GEORGIA STATE UNIVERSITY

Measuring the Effectiveness of Web Based Collaboration with Latent Semantic Analysis

---

BY

KIREET KOKALA

MS Project Report Submitted in the Partial Fulfillment of the Requirements
For the degree of

Master of Science

Department of Computer Science
College of Arts and Science

ACKNOWLEDGEMENTS

The achievements of this project were largely based on the timely encouragement, help, and constructive criticism of many individuals. I would like to thank my project advisor Dr. Ying Zhu for introducing me to a highly interesting and promising subject, "Latent Semantic Analysis". I would also like to thank Dr. Rajshekar Sunderraman for always encouraging me and providing guidance throughout all stages of my Master's journey. Thanks to Dr. Michael Weeks for reviewing the project source code and helping with critical optimization pieces, and for being a part of my MS project committee. I am very grateful to my best friend Mohammed Ali Khokar for his involvement with the project and all his direction with the codebase. Special thanks to my mentor Ed Jenkins for the indepth discussions on architecture and other related technologies. Some of my greatest strengths are resourcefulness and a keenness to learn; I am deeply humbled to have these values instilled in me by my parents Rajani Kokala and Satyanarayana Kokala who have supported me in all my endeavors.

Measuring the Effectiveness of Web Based Collaboration with Latent Semantic Analysis

Abstract—We propose the use of a statistical analysis technique called Latent Semantic Analysis (LSA) for analyzing collaboration between online groups. LSA is a strong method for approximation in natural language processing given a large text corpora. The technology has been proven effective in analyzing complex patterns that are part of composite conversations. It is our aim to detect patterns in online collaboration and/or sample email data by capturing several inputs and subjecting them to LSA enabled programs in MATLAB. Our method of analyzing a large text corpus by the co-occurrence frequency accounts for word usage in several documents while representing their semantic meaning. We show the effectives of the collaboration, uncover a few limitations of the technique, and establish basic trends in the context of the overall conversation.

Index Terms—LSA, online collaboration semantic analysis, co-occurence frequency, pattern searching.

1 Introduction

With the increasing demand for reusable and easily accessible data in domains like health-care, military fields, scientific research, and the commercial industry, the information exchange rate online has increased with services such as e-mail, website forums, blogging, and social networking. Collaboration between individuals often leaves a highly traceable data trail and opens up the doors to statistical analysis on conversations. Latent Semantic Analysis (LSA) is one such technique which has been employed for that task. In this paper we will discuss the merits and shortcomings of experimentation with the LSA technique. LSA is a statistical method for extracting and representing the contextual meaning of words. The main idea driving this technique is that the entire word collection of all words' context within the text corpus provides mathematical cues to help determine the similarities of the words and their meanings [1, 2]. We benefit from LSA in that it provides passage-based coherence, has mechanisms that can account for noisy data, and reduce computing resources. We then introduce our solution to obtaining important result fields that help summarize the 'effectiveness' of the analyzed conversational data [2]. Briefly, our implementation of LSA starts by creating a database of the words gathered from the IBM Many Eyes Visualization website forum via Google RSS reader. The text is formatted slightly to form lexemes and regular expressions are employed to determine the total number of objects in the title arrays for comparison. Pairings of the article titles and comments are generated, and the database is updated. The term frequency, the most important field is calculated by obtaining the number of times a word occurs in a document (i.e. article). To distinguish words from other documents, we alternatively count the number of times each term occurs in each document and sum them all together. The second part of the term frequency is the inversere document frequency, which diminishes the weightage of words occurring very frequently in the wordbank. Together, the term frequency-inverse document frequency (tf-idf) helps normalize the weight of infrequently occurring words [5]. The results are stored in a matrix and are visualized in a sparse plot with density markings indicating the rate of collaboration between authors on articles [8]. We mainly find that articles with many uinique comments generally point to a strong collaboration between authors.

