Evaluating Semantic Indexing Techniques through Cross-Language Fingerprinting

Eduard Hoenkamp
Nijmegen Institute for Cognition and Information
Montessorilaan 3
6525HR Nijmegen, the Netherlands
hoenkamp@acm.org

Sander van Dijk
Nijmegen Institute for Cognition and Information
Montessorilaan 3
6525HR Nijmegen, the Netherlands
SandervanDijk@student.ru.nl

ABSTRACT

Users in search of on-line document sources are usually looking for content, not words. Hence, IR researchers generally agree that search techniques should be geared toward the meaning underlying documents rather than toward the text itself. The most visible examples of such techniques are Latent Semantic Analysis (LSA), and the Hyperspace Analog to Language (HAL). If these techniques really uncover semantic dependencies, then they should be applicable across languages. We investigated this using electronic versions of three kinds of translated material: a novel, a popular treatise about cosmology, and a data base of technical specifications. We used the analogy of fingerprinting used in forensics to establish if individuals are related. Genetic fingerprinting uses enzymes to split the DNA and then compare the resulting band patterns. Likewise, in our research we use queries to split a document into fragments. If a search technique really isolates fragments related to the query, then a document and its translation should have similar band patterns. In this paper we (1) present the fingerprinting technique, (2) introduce the material used, and (3) report preliminary results of an evaluation for two semantic indexing techniques.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Experimentation, Languages, Performance

Keywords

Semantic indexing, CLIR, Evaluation, Visualization

1. INTRODUCTION

The aim of cross-language information retrieval (CLIR) is to use a query in one language to search a corpus in a different language. Researchers have used various language pairs and compared different IR techniques in search of the most effective approach. An example that can stand for several others is Yang et al.'s [4] study of bilingual corpora. An experiment in the vector space paradigm, it compares traditional IR approaches such as the Generalized Vector Space Model (GVSM), Latent Semantic Analysis (LSA), relevance feedback, and term in context translation. The evaluation is based on the usual recall/precision metrics. In contrast, our present interest is not so much in *which* is the most effective IR system, but *why* is it the most effective. More precisely, we are interested to discern whether a system is successful because it handles the underlying concepts that were communicated, or because it excels in statistical sophistication. Hence, we shift the focus from comparing how well techniques work for CLIR, to using CLIR to compare which ones best handle underlying concepts. If the success of a technique for a corpus can be attributed to its handling of the underlying concepts, then it should be successful for a translation of the corpus as well.

We will describe a technique to assess this invariance under translation, with an illustration from two approaches to IR: the traditional vector space model, and the more recent probabilistic language modeling approach.

2. PARADIGMS IN SEMANTIC INDEXING

An important approach in the vector space model that tries to target the underlying semantics of a corpus is LSA. Empirical evidence suggests that a lossy compression [3] of the high dimensional document space spanned by the terms, will result in a lower dimensional space spanned by 'latent' semantic factors [2]. The technique has met with success in CLIR experiments like the one mentioned above [4]. An approach that even more directly tries to incorporate semantic relationships in the corpus is the 'Hyperspace Analog to Language' (HAL). The method is based on the observation that distance between words in text is an indicator of how related the words are in meaning [1]. The HAL representation is computed by sliding a window over the documents and assigning weights to word pairs, inversely to the distance from each word to every other word in the window. These distances have been mapped to conditional probabilities used for experiments in probabilistic language modeling. The latter is a more recent development in IR, which views documents as samples from a source that stochastically produces terms.

