is a "content filter." The term "information filtering" is now commonly identified with the idea of text selection based on content.

Over the subsequent decade, occasional papers reporting the performance of a variety of information filtering applications have appeared. While electronic mail was the original domain about which Denning had written, subsequent papers have addressed newswire articles, USENET news articles, technical reports, and broader network resources. A significant contribution to the emergence of information filtering as a research area has been the five Message Understanding Conferences (MUC) sponsored by the Defense Advanced Research Projects Agency (DARPA)

(now known as ARPA) each year since 1989 [3]. The principal thrust of the MUC project has been the application of natural language techniques to the extraction of specific information from messages. In 1990 DARPA established the TIPSTER project to fund the research efforts of several of the MUC participants [8]. TIPSTER added an emphasis on the use of statistical information retrieval to preselect messages for natural language processing.

In November of 1991 Bellcore and ACM SIGOIS jointly sponsored a workshop on "High Performance Information Filtering" that brought together a substantial quantity of research to establish a basis for more rapid development of the field. A group of significant papers from that workshop were published in a special issue of the Communications of the ACM in December, 1992 [1, 7]. In 1992 the National Institute of Standards and Technology (NIST) capitalized on DARPA's experience with TIPSTER by cosponsoring an annual Text Retrieval Conference (TREC) focused on information filtering and retrieval [9].

2.1 Terminology

The term "Information Retrieval" has been commonly applied to the retrieval of relevant information from relatively static collections that are not amenable to organization as a relational database. When retrieval of text information is intended, the term "text retrieval" is often used.

The terminology in information filtering is less well standardized. A closely related concept is "routing." In routing the goal is to ensure that each document is routed to at least one member of a specified group. Routing is a special case of the more general concept of information filtering since in information filtering we allow for the possibility that some documents may not be selected for display at all. Sometimes the term "selective dissemination of information" is used to refer to information filtering, although other authors use that term to refer to routing.

In information filtering applications the specification of a user's interests is often referred to as a "profile." Each profile may include several interest specifications, which are analogous to the queries posed by the user in an information retrieval context. In an analogy to the function performed by a newspaper editor, the texts to be filtered are frequently referred to as "articles" and the text information stream itself is called "news." Although filtering can also be applied to journal articles, technical reports, and many other types of documents, we have adopted this "news" terminology in the discussion which follows.

2.2 Multilingual Retrieval

Nevil defines three types of text retrieval systems: monolingual systems, multilingually searchable systems, and fully multilingual systems [11]. Some so-called multilingual information retrieval systems are actually a collection of monolingual systems that are all developed using the same tools. Such systems allow the user to choose among several languages, but once that choice is made, both the query and the documents are restricted to that language. Our use of the term "multilingual" corresponds to Nevil's concept of a multilingually searchable system in which the query may be expressed in a language different from that of the document. A fully multilingual system, in which the user may choose to display any document in any language, could in principle be built upon the foundation of such a multilingually searchable system through automatic machine translation of the selected documents.

Interest in information retrieval grew out of the increasing ability to archive enormous amounts of information as media costs decreased and communications capacity increased. Much of the research in information retrieval has been conducted from the perspective of library science. Salton began the study of multilingual information retrieval twenty five years ago. In those experiments he used a manually constructed multilingual thesaurus to map terms from either English or French documents to a common set of concepts. Natural language queries were similarly mapped and his SMART vector space information retrieval system was used to select appropriate documents. The reported results indicate that retrieval performance is not affected significantly by the choice of query language. Salton later found the same result for documents accessed by queries in English and Russian [14].

Subsequent researchers have applied essentially the same approach, although increasingly sophisticated algorithms and user interfaces and rapid advances in hardware capabilities have significantly improved performance. Recent results from the European Multilingual Information Retrieval project are reported in [13]. Pollitt and Ellis argue forcefully for the effectiveness of techniques based on a multilingual thesaurus [12] and Nelson has integrated thesaurus-based retrieval with machine translation to develop a fully multilingual information retrieval system for English readers of Japanese texts [10].

