Inquiry: The University of Arkansas Undergraduate Research Journal

Volume 4 Article 22

Fall 2003

Investigation on How to Improve Latent Semantic Analysis Performance

Thao Doan
University of Arkansas, Fayetteville

INVESTIGATION ON HOW TO IMPROVE LATENT SEMANTIC ANALYSIS PERFORMANCE

By Thao Doan
Department of Computer Science and Engineering

Faculty Mentor: Dr. Russell Deaton
Department of Computer Science Computer Engineering

Abstract:

Latent Semantic Analysis (LSA) is a matching technique capable of recognizing the semantic relationships of data that ordinary techniques such as string matching cannot. This is especially valuable for data integration applications, like those of Acxiom, where data items are usually related by context, rather than in a literal match. Even though it has been shown that LSA is 30% more effective in finding and ranking relevant pieces of information than existing string-by-string matching techniques (Deerwester et al., 1990; Dumais, 1995), the performance of the LSA seems to be affected by the presence of shared words, or "noise", in data. The objective of this research is to study the influence of noise on the LSA performance quantitatively and analytically, which provides understanding for the following researches to develop a noise-filter method used to improve LSA performance. Our research shows that shared terms degrade the performance of LSA for matching queries to documents from the same category, and result in increased misclassification. In addition, share terms change the document that best matches the query.

Introduction:

This project studies the effects of shared terms (noise) on the performance of the LSA technique by controlling the proportion of noise in the data and analyzing how the LSA outcomes are affected by the adjustment. The experiments were performed on a test database resembling Acxiom's business term database. Then, several metrics were applied on the generated results in order to understand the influence of noise on the technique.

Background:

Since the inception of the Internet and other achievements in Information Technology, people have gained access to a greater amount of data from different sources and on different subjects every day. However, not all of the information given is relevant. Hence, with the growth of information, people spend more time filtering, grouping and organizing useful pieces of information. Latent Semantic Analysis[1,2,3] is a data retrieval technique widely used for many purposes from text matching to image processing applications, in business as well as in academia. It is best known for being able to integrate data items of similar latent semantic content and otherwise different in literal presentation. This reduces dramatically, matching errors caused by words with unrelated meanings but having the same spelling or by different words presenting the same information. This ability to handle polysemy and synonyms makes the LSA technique more effective than other string-matching methods. At the core of LSA is a term-to-document matrix in which every value represents the occurrence of a specific term in the corresponding document. Then, this matrix is scaled down to a set of k factors (usually between 100 and 300) by the Singular Value Decomposition method. This dimension reduction brings together words of the same context (document). Therefore, the matching criterion is based primarily on the latent semantic content of the terms rather than their literal presentation.

Methods and Materials:

1.Test Database

The test database used in this study includes a set of short terms resembling Acxiom's business name data. The individual terms were letters from the alphabet, and a document or query, a string of letters. These terms were divided into five categories, four of which are specific term categories, and one is a shared term category. The latter represents shared or noise words that are common in many documents of a database and that affect the outcomes of LSA. With four terms in each category, we constructed 60 queries (4 specific categories * (2^4 –1) possible combinations = 4*15) and 16 documents (4 documents for each specific category). During the experiments, shared terms were added systematically into these queries and documents so that the results reflect the influence of noise in queries and documents on the LSA performance. In all cases, the LSA score was the cosine of the angle between the query and document vectors in the reduced LSA space.

2.Software

In this study, we used Telcordia Latent Semantic Indexing Software (LSI) implementing the concept-based retrieval method found and developed by Telcordia ™ Technologies. This method, Latent Semantic Analysis, was proved to be 30% more effective than existing string matching techniques, which match data based on their literal representation rather than their semantic context. According to Telcordia, LSI is especially beneficial when: high recall is needed (i.e., not only the exact match of query information but also related data is desired), limited description of query information is provided (e.g., figure captions and short business names), or cross-language retrieval is demanded.