[1]Polysemy: the ability of words to have multiple meanings or interpretations [3, 8].

2 Background

2.1 Latent Semantic Analysis

LSA, the method for extracting and representing the contextual meaning of words through statistical computations over a large text corpus, has been applied in fields such as psychology, sociology, data mining, and theoretical applications with generally accepted success and modest criticism [2]. It was first applied to Information Retrieval (IR) and was known as Latent Semantic Indexing in the late 1980s. Later, it was used to relate the synonym and polysemy[1] problems in IR [1]. LSA has been studied rigorously since then and several variations of it have emerged [5, 7, 10]. The process starts by using an algebraic method called Singular Value Decomposition (SVD) to condense the large input data into smaller and manageable rectangular matrices of words grouped by logical passages. Each cell of the matrix contains a transform of the frequency of the given word in the passage. Next, the matrix is decomposed so that each passage is represented as a vector whose value is the sum of all vectors representing its component words.

The similarities between words-to-words, passages-to-words, and passages-to-passages are computed as cosines, or dot products, etc. [2,8]. The competency of the approach has been well correlated with several human phenomena relating association or semantic similarity [3]. For instance, its scores are similar to those of humans on standard subject matter tests and it simulates word-to-word and passage-to-word priming data. An important point about the nature of LSA is that the estimates generated are not mere frequency counts or correlations based on word usage. Instead, the results offer a semantic or deeper meaning of the arguments provided in the mathematical analysis. The details of prior reportings in the literature will be examined in the data analysis section.

2.2 Related Works

Several LSA variants and experiments are available for research and practical use. We start our focus on the work done by Wang et. al [5] to showcase a successful multiple-LSA (M-LSA) technique that was used in establishing multiple co-ocurrence relationships between types of objects analyzed. The problem dissected was that, for instance, multiple co-occurrence relations need to be represented by multiple co-occurrence matrices. The researchers constructed an undirected graph G(V,E) to show this relationship. Specifically, the goal was to find the latent semantic representations for each type of object. And, based on the co-occurrence data of G, they identified the most significant concepts based on the mutual reinforcement principle. These concepts span a semantic space. Finally, each object is represented in a unified low-dimensional space. The results of their experimentation show that the M-LSA variant outperformed standard LSA results and was applicable

towards applications, including collaborative filtering, text clustering, and text categorization [5].

In another study, Pino and Eskenazi demonstrated the use of LSA in word sense discrimination for words with related and unrelated meanings within a tutor application for English vocabulary learning for non-native speakers [10]. An indexed database containing manually annotated documents was used and the documents used by the tutor contained the target words. LSA performance for words with related meanings and for words with unrelated meanings was investigated. Lastly, they examined if reducing the document to a selected context of the target word improved performance. Their method overcame sparsity of short contexts such as questions and resulted in an improvement over the exact match baseline [10]. Comparitively, our method is more general and along the standard LSA computational lines. We performed LSA by constructing a similar database [10]. However, the input we analyzed was an amalgamation of cross domain topics, which, in the beginning phases did not readily result in any trends versus the predetermined dataset in the earlier study [10]. Secondly, our method varies in the absence of a rank-lowering algorithm. The main reason for this deviation was that the dataset was minimally sized and secondly, we did not assume the dataset to be noisy.

Our work is related to the traditional LSA algorithm [3] and features in-house optimization of the term frequency [4]. Incorporating an initial database powered co-occurrence matrix and tf-idf solution, our paper models its solution on a smaller scale when compared with M-LSA [10]. We use the technique to establish, ascertain, and visualize the rate of collaboration between authors on articles.

