From: FLAIRS-01 Proceedings. Copyright © 2001, AAAI (www.aaai.org). All rights reserved.

Assessing the Effectiveness of LSI in Approaching the Intention of a User's Query

Massoud Moussavi[1] and Robert Najlis[2]

The World Bank*[1]
1818 H Street N.W.
Washington, D.C. 20433
mmoussavi@worldbank.org

Computer Science Department[2]
Indiana University
Lindley Hall, Rm. 215
Bloomington, Indiana 47405
rnajlis@cs.india.edu

Abstract

Documents retrieved in response to a user's query should reflect the intention of the user. Keyword searches are not sufficient to accomplish this task. This paper considers Latent Semantic Indexing (LSI) as a possible solution. LSI is considered in the context of a large set of World Bank documents. LSI is shown to be useful in this regard, but not a complete solution. Additional, domain specific, information would be needed as well.

Introduction

The World Bank has a very large repository of case studies and reports on various economic and social development projects undertaken in the developing countries. Staff working on a new development project often have to perform a keyword search to find an example that is similar to their situation. But keyword based searches can miss pertinent documents because those documents do not contain the presented keywords. Furthermore, they might retrieve documents which, although containing the keywords, in fact are unrelated to the actual search (Moussavi 1999). This means that the difficult task of trying to find the correct words to enter for a search is placed on the user. It would be preferable to give as much of the job as possible to the computer instead. The first place to look for solutions is how the knowledge base is indexed. Keywords are an essential building block to any indexing scheme, but they are not enough by themselves. Correlations between the keywords and documents also need to be found. One system for doing this is Latent Semantic Indexing (LSI).

LSI groups words by their relationships to other words in and across documents. This allows the system to place related documents in close proximity, and unrelated ones further away. Thus, although some actual keywords in the documents might differ, if there is enough similarity across documents, they will be grouped together. Furthermore, a search on one of the keywords will also yield documents containing the other related words. LSI works by taking a matrix of terms and documents and analyzing it by singular value decomposition (SVD). Each term and document is indexed by its vector in the matrix (Deerwester, et al. 1990) (Laham 1997). The result is a geometric space correlating terms and documents.

Research

This paper describes findings from an ongoing examination of application of LSI to indexing (and concept generation) a large set of World Bank documents on economic and social development issues. Our research objective is to develop a conversational system (Aha and Breslow 1997) via which the users can retrieve the most relevant documents by answering questions on indexed features. This paper focuses on the efficacy of LSI in the retrieval of documents in a manner which more closely fits with the user's concept for the query. Queries of over 500 abstracts of documents on the topic of economic growth in East Asia, and especially

* The findings, interpretations, and conclusions expressed in this paper are those of the author and should not be attributed in any manner to the World Bank, to its affiliated organizations, to the Board of Directors or the countries they represent.

Korea, were conducted. Each query returned 100 documents. The correlation of the queries and the documents retrieved by LSI will be examined.

Three of the queries that were presented to LSI :

(1) How did East Asian countries achieve their economic growth
(2) How did East Asian countries achieve their growth rates
(3) How did Korea achieve its economic growth

LSI transformed those user entered queries into the following forms:

(1) east asian countries achieve economic growth
(2) east asian countries achieve growth rates
(3) korea achieve economic growth

Thus, all commonly used words were removed from the query before the search was begun. The phrases were not treated as sentences, but rather as a conglomeration of keywords. Queries could be presented in the form of sentences, or even as documents, but LSI would base the search solely on the relevance of the keywords, not by attempting any kind of linguistic parsing of the query.

Results

The documents returned for queries (1) and (2) were quite similar in many ways, but query (2) also included many documents relating to interest rates and other topics which were not quite as relevant to economic growth on the whole. Results of query (3) were highly focused on Korea, at times even to the exclusion of economic growth.