3 Multilingual Filtering

The objective of multilingual information filtering is to evaluate articles from an information stream and select those which are relevant to a user's interest regardless of their source language. Once a relatively small set of relevant articles is identified, scarce translation resources can be devoted to only those articles. We have developed three approaches to multilingual information filtering: text translation, term vector translation, and latent semantic coindexing.

3.1 Text Translation

The goal of text translation is to transform each term in the source language into a form useful for retrieval. The multilingual thesauri which have shown good results in multilingual information retrieval are one example of such a technique. A multilingual thesaurus maps each of the terms in a source language to one or more members of a set of manually selected concepts in such a way that the concept representation is language independent. When the mapping is one-to-many, word sense disambiguation is required to determine which mapping should be chosen for a specific instance of a word. Queries are similarly mapped, and the relevance determination is made in the manually constructed concept space.

Machine translation offers a more sophisticated approach to text translation. Each arriving article can be automatically translated into the user's preferred language. Although present machine translation systems may not produce perfect translations, the sophisticated linguistic analysis required for machine translation may result in better input to the filtering process than could be obtained using a multilingual thesaurus.

Many machine translation systems generate an intermediate representation of the text and then generate the translated text from that intermediate representation. When the intermediate representation is independent of the specific language pair selected it is known as an interlingua [5]. A sophisticated data structure is frequently used to preserve both structural information and the unresolvable ambiguity that must be known to accurately generate the target language output. It is possible to extract the words (or word senses) from the data structure, but it may be more effective to exploit other aspects of the representation to add a natural language component to the filtering process. If an interlingual representation could be exploited effectively to improve relevance determinations, a variety of source and user languages could be accommodated seamlessly.

3.2 Term Vector Translation

Errors in word sense selection can adversely affect filtering performance when text translation is employed. If, instead of transforming each word in the source language into a unique word in the target language, we were to transform it into a distribution on words in the target language then the performance degradation associated with incorrect word sense selection might be mitigated. We call this approach term vector translation. While text translation can be used with almost any filtering technique (boolean keyword matching, vector space mapping, inference networks, etc.), the term vector translation approach will work only with vector space mapping techniques.

Vector space mapping is an approach originally developed for information retrieval technique which provides a basis for ranking documents by their degree of similarity with a query. In the vector space technique the set of terms in a document is represented as a vector, where each component of the vector is a function of the frequency with which that term appears in the document. Step functions, logarithms and functions which account for the frequency of the term in the entire collection are all commonly used. Indexing terms are often chosen as a subset of the words in the collection of documents. In large collections relatively few of the terms will appear in each document, resulting in sparse term frequency vectors. One commonly used similarity measure is the cosine of the angle between two vectors which is computed as the inner product of two normalized vectors. The vector space technique was introduced in Salton's SMART system [15].

Term vector translation avoids the complexity of the word sense disambiguation required in text translation by using a statistical word-by-word machine translation algorithm in which the frequency of each term in a source language term frequency vector is used to find the expected frequency of terms in the target language term frequency vector. 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.

This approach requires a multilingual lexicon that specifies a distribution on the possible translations for every term. Methods for collecting similar statistics from a large bilingual document collection have been developed for statistical machine translation research [2]. Since term vector translation does not require the complex mathematical model used in statistical machine translation, we expect that the parameter estimation task will be quite tractable if a rep-

resentative parallel bilingual training corpus is available.

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. This would result in a more nearly uniform in the target language than in the source language. As a result, the amount of structural information available for a vector space document selection method to exploit will likely be reduced. While this may adversely affect retrieval performance, we expect that the term vector translation technique will achieve a significant reduction in computational complexity when compared to automatic machine translation. Furthermore, this slight flattening of the distribution may be less harmful to retrieval performance than incorrect selection of a word sense would be when the text translation approach is used.

3.3 Latent Semantic Coindexing

The term cooccurrence information present in a parallel bilingual training corpus can be exploited more directly by an approach we call latent semantic coindexing. Latent semantic coindexing is an extension of Latent Semantic Indexing (LSI), a vector space information retrieval method based on factor analysis which has demonstrated improved performance over the original vector space technique used in Salton's SMART system [4].