3.Experiments

In the first experiment (LSA 1), no noise word was allowed in documents or queries. The result reflects LSA performance in a noise-free environment, and thus, served as both a baseline and control for subsequent tests. From the second experiment, the proportion of shared words in queries as well as documents were raised gradually as follows:

·2 nd experiment (LSA 2): 25 % noise in documents, 0% in queries
·3 rd experiment (LSA 3): 25% noise in documents and queries
·4 th experiment (LSA 4): 50% noise in documents and queries

After generating the LSA results, we made several observations on the series of changes from one experiment to the next and applied different metrics to analyze the influence of noise words on the results. One type of data matching of interest is matching a query with its proper category. For instance, the query might represent a bank that should be matched with entries in the financial institution category. The metrics for this type of match were in-category average and out-category average LSA scores, which indicate how the performance changes within the group of queries of one category and the group of queries out of that category. A second type of result is to match a query with best scoring document in the database. For example, we would like to match Computer Science with the CSCE department at The University of Arkansas, since both sets of terms represent the same entity. Therefore, we also analyze how noise adjustment affects the best-matched documents of each query.

Result Analysis:

1.Results

Due to the length limits of this paper, here are presented only representatives of the results.

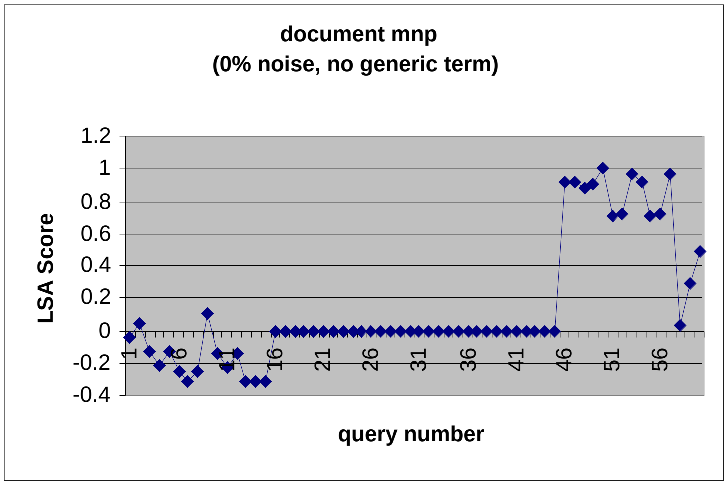

Figure 1: LSA Score versus query number for document "mnp."

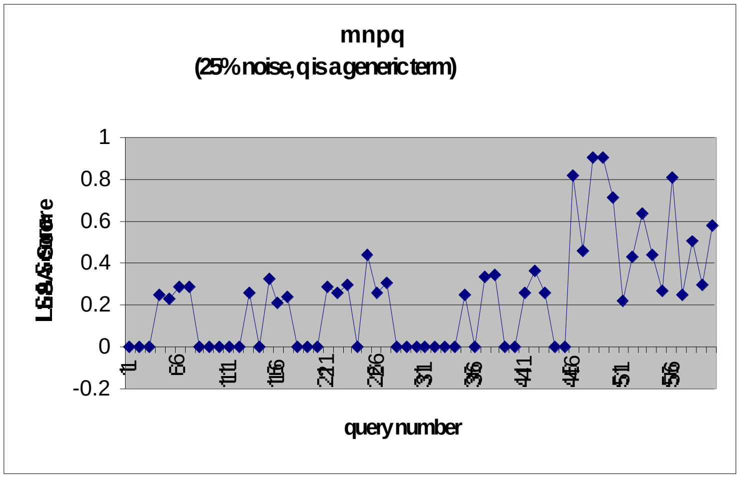

Figure 2: LSA Score versus query number for document "mnpq."