Queries (1) and (2) both returned with document 489 as the most relevant, giving it a weight of 0.740514 and 0.712007 respectively. As can be seen, the document refers to the effect of credit policies in the development of East Asian countries:

> Credit policies : lessons from East Asia -- Directed credit programs were a major tool of development in the 1960s and 1970s. In the 1980s, their usefulness was reconsidered. Experience in most countries showed that they stimulated capital-intensive projects, that preferential funds were often (mis)used for nonpriority purposes, that a decline in financial discipline led to low repayment rates, and that budget deficits swelled. Moreover, the programs were hard to remove. But Japan and other East Asian countries have long touted the merits of focused, well-managed directed credit programs, saying they are warranted when there is a significant discrepancy between private and social benefits, when invesment risk is too high on certain projects, and when information problems discourage lending to small and medium-size firms. The assumption underlying policy-based assistance and other forms of industrial assistance (such as lower taxes) is that the main constraint on new or expanding enterprises is limited to access to credit. The authors give an overview of credit policies in East Asian countries (China, Japan, and the Republic of Korea) as well as India, and summarize what these countries have learned about directed credit programs. Among the lessons: 1) Credit programs must small, narrowly focused, and of limited duration (with clear sunset provisions); 2) subsidies must be low to minimize distortion of incentives as well as the tax on financial intermediation that all such programs entail; 3) credit programs must be financed by long-term funds to prevent inflation and macroeconomic instability, recourse to central bank credit should be avoided except in the very early stages of development when the central bank's assistance can help jump-start economic growth; 4) they should aim at achieving positive externalities (or avoiding negative ones), any help to declining industries should include plans for their timely phaseout; 5) they should promote industrialization and export orientation in a competitive private

This document seems to fit fairly well into the concept of the query, so this is an appropriate retrieval from LSI. Notably though, this document was not considered terribly relevant to query (3), where it was only given a weight of 0.279761, quite a low score, especially given the strong weighting it received for the other two queries. This, despite the fact that Korea is even mentioned explicitly in the document.

Documents 512, 516, and 517 were all considered to be highly relevant to query (3), being given weights of 0.690721, 0.619509, and 0.596581 respectively. All of the documents refer to Korea's economic growth, thus one would expect this to be the case.

Document 512:

> Korea - Problems and issues in a rapidly growing economy -- The phenomenal economic progress of the Republic of Korea during the past decade is one of the outstanding success stories in international development. Despite a lack of natural resources, Korea has made the transition from a largely agricultural society into a semi-industrialized country in a relatively short time. This book traces the impressive industrial growth and discusses the issues that Korea's five-year plan for 1977-81 must try to resolve. It focuses on the problem of mobilizing financial resources - both foreign capital and domestic savings - for future growth and on the problem of increasing industrial exports without undermining efforts

toward import substitution. This study of Korea's plans for industrial expansion emphasizes the textile, electronics, and shipbuilding industries, as well as the contribution that small and medium-size firms can make. It also considers the social goals of more even distribution of the benefits of growth, increased employment, and greater parity between urban and rural incomes.

One would also expect that any documents so relevant to a query on the economic growth of Korea, would also be of value in a query pertaining to the economic growth of East Asia. However, LSI did not weigh these documents heavily for queries (1) and (2). For query (1), they were given weights of 0.406988, 0.204237,and 0.221794, while for query (2) they were given weights of 0.356005, 0.169973, and 0.276461 respectively. Likely, these documents might not even be considered relevant in a search for queries (1) and (2).

In the case of document 504, LSI rated it similarly for queries (1), (2), and (3), 0.414071, 0.370077, and 0.372094 respectively. This despite the fact that the document does not mention East Asia or Korea at any point in the document.

Nepal - 1997 Economic update: the challenge of accelerating economic growth -- This report was prepared for the Nepal Aid Group meeting scheduled to be held in Paris in early 1998. The report focuses selectively on a limited set of key issues critical to accelerating economic growth and improving development management in the short-to-medium term. The report is organized into two chapters. Chapter 1 reviews recent economic developments, focusing particularly on economic growth, fiscal management, the financial sector, and progress in economic reforms. Chapter 2 looks at the development agenda which will accelerate economic growth and development on a sustainable basis. The main elements of the development agenda include: 1) Aim for sustained high rates of broad-based and more equitable economic growth, and make strong efforts to reduce the population growth rate in order to improve the living conditions of the predominantly rural poor. 2) Make concerted efforts to promote human resource development and to provide basic infrastructure and services. Improving education, health, and nutrition are essential for enhancing literacy, skills, productivity, and income-earning capacity. Increasing basic infrastructure would directly help support economic activities and improve living standards. 3) To take care of those not directly benefiting from the growth process, especially vulnerable and underprivileged groups, including women.

In fact, perhaps it is because this document mentions neither East Asia or Korea that the weights are so similar. It is not pulled off in any direction by the inclusion of any of these terms.

Another interesting result is the case of document 230. It is a very short document, without even an abstract.