The first step in LSI is to perform a singular value decomposition on a set of representative documents and then set the smallest singular values (which are thought to correspond to term usage variations) to zero. This representative collection can be obtained through random selection of documents arriving over an arbitrary period. The singular value decomposition and selection results in a matrix that can be used to map newly arriving articles into a "concept vector" space. These concept vectors are then compared to a set of interest concept vectors that are developed in the same way from natural language specifications of user interests. Cosine similarity measures are used to score the article against each interest in the profile so that a pointer to the article can be added to the "reading list" for each interest in decreasing order of similarity to that interest. Foltz reports results using LSI for filtering news articles in [6] and Foltz and Dumais apply LSI to filtering technical reports in [7].

Both text translation and term vector translation use some form of automatic translation to transform each article into a common representation and then make relevance determinations on what is essentially a monolingual collection of articles. 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 version 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 resulting matrix regardless of their source language. We expect that similar articles in different languages would be transformed directly into similar concept vectors.

Suboptimal alignment of the concept space can, however, become a problem when LSI is used for information filtering. If new articles diverge significantly from the set of concepts in the original collection, the set of singular vectors which represent "typical" term frequency and span the concept space may become less relevant to the newly arriving articles. Periodic recomputation of the singular value decomposition can mitigate this difficulty, however.

4 Discussion

One of the key advantages of multilingual information filtering over a set of monolingual information filtering systems is that all user relevance judgments are available to improve future relevance feedback regardless of the source language of any particular article. Thus, even users able to read multiple languages may benefit from the wider scope of relevance feedback data available from the larger volume of articles. Two issues remain to be addressed: the availability of a parallel multilingual training corpus and a methodology for performance evaluation.

The availability of a multilingual training corpus is a significant issue for both term vector translation and latent semantic coindexing. Several such corpora are available in electronic form, but the relatively narrow scope of some of these corpora could limit filtering effectiveness over broader domains. The Canadian Parliament has produced a collection with relatively broad coverage, although only in a single language pair (French and English) that should be useful for comparing the performance of our three approaches.

Another significant issue is performance evaluation. Two common measures of effectiveness in information filtering and retrieval are recall and precision. Recall is defined as the ratio of relevant texts that are retrieved to the relevant texts that are avail-

able. Precision is defined as the ratio of relevant texts that are retrieved to the total number of texts that are retrieved. During design of an information filtering system it is often possible to increase either recall or precision at the expense of the other, so it is common to report the precision for a range of values of recall.

Although the concepts of recall and precision are well defined, there are two difficulties with their application to information filtering. The most basic problem is that relevance itself is not easily determined because human relevance judgments exhibit significant variability between observers. Furthermore, evaluators sometimes find it difficult to render a binary relevance judgment on a specific combination of a text and a query. The second difficulty has a more practical basis. Although precision can be evaluated by making relevance judgments on the relatively few articles that are selected by a filtering algorithm, relevance judgments must be made for every article in order to evaluate recall. Since the number of articles can grow without bound, some sampling approach is required to develop an estimate of recall.

Precision and recall help to characterize the performance of an information filtering system, but computational complexity is also a significant issue in practical applications. Typically, information retrieval system architectures are optimized for execution of changing queries against relatively stable text corpora. In information filtering the typical situation is just the opposite. Algorithms and data structures must be constructed with this in mind. Furthermore, in some information filtering applications (such as newswire distribution within a newspaper office), the filter must operate in near-real time. Many application environments also experience rapid growth in news volume, placing a premium on techniques which can be applied on progressively larger scales. Thus there is a three dimensional tradeoff in information filtering system design between recall, precision and computational complexity.

While it would be desirable to evaluate the precision and recall of each technique over the same set of articles and then compare their performance. the lack of a scored test corpus over any domain similar to that of an existing parallel multilingual training corpus makes that approach impractical at present. A more feasible evaluation technique is to compare the performance of each new technique with a baseline technique for which the recall and precision have been well established. We will describe such an approach using the parallel English/French corpus produced by the Canadian Parliament. Because two of our approaches require training data, we plan to partition that parallel bilingual corpus into separate training and evaluation sets.

