Influence of domain information on Latent Semantic Analysis of Hindi text

Dr. P. Vijaya Pal Reddy
Professor, Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad

Abstract— The work presented in this paper is to evaluate the performance of Latent Semantic Analysis (LSA) model in capturing word correlations within text by including domain information in the process. The performance of the model is empirically evaluated by classification of Hindi text. The accuracies of classification are compared against plain LSA. An increase of 1.25% classification accuracy is achieved when compared to plain LSA.

Keywords— Dimensionality Reduction, Document Classification, Domain Information, LSA, SVD

1. INTRODUCTION

In the recent years, large volumes of text in Indian languages has been added to the Internet. With such huge amount of data a lot of research is going on in the direction of extracting meaningful information from text. Latent Semantic Analysis (LSA) is a mathematical technique that captures the semantics of text based on word co-occurrences within them [1]. LSA has no information about semantics like word definitions, word order, parts-of-speech or grammar rules, etc. and yet it performs fairly. The work in this paper is to enhance the performance of LSA by including domain information in the process. The performance of this enhancement is evaluated by classifying Hindi text. The classification accuracies of the enhanced model is compared against plain LSA.

This paper is organized as follows. Section 2 explains about LSA. Section 3 is a discussion about document classification using LSA. Section 4 describes the dataset used for the experiments. Section 5 presents the results of experiments by including domain information in LSA for document classification. Section 6 briefs the previous related work in literature. Section 7 presents the conclusions and further scope of work.

2. LATENT SEMANTIC ANALYSIS

The basic idea behind LSA is that words that have similar meanings tend to appear in similar contexts across text. LSA works by capturing prominent word co-occurrence patterns scattered across text. It also includes those correlations from word usages that may seem trivial but actually contribute to give unique concepts. LSA makes use of the implicit higher order structure that exists in the association of words with documents to correlate multiple words or documents referring to the same concept. LSA uses Singular Value Decomposition (SVD) to capture all correlations latent within a document by modelling interrelationships among words so that it can semantically cluster words and documents.

A. Singular Value Decomposition

Singular Value Decomposition is the core process of LSA. It is a technique in linear

algebra for matrix decompositions that breaks down a matrix A into three matrices U, S and V. Each of these matrices represents a different interpretation of the original matrix. According to the theorem stated by [2], a rectangular matrix A can be broken down into the product of three component matrices – an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix

The theorem is usually presented as follows:

$$A_{mn} = U_{mm} S_{mn} V^{T}_{nn}$$

where $U^TU = I$, $V^TV = I$; I being an identity matrix, the columns of U and V are ortho-normal eigenvectors of AA^T and A^TA respectively, and S is a diagonal matrix containing the square roots of eigen values from U or V, known as singular values, sorted in descending order. For SVD to work on a dataset with m words spread across n documents, a vector space representation is constructed as a term-by-document matrix A_{mn} of order m × n. Each row of this matrix represents a unique word and each column represents a document. Each cell i, j of this matrix contains the frequency with which word i occurs in document j. This data in matrix A is analysed for capturing significant patterns of word combinations. Matrix A is subject to dimensionality reduction which is the most crucial step in LSA.

B. Dimensionality Reduction

In the SVD process, a matrix is constructed as a product of three matrices obtained upon its eigen decomposition. In the context of LSA, the underlying principle is that the original matrix is not perfectly reconstructed. Rather, a representation that approximates the original matrix is reconstructed based on a reduced number of dimensions of the original component matrices. The original representation of data in matrix A_{mn} is reconstructed as an approximately equal matrix Ak_{mn} from the product of three matrices U_{mk} , S_{kk} and and V^T based on just k dimensions of the component matrices U_{mm} , S_{mn} and V_{nn} of the original matrix A. The diagonal elements of matrix S are non-negative descending values. If S is reduced to a k × k order diagonal matrix S_{kk} , then the first k columns of U and V form matrices U_{mk} and V_{nk} respectively. The reduced model is:

$$Ak_{mn} = U_{mk} S_{kk} V^{T}_{kn}$$

This approximate representation
of the original document after dimensionality reduction reflects all the underlying correlations. Words that occurred in some context prior to dimensionality reduction, now become more or less frequent, and some words that did not appear at all originally may now appear significantly or at least fractionally. This lower- dimensional matrix representation of the text is called as “Semantic structure” or “LSA space” or “Semantic space” in the literature [3]. In this space, the relevance of words to documents are based on not just their mere appearances but through the “concepts” that they describe in the documents. Thus, documents that may not contain a word may still be relevant to that word based on its correlation with other words used in similar contexts in those documents.