Date: 1997-06-26
Report#: 16812
Title: Korea- Public Hospital Modernization Project
Abstract:

There is very little, if any relation to economic development. Why then would it receive a weight from query (3) as high as 0.684027? One possible explanation is that of the few terms present in the document, Korea is one. Thus the relation of the occurrence of the term Korea to the total number of terms in the document is quite high. However, there is no correlation to the concerns of economic growth mentioned in the query. Not surprisingly document 230 was not one of the documents retrieved in query (1) or query (2) (each query retrieved 100 of the over 500 documents).

Discussion

While LSI seems quite able to retrieve documents relevant to the keywords given in a query and in this respect its performance is superior to keyword-based search engines, it is still not satisfactory in dealing with the intention of the user. Our last example where LSI returns a document that has no relation to economic development in Korea brings up the following interesting question: What additional (external to LSI) criteria should be used for rejecting or selecting a document? We could consider several parameters: length of a document, relative weights of terms within a query (for instance, given our domain knowledge, we might associate a higher weight to 'economic growth' than 'Korea'), etc. Moreover, we can take advantage of subsumption axioms (Woods 2000) and rules that capture a good deal of domain knowledge about relationships between concepts to preprocess queries and achieve a better match to user's intention. For example, our results were not quite as expected in so far as returning similar documents for queries on economic growth in East Asia and Korea. Given that Korea is an East Asian country, one might well expect to see a higher correlation of weights for retrieved documents. Preprocessing of the query in this case amounts to utilizing a keyword based concept hierarchy to insure that related keywords are included in any given query. Thus for example, Korea would be included in a query on East Asia.

It is clear that, despite LSI's ability to find correlations between terms and documents, there is still a

great need for storing domain knowledge in terms of heuristic rules and relations between concepts (e.g., relation between countries and regions). Such a system would consist of three integrated components: A *case library* containing concrete cases of development projects; A *knowledge base* that contains general domain knowledge in the form of relations among features and heuristic rules; and A *situation assessment user interface* that derives more meaningful features before attempting retrieval of a useful case (Moussavi 1999). LSI would be integral to the system in a number of ways. For one thing, it would be used to index the case library, thus extending the indexing of the system to include more than just the knowledge base.

One, very useful, ability of many CBR applications is the ability to adapt old cases to new ones. Given a knowledge base of documents, no adaptation is possible, as the documents cannot be changed. However, with a system that is constructed to handle user queries of the knowledge base, the queries themselves can become cases. Thus adaptation of old queries to new ones is a possibility. Furthermore, this focuses the adaptation process on the desires of the user, rather than the facts of the knowledge base, which seems appropriate for improving system response to user needs. In order to do this, the queries, both old and new, would need to be indexed. Again, LSI is a powerful tool for accomplishing such indexing. LSI would be useful for this in two ways. One, the actual queries could be indexed. While the queries are quite short, over time, quite a number of them would be built up. Secondly, the results of the queries could be indexed. This would allow for the generation of a knowledge concerning query results. Such knowledge could be used as feedback for the system in the creation and maintenance of concept categories, types of information found useful by users given with different searches, and other information related to search cases. Such information would be invaluable to the long-term maintenance and growth of such a CBR system. Indexing by a system such as LSI would make the information eminently more accessible.

Summary

This paper looked at the effectiveness of LSI in returning documents that are close to what the user was looking for. The performance of LSI in this task has ramifications for its use in that task, as well as other related. In order for LSI to be of use in any of those tasks, it must return results true to the user's intentions. It was seen that while LSI was good at this task, it would likely need to be augmented by other, domain specific knowledge.

Acknowledgements

The authors would like to thank the FLAIRS reviewers for their helpful comments and feedback. We would also like to thank David Leake for his help and suggestions.

References:

Aha, D. W. and Breslow, L.A. (1997). Refining Conversational Case Libraries. In *Proceedings of the Second International Conference on Case-Based reasoning*, pages 267-278. Providance, RI: Springer-Verlag.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, "*Indexing by Latent Semantic Analysis*", Journal of the American Society for Information Science, 41(6), 1990, pp. 391-407.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (p. 979). Mawhwah, NJ: Erlbaum

Moussavi, Massoud. (1999). *A Case-Based Approach to Knowledge Management*. In (Aha & Muñoz-Avila, 1999)

Woods, William (1997). *Conceptual Indexing: A better way to organize knowledge*. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories, Mountain View, CA. www.sun.com/research/techrep/1997/abstract-61.html.