3 Methodology

We commence the process by gatherting data textual from the IBM Research website Many Eyes due to the large number of comments available (http://manyeyes.alphaworks.ibm.com). Comments are gathered via Google RSS reader and saved into a raw data file. Extraneous fields are stripped out, leaving intact the essentials: Article Title, Comment, Author, and Data. The data is categorized mainly by comments being grouped by the Article title to simplify later computations in MATLAB. Our *workBank* averaged a generation time of 14 seconds on a Dell PC with a 3.2 GHz processor and 2GB RAM. The implementation commences by the use of static methods such that an object is not needed to utilize laborious functions and optimize the program structure. A pre-sorted database is created from the dataset [5] with regular expressions being employed to decompose sentences into tokens of strings using quotes as the delimiter. With this approach, permissions are set appropriately to access the *wordBank* for display purposes, but also to disable the ability to add any data externally. From experimentation it is seen that the second index of the breakline is the title and fourth index is the comment associated with the title. Hence, the text is converted to lowercase for comparison. Next, each word is compared to all *TitleComments* objects in the data structure (i.e. via arrays). If one of the object's title matches with the title of current line, then this line's comment is added to the current object and stored. However, if the title does not match any of the objects in the array, then a new object is created with that title and the corresponding comment is added to the object. The process continues recursively until the end of the file. Next, the term frequency-inverse document frequency (tf-idf) is calculated. Each element in this matrix corresponds with the word count matrix. The formula used is tf*idf, where, tf = wordCount / total numWords in the comments. Thus, idf = log (the total document count divided by the number of documents the given word appears in + 1). Finally, each comment is compared to the other comments corresponding to an article by means of the tf-idf to establish their similarities. If an article has just one comment it is correlated directly with the article title. To illustrate, suppose we encounter *N* types of objects (X_1, X_2…, X_N), then each pair can have a co-occurrence relation [5,10]. The occurrence frequency of each object (α) can be individually examined or the summation ($\sum \alpha_{X1} + \alpha_{X2} + \ldots \alpha_{XN}$) can be obtained for a larger context of the dataset. The SVD technique mentioned earlier is used to group the larger dataset into smaller, manageable matrices. For a given matrix, the count of the common words was analyzed to establish the frequency specific to the relevant passage. Since every comment has a title and user, some of the patterns are propogated via relations such words → queries, words → users, words → title and so forth. Several propogations were possible and they can get more complex as a result of the model.

4 Data Analysis with LSA

A major consideration when working with large amounts of data is visualizing the results provided by LSA. Similar to M-LSA visualization and LSA analysis [6,7] we incorporated sparse graphs using an adapted visual marker technique [8]. There are advantages to using sparse graphs for dense data versus exploring visual trends for broad categories in online activities such as blogging, reading, and commenting. Mainly, we were able to readily identify comments that are most closely related to the article. In Fig. 1, the plot of comments on the Article "US Government Expenses" is seen. The original comment is the red line (C1) that is crowded with densely populated comments.

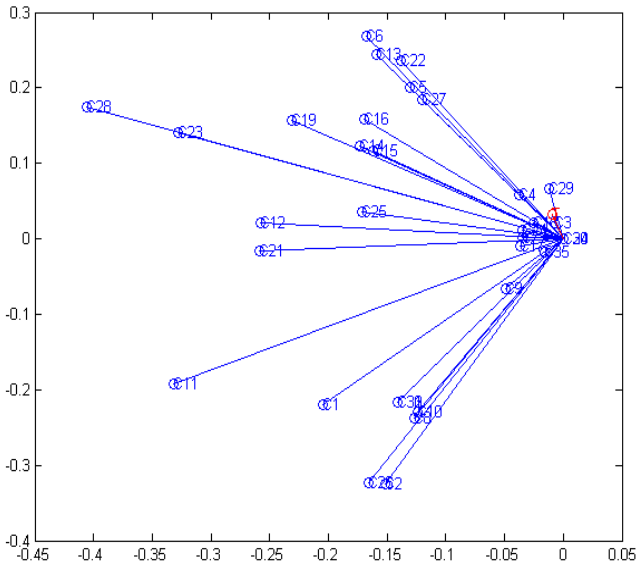

Fig. 1. US Government Expenses 1962-2004. Red line is the title line or baseline. C3 is the most closely related comment marker.