3. DOCUMENT CLASSIFICATION

The semantic space obtained after dimensionality reduction through LSA is used for document classification. In order to find the category of the test document it is first represented in the

reduced LSA space using a process called "Fold-In" [4]. To fold-in an m × 1 test document vector d into the LSA space of the lower dimensions k, a pseudo-document representation d_s based on the span of the existing term vectors (the rows of U_{mk}) is calculated as:

$$d_s = d^T U_{mk} S^{-1}$$

This pseudo-document is then appended to the set of document vectors as a row in V_{nk} . It can then compared with all the other rows representing each document in the training set using any of the standard measures of similarity like Cosine measure, Euclidean distance, etc. The category of the document which has the highest similarity with the pseudo- document is assigned to the test document d. Any of the standard approaches for document classification like k- Nearest-Neighbor (kNN), Decision Trees, Naive Bayes, Support Vector Machines (SVM), etc. are applied for classification purposes. For the present work, document classification is used to assess the modified model. One of the kNN type classifiers i.e. 1-Nearest-Neighbor (1NN) classifier is used for its intuitiveness. For measuring closeness, Cosine similarity has been used in the experiments.

4. DATASET

Many online news providers like BBC Hindi, Dainik Bhaskar, NDTV Khabar, etc. provide news articles in Hindi from a broad range of categories such as business, politics, sports, entertainment, education, technology, etc. In the present research work, all the experiments are conducted on a dataset consisting of BBC Hindi news articles from its science, sports and entertainment categories. There are many advantages of choosing BBC Hindi news as the in-house dataset. Firstly, it is a popular and rich resource of news articles available in abundance for Hindi and is freely accessible. So they are easy to collect from the Internet. Secondly, the news articles are essayed by journalists with the aim of highlighting important insights of the news story. Such articles have a lot of scope to contain natural co-occurrences of words. These natural co-occurrences provide scope for modelling word correlations. Thirdly, the rich linguistic information naturally embodied in the Hindi language text allows to gather syntactic and lexical knowledge necessary for extracting words and documents that are close to the concepts grasped by humans.

The chosen dataset contains 900 news articles downloaded randomly from the BBC Hindi news website with 300 articles in each category - "science", "sports" and "entertainment". Each document was associated with a category label based on the categorisation of the articles on the BBC website. The documents were further validated for its category against its content by a human expert. From each category 50 documents were randomly selected to provide domain information about that category. 50 articles from each category were randomly selected for performance testing and the remaining documents of each category were used for training. Table I presents the statistics of the BBC Hindi news dataset.

TABLE I
STATISTICS OF THE BBC HINDI NEWS DATASET

Document attributes	Values
Number of documents in dataset	900
Number of categories	3
Number of documents per category	300
Number of documents in training set	600
Number of documents in test set	150
Number of documents used as domain information	150

During pre-processing of documents in the dataset, initially the corpus was divided into individual documents. Then each document was broken down to a list of words. Then the punctuations, special characters and numbers were removed. Subsequently, the stop-words that were used across all the documents just as language constructs were removed as they cannot actually infer any meaning. This elimination was based on the stop-word list provided by the University of Neuchatel After this, the duplicate occurrences from the remaining word set were removed leaving only unique words. These words were further stemmed to their root forms. In order to stem the remaining words to their root-word forms, the work of Ramanathan [5] was used in which suffixes are striped off on a longest match basis. After all the pre-processing, the dataset contained only unique root words spread across multiple documents.

5. Including Domain Information in LSA

By just relying on a mathematical approach, LSA is able to capture the subtle word co-occurrence patterns including even those words that never occurred together within a single document in a collection. This way LSA performs fairly even without using any external sources that convey semantic information about documents like word definitions, parts-of-speech or grammar rules, etc. Intuitively, when additional information is added into the process, LSA's capability to understand document semantics should improve. The present work is an attempt to apply LSA on Hindi documents by providing domain information in the process.

When the training set is small, the documents in it may not be sufficient to include more number of words that are important within a domain. So any extra information about the domain of the training documents may contain some extra words related to the concepts within that domain but never used in the training set. Such words may hopefully provide significant patterns of word combinations by forming paths of higher order correlations between words the given domain. This domain information is given to LSA by adding extra documents other than the existing training set but contextually similar to the existing training set. So extra rows and columns get added to the original term-by-document matrix as shown in Fig 1.

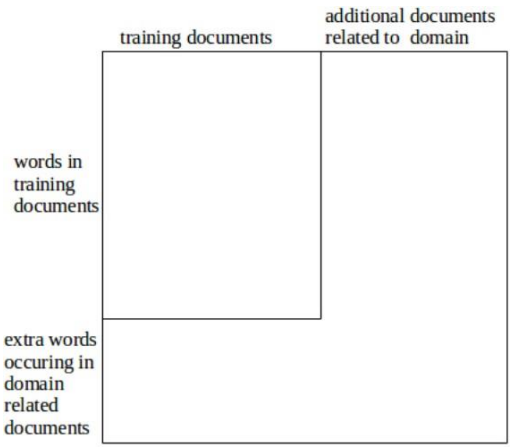

Fig. 1 Including domain information in LSA

To the existing training set of 600 documents, extra 150 documents, 50 from each category science, sports and entertainment are included as domain information into the original term-by-document matrix. For classification, the 150 test documents are folded into the LSA space reconstructed along with the added domain information and then compared only to the columns corresponding to the original training documents. The results are shown in Fig 2. In this figure the graphs for plain LSA and LSA with domain information are shown as LSA and SLSA respectively. It is seen that the performance of SLSA with domain information is better than plain LSA along a majority of the dimensions. Even at lower dimensions, SLSA does not sacrifice the performance. The average classification accuracies are obtained as 90.94% for plain LSA and 92.19% for LSA with domain information. An increased performance in document classification by 1.25% using domain information is achieved.