We first compute the baseline set of similarity measures using the baseline technique and an arbitrary query. We then compute a set of similarity measures for each of our three information filtering approaches using that same query. As a measure of similarity between a vector of scores computed using one multilingual filtering approach and the vector of scores computed using our baseline approach we propose to compute the cosine similarity measure between the two score vectors.

For each technique that produces a large cosine (near 1.0) we will conclude that the scores (and hence the ranking) produced by the multilingual technique is very similar to those produced by the baseline technique and therefore similar recall and precision can be expected. Since the performance of the baseline technique is well characterized, such a result would provide useful information about the performance of the the multilingual information filtering approach being evaluated.

If the cosine turns out to be too small we would only be able to conclude that a significantly different ranking would likely be produced by the two techniques. It would not, however, be possible to determine the effect of these differences on recall and precision because it is possible for quite different rankings to generate similar recall and precision. In such cases it would be necessary to obtain a scored corpus and compare the recall and precision of the various approaches directly.

Our text translation approach allows the application of any filtering technique to the translated texts, and term vector translation allows the application of any vector space mapping technique. Since latent semantic coindexing incorporates LSI, we believe that the use of LSI as the filtering technique for other two approaches will facilitate performance comparisons. We also believe that application of LSI to just the English portion of a bilingual test corpus will establish a useful baseline score vector. For this baseline case the singular value decomposition should be computed using only the English language portion of the training corpus to prevent cooccurrence information in the French texts from influencing the choice of singular vectors.

We expect to gain some insight into the set of acceptable values for our similarity measure from a separate experiment on the test portion of the parallel bilingual corpus. Our intent is to determine the variation that results from the manual translation used to

construct the corpus. We will first manually translate the query into French and then construct a vector of scores using LSI on only the French texts. The cosine between this vector and the baseline score vector represents the limiting performance the applying LSI to manual translations can achieve.

5 Conclusion

The explosion of digital information and the importance of that information for both social and commercial purposes motivates our study of multilingual information filtering. The recent development of latent semantic indexing and statistical machine translation provide tools with the potential to achieve adequate performance at moderate cost. At the same time, continuing improvements in machine translation performance have led us to consider application of that technology to multilingual information filtering when performance requirements justify the substantial investment required. We believe that the three approaches we have proposed have the potential to satisfy a broad range of requirements when information in multiple languages must be monitored and disseminated.

References

[1] N. J. Belkin and W. B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, Dec. 1992.

[2] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] DARPA. *Fourth Message Understanding Conference (MUC-4)*. Morgan Kaufmann, June 1992.

[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, 1990.

[5] B. J. Dorr. The use of lexical semantics in interlingual machine translation. *Machine Translation*, 7:135–193, 1993.

[6] P. W. Foltz. Using latent semantic indexing for information filtering. In F. H. Lochovsky and R. B. Allen, editors, *Conference on Office Information Systems*, pages 40–47. ACM, April 1990.

[7] P. W. Foltz and S. T. Dumais. Personalized information delivery: An analysis of information filtering methods. *Commun. ACM*, 35(12):51–60, Dec. 1992.

[8] D. Harman. The DARPA TIPSTER project. *ACM SIGIR Forum*, 26(2):26–28, Fall 1992.

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

[10] P. Nelson. Breaching the language barrier: Experimentation with Japanese to English machine translation. In D. I. Raitt, editor, *15th International Online Information Meeting Proceedings*, pages 21–33. Learned Information, Dec. 1991.

[11] H. Nevil. Session V report of the English language discussion group. In *Second European Congress on Information Systems and Networks*, pages 162–164. Verlag Dokumentation, May 1975.

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

[13] K. Radwan, F. Foussier, and C. Fluhr. Multilingual access to textual databases. In A. Lichnerowicz, editor, *Proceedings of a Conference on Intelligent Text and Image Handling (RIAO 91)*, pages 475–489. Elsevier, Apr. 1991.

[14] G. Salton. Experiments in multi-lingual information retrieval. *Information Processing Letters*, 2(1):6–11, 1973.

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.