Many related comments show similar slopes and characterstics of the original title or baseline marker. Further scaling of the graphs reveals the most closely related comments; C3 was the closest and most effective comment for this article.

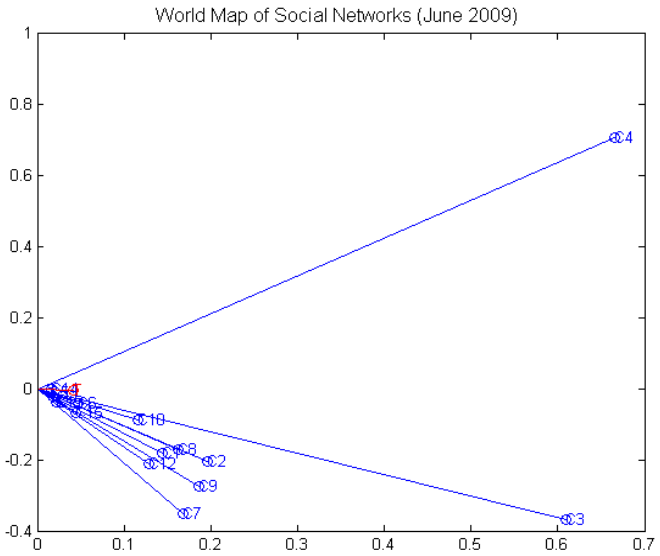

Fig. 2. World Map of Social Networks in June 2009.

Table 1 supports the findings reported in Fig.2. As seen in the table and figure, C9 is most closely related to the baseline. The slopes of C6,C8,C12 are similar and their similar word counts are also high. Of the total 11 words in the baseline article title, the highest tf-idf corresponds to C9, which is the most effective in terms of semantics. A very stark comparison of titles and comments is seen in spam filtering that has been identified in Fig. 3. The article in question features comments that are unrelated to the baseline marker. Hence, the slopes of those lines do not match, and we were able to conclude that disparate markers could be the result of: a) Spam, b) Too little data, c) Outliers. Closer inspection of the data revelaed that the comments were *Spam* due to the following raw comments, "a new version... title is available!"

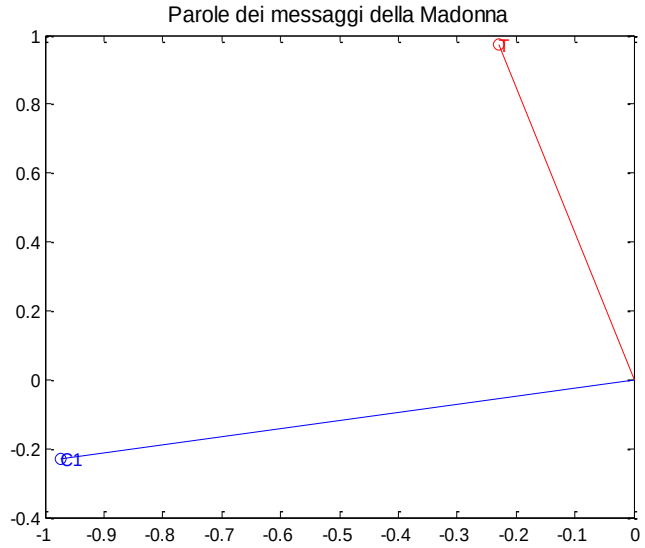

Fig. 3. Parole dei messaggi della Madonna (*Words of Our Lady's messages*). Spam identification via disparate slopes and comment magnitude.