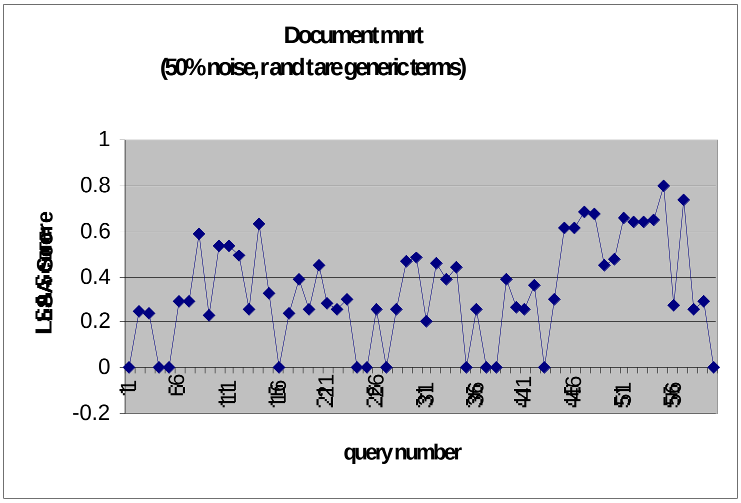

Figure 3: LSA Score versus query number for document "mnrt."

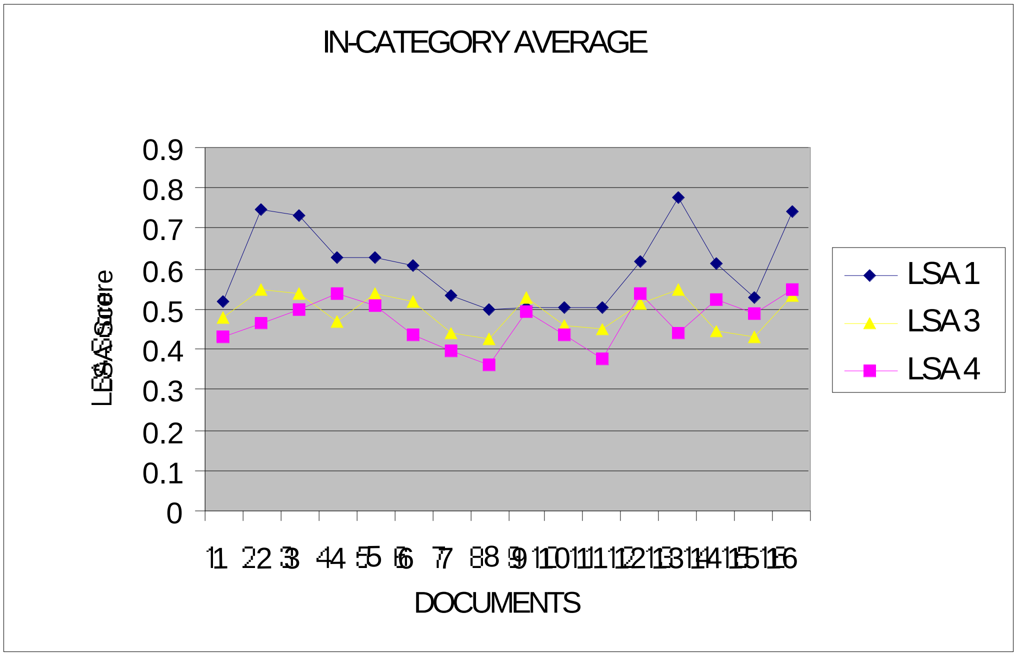

Figure 4: The average scores in experiments 1, 3, and 4 for matching to correct category.