In the remainder we will see how experiments in these completely distinct paradigms can both be accommodated by the fingerprinting technique we are about to describe.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

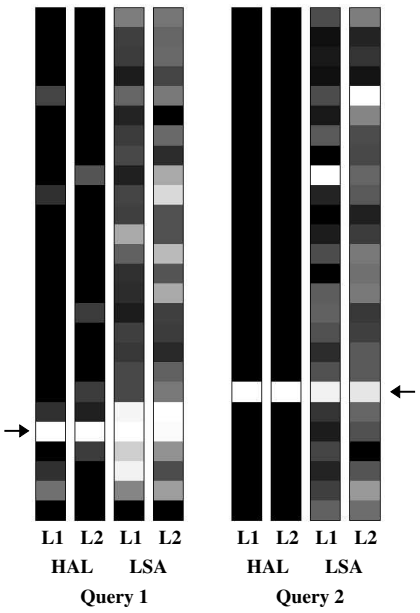

Figure 1: Comparing fingerprints for two queries about "The old man and the sea". L1 is the original, L2 its German translation. For each language pair, two different kinds of semantic indexing were used. Gray scale indicates computed relevance (with white most relevant). The vertical axis is the location of passages in the book (with page 1 at the top).

3. CROSS-LANGUAGE FINGERPRINTING

CLIR experiments in the literature have used multilingual, document-aligned corpora, where documents in one language are paired with their translation in the other. To understand the fingerprinting analogy, imagine the documents of one language stacked on a pile, next to a pile that has the translations in the same order as the original. For a given query, a search technique will assign relevance weights to the documents. These weights can be expressed with a color for each document, from black (not relevant) to white (highly relevant). The pile with original documents will show bands reminiscent of the bands in a DNA fingerprint. If the search technique is invariant under translation, than the bands in the piles should align. Weaker invariance will show up as less overlap in the band pattern.

3.1 Guideline for experiments

An example will sketch the experimental paradigm.
Material. Instead of paired corpora, we use three other kinds of translated material: (1) Hemingway's "The old man and the sea" with translations in German and Italian, (2) Hawking's "A brief history of time" also with German and Italian translations, and the (3) the NVDIA site for graphics cards, which gives their specifications in nine different languages. To illustrate the procedure we take "The old man and the sea" and the languages English and German. The novel was split in parts of about 1000 words each, so that the book and its parts play the roles of the corpus and its documents. We removed the stopwords.
Procedure. We compared the HAL approach and LSA. As queries we selected sentences from the book. The queries used for Figure 1 are located at the arrows. Ranking is done via pseudo-relevance feedback with re-weighting.
Results. Under these conditions, the HAL approach performs better than LSA in terms of band overlap. This shows up especially with Query 2, which we made extremely specific using Hemingway's description of fish eyes: "as detached as the mirrors in a periscope or as a saint in a procession". We do not claim that HAL works better in general than LSA as a search technique, only that under the confined conditions given here, HAL is more invariant under translation than LSA. (This is only meant to explain the role of the fingerprints, not as a generalization.)

3.2 Future experiments

The example just given requires us to select queries by hand, and to visually make the comparison of the results. We are currently investigating metrics to define a distance between band patterns. The analogy would be the kind of comparison made between fingerprints (RFLP's) in case of a paternity dispute. Once we have an appropriate distance measure (Kullback-Leibler divergence is a candidate), we can use a Monte-Carlo method to repeatedly select a random passage from the novel (or other collection), run the comparison, and collect statistics about the distances.

4. CONCLUSION

We have presented a method to compare the efficacy of semantic indexing techniques, based on the assumption that they should be invariant under translation. We have shown how the analogy with DNA-fingerprinting can give an initial assessment of the techniques under study. In our small-scale investigation of two techniques that claim to operate on the semantic level, HAL and LSA, the former looks more propitious. If LSA would give overall better performance in terms of recall/precision, than this would have to be attributed to statistical sophistication more than to the uncovering of semantic relationships. This requires a more thorough experimental design that takes language specific issues into account. For example, German might require splitting up compound nouns (as they are spelled as one word), and for Italian perhaps the window size for HAL has to be increased (as it has a freer word order than English). To design such a robust and reliable experimental paradigm, the next step is to develop a metric to compute a distance between the band patterns produced in the experiment.

5. REFERENCES

[1] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25:211 – 257, 1998.
[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
[3] E. Hoenkamp. Unitary operators on the document space. *Journal of the American Society for Information Science and Technology*, 54(4):314–320, 2003.
[4] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligence*, 103(1-2):323–345, 1998.