World Map of Social Networks (June 2009)

	Total Words	Words in Common	Percent in Common
C1	41	3	7
C2	25	0	0
C3	86	5	6
C4	107	8	7
C5	11	0	0
C6	21	2	10
C7	49	3	6
C8	39	3	8
C9	26	5	19
C10	30	1	3
C11	12	0	0
C12	23	2	9
C13	7	0	0
C14	11	0	0
C15	28	2	7
T	11	11	100

Table 1. World Map of Social Networks_June 2009 metrics. The total words in the article and comments closely related are shown.

We found the results promote a strong corollation between catch words and phrases in article comments by users. The frequency of the comments, number of words in common between the baselines and comments, shaped the visualization of the data to establish basic trends. Typically, it was observed that articles with many comments generally point to a strong collaboration between authors. However, articles with fewer yet strong comments also fell into the above scneraio, which could be attributed to the overall quality and depth of the semantic meaning contained within such data.

Though we were able to capture the semantics in the word pairings, there were gaps in the analysis, which could have been the reason for some findings such as the spam comment classification. For instance, multiple meanings of words could not be ascertained due to limitations of the LSA technique [3]. The interpretation of a word occurrence having a singular meaning has been attributed to the shortcoming. A similar limitation is that the order of the words does not play a role when forming the word pairings. This is limiting in that sentence grammer is entirely skipped, which curbs the semantic limitations of the comments, particularly the lengthy ones which can pose multiple meanings. However, the above points have been discussed in the literature and we have taken account of them in our analysis of the data. In summation, we showed that these pairings from LSA were effectively exploited due to the salient words present in the comments [5].

5 Conclusion

We successfully applied traditional LSA techniques and optimized the term-frequency relations while analyzing online collaboration. In the paper, we developed a robust database powered co-occurrence matrix and tf-idf solution, to analyse more than two thousand comments. Based on the co-occurrence frequencies, we were able to analyze multiple comments on articles and show their effectiveness by broadly categorizing them as valid or spam. Our experiments showed that the LSA technique is highly valuable in classifying comments by authors by providing accurate visual cues to the user. Future improvements upon our technique can provide a more robust solution within one application that can analyze, collect data, and subject it to the LSA algorithm while providing visualization models to choose from. We have shown LSA to be highly effective in utilizing all the information from well constructed matrices and data structures. We conclude with the

affirmation in the increased accuracy of the approach and wide applications across multiple domains.

6 References

[1] 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 for Information Science and Technology (JASIS), 41(6):391-407, 1990.

[2] Landauer, T., Foltz, P., Laham, D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284, 1998.

[3] Landauer, T. and Dooley, S. "Latent semantic analysis: theory, method and application," *Proc. Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community*: 742-743, 2002.

[4] Landauer, T. K. From paragraph to graph: Latent semantic analysis for information visualization. *PNAS* April 6, 2004 vol. 101 no. Suppl 1 5214-5219.

[5] Wang, X., J.-T. Sun, et al. "Latent semantic analysis for multiple-type interrelated data objects," *Proc. 29th annual international ACM SIGIR conference on Research and development in information retrieval*: 236 – 243, 2006.

[6] Landauer, T. K. *Handbook of Latent Semantic Analysis*, 2007.

[7] Zhu, W. and C. Chen. Storylines: Visual exploration and analysis in latent semantic analysis. Computers & Graphics 31: 338-349, 2007.

[8] G.. Gorrell, *"Latent Semantic Analysis: How does it work, and what is it good for?," LSA Tutorial*, http://www.dcs.shef.ac.uk/~genevieve/lsa_tutorial.htm. 2005.

[9] Larusson, J. A. and R. Alterman. "Visualizing student activity in a wiki-mediated co-blogging exercise," *Conference on Human Factors in Computing Systems Proc.27th international conference extended abstracts on Human factors in computing systems*: 4093-4098, 2009.

[10] Pino, J. and M. Eskenazi. "An application of latent semantic analysis to word sense discrimination for words with related and unrelated meanings," *Proc. Fourth Workshop on Innovative Use of NLP for Building Educational Applications* 43-46, 2009.