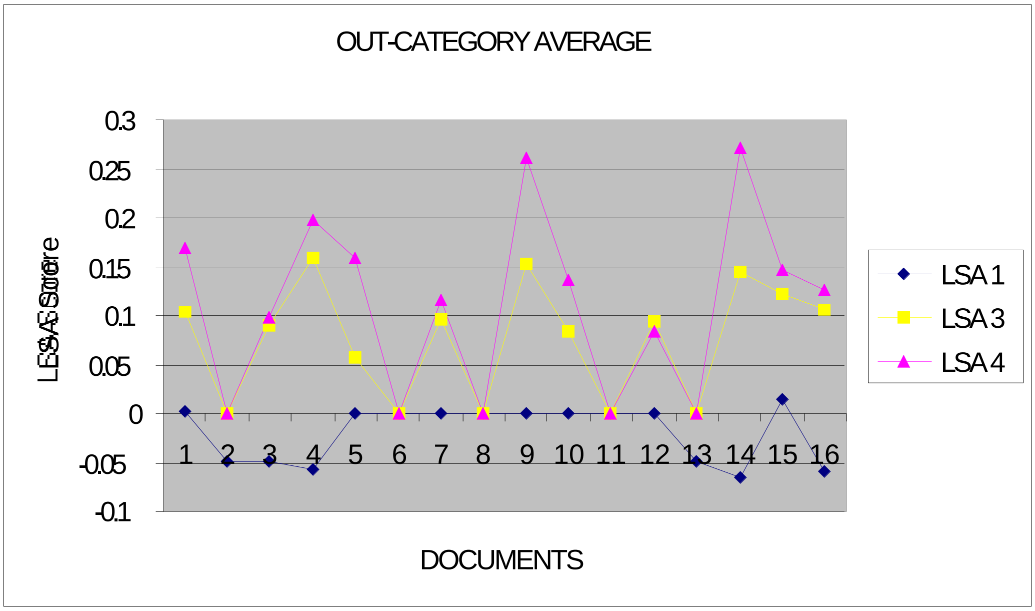

Figure 5: The average scores in experiments 1, 3, and 4 for matching to incorrect category.

Best-matched changes: For each query, there is a document scoring the highest point. However, this best-matched document varies with the adjustment of noise in the database. The following data indicates the magnitude of this change in the results of the experiments.

- **LSA 3 VS. LSA 1:**

Number of best-matched documents changed (/60): 42

- **LSA 4 VS. LSA 1:**

Number of best-matched documents changed (/60): 44

- **LSA 4 VS. LSA 3:**

Number of best-matched documents changed (/60): 33

2.Discussion

The results show that shared words in documents and queries have a significant impact on the outcome of the LSA technique. Moreover, the higher the proportion of noise in data, the smaller the gap between in-category query and misclassified query matching probabilities. For example, the difference in average performance between the two groups in the first experiment (0% noise) is 0.6247, whereas it is only 0.3570 in experiment number four (50% noise). This reduction apparently makes it more difficult to differentiate one group from the other. Another interesting observation from the results is that shared terms had more influence on misclassified query performance than on the correct category classification. LSA score changes of in-category queries are smaller than those of misclassified queries corresponding to the increasing amount of noise.

Conclusion:

Latent Semantic Analysis is a promising technique for data integration. Nevertheless, the results of this study show that increasing the proportion of shared terms in the data causes degradation in the performance of LSA. For the most part, this degradation is the result of more misclassification with increasing percentages of shared terms. In addition, the best match to a query is highly variable with changing percentages of shared terms. Thus, indentification and elimination of shared terms is key to increasing LSA performance.

References:

[1] S. Deerwester, S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” Journal of the Society for Information Science, vol. 41, pp. 391-407, 1990.

[2] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," Communications of the ACM, vol. 35, pp. 51-60, 1992.

[3] http://elib.cs.berkeley.edu/papers/clustering/bayesian/

[4] Wittenburg, K. and Sigman, E. "Integration of Browsing, Searching, and Filtering in an Applet for Web Information Access." CHI'97 short paper.

[5] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988), "Using latent semantic analysis to improve information retrieval." In *Proceedings of CHI'88: Conference on Human Factors in Computing,* New York: ACM, 281-285.