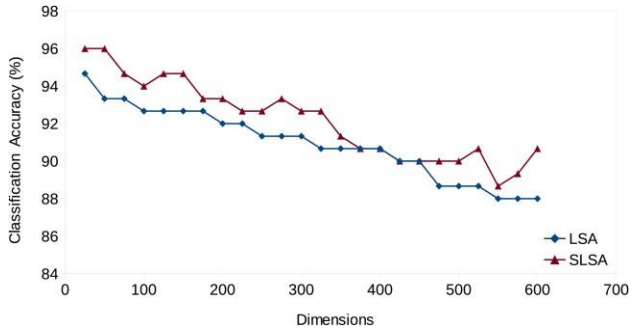

Fig. 2 Classification accuracies across various dimensions for LSA and SLSA with domain information

6. Related Work

There are several extensions of LSA that have been empirically shown to perform better for a variety of research tasks. Many of these have been specifically extended for classification problems. Relevant prior work is that of Hastings [6] in which surface parsing is employed in LSA by replacing pronouns in the text with their antecedents. The model has been evaluated as a cognitive model. Zelikovitz [7] used LSA for document classification by accommodating background knowledge for constructing the semantic space. SVD is performed on a term-by-document matrix that includes both the training documents and background knowledge. The work reported increased accuracy rates in classification. Serafin [8] suggested that an LSA semantic space can be built from the co-occurrence of arbitrary textual features which is used for dialogue act classification. Kanejiya [9] attempted to capture syntactic context in a shallow manner by enhancing target words with the parts-of-speech of their immediately preceding words. The syntactically enhanced LSA model is used in the context of an intelligent tutoring

system. The results reported an increased ability to evaluate more student answers. Rishel [10] achieved a significant improvement in classification accuracy of LSA by using part- of-speech tags to augment the term-by-document matrix and then applying SVD. The results of the work showed that the addition of parts-of-speech tags can decrease word ambiguities significantly.

7. Conclusions and Future Scope

The work presented in this paper is to determine whether including domain information has any influence over LSA's performance in capturing semantics of Hindi documents. Domain information is added into LSA by adding extra rows and/or columns to the initial term-by-document matrix from where LSA's processing starts. Along with domain information the model's accuracy of classifying documents information the model's accuracy of classifying documents .

Future scope of work is in the lines of considering category information of documents in LSA for document classification. Category information may improve the semantic space by increasing the correlations between words in text. As part of extended work, the experiments are being carried out in this direction.

References

[1] S. Deerwester, S. Dumais, G. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis", *American Society for Information Science*, pp. 391–407, 1990.
[2] K. Baker, "Singular Value Decomposition Tutorial", *Electronic document*, 2005.
[3] T. K. Landauer and P. W. Foltz, "An Introduction to Latent Semantic Analysis", *Discourse Processes*, pp. 259–284, 1998.
[4] M. Berry and S. Dumais, "Using linear algebra for intelligent information retrieval", *SIAM Review*, pp. 573–595, 1995.
[5] A. Ramanathan, "A Lightweight Stemmer for Hindi", *Workshop of Computational Linguistics for South Asian Languages Expanding Synergies with Europe*, pp. 42–48, 2003.
[6] P. W. Hastings, "Rules for syntax, vectors for semantics", *Cognitive Science Society*, 2001.
[7] S. Zelikovitz, "Using LSI for Text Classification in the Presence of Background Text", *ACM International Conference on Information and Knowledge Management*, pp.113–118, 2001.
[8] R. Serafin, B. D. Eugenio and M. Glass, "Latent semantic analysis for dialogue act classification", *Association for Computational Linguistics on Human Language Technology*, pp. 94–96, 2003.
[9] D. Kanejiya, A. Kumar and S. Prasad, "Automatic Evaluation of Students Answers using Syntactically Enhanced LSA", *Workshop on Building Educational Applications using Natural Language Processing*, pp. 53–60, 2003.
[10] T. Rishel, A. L. Perkins, S. Yenduri and F. Zand, "Augmentation of a Term-Document Matrix with Part-of-Speech Tags to Improve Accuracy of Latent Semantic Analysis", *International Conference on Applied Computer Science*, pp. 573–578, 2006.