ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 7, July 2015*

# Influence of Supplementary Information on the Semantic Structure of Documents

**Karthik Krishnamurthi[1], Vijayapal Reddy Panuganti[2], Vishnu Vardhan Bulusu[3]**

Department of Computer Science, Christ University, Bangalore, India[1]

Department of Computer Science and Engineering, GRIET, Hyderabad, India[2]

Department of Information Technology, JNTUHCEJ, Karimnagar, India[3]

**Abstract**: Latent Semantic Analysis (LSA) is a mathematical model that is used to capture the semantic structure of documents based on word correlations in them. In-spite of being completely independent of any external sources of semantics, LSA captures the semantic structure quite well. However, previous work in the literature show that including any supplementary information in LSA influences the model's ability to capture the semantic structure of documents. The work presented in this paper is to investigate how supplementary information influences the semantic structure of documents.

**Keywords**: Dimensionality Reduction, LSA, Semantic Structure, Supplementary Information, SVD.

## I. Introduction

Since the invention of Internet till the present age, huge number of documents are being added to the world-wide information repositories on a daily basis. With the availability of such huge data on the Internet, the present day research is progressing towards developing methods for machines to understand, learn and extract meaningful information from documents. Among various approaches found in the document analysis literature, Latent Semantic Analysis (LSA) is one technique that captures the semantic structure of documents based on word co-occurrences within them [1]. In-spite of being completely independent of any external sources of semantics, it performs quite well. However, any extra information included in LSA influences the model's ability to capture the semantic structure of documents. The contribution of the present work is to study the influence on the semantic structure of documents by supplementing LSA with extra information. The rest of the paper is organised as follows. Section 2 presents the related work in the literature. Section 3 is a discussion on LSA. Section 4 explains co-occurrence patterns in LSA. Section 5 interprets the working of LSA using coordinate geometry. Section 6 discusses the influence of supplementary information on LSA. Section 7 presents the conclusions.

## II. Related Work

There are several extensions of LSA that were empirically shown to perform better in classification problems. Relevant prior work is that of Wiemer-Hastings et al. [2] in which surface parsing is employed in LSA by replacing pronouns in the text with their antecedents. The model was evaluated as a cognitive model. Serafin et al. [3] suggested that an LSA semantic space can be built from the co-occurrence of arbitrary textual features which can be used for dialogue act classification. Kanejiya et al. [4] 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 et al. [5] 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. Eugenio et al. [6] used LSA in a text classification application to capture the higher order structure of dialogue contexts by adding richer linguistic features to LSA. The results showed better performance when classification was carried out on the reduced semantic spaces generated by feature-LSA compared to plain LSA. Krishnamurthi et al. [7] used LSA for Hindi document classification by accommodating domain information for constructing the semantic space. SVD was performed on a term-by-document matrix that includes both the training documents and domain information. The work reported increased accuracy rates in classification. Krishnamurthi et al. [8] suggested Supplemented Latent Semantic Analysis, a modification over LSA and applied it for classification of Hindi documents to improve classification accuracies.

## III. Latent Semantic Analysis

LSA uses Singular Value Decomposition (SVD) followed by Dimensionality Reduction to capture all correlations latent within a document by modelling interrelationships among words so that it semantically clusters words and documents. SVD 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. Rectangular matrix A is broken down into the product of three component matrices – an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V. The theorem is usually presented as follows [9]:

DOI 10.17148/IJARCCE.2015.4750

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

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.

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. Mathematically, 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 $V^T_{kn}$ 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 texts is called as “Semantic structure” or “LSA space” or “Semantic space” in the literature [10]. 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.

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” [11]. 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 $V_{nk}$ and 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.

## IV. Co-occurrence Patterns in LSA

In a document collection a word may co-occur with many words. Even if words do not directly co-occur in any document in the dataset they may still be related transitively. Suppose a document contains the words x and y then x and y have a first order co-occurrence. But if x co-occurs with z in document d1 and suppose z co-occurs with y in another document d2 then x and y have a second order co-occurrence via z. Further, z being a common attribute between d1 and d2, it establishes a second order co-occurrence path or a connectivity chain x-z-y binding d1 and d2. Suppose that a word combination pattern with first order co-occurrence appears in r documents. Then there exists r first order co-occurrence paths for that word combination. Now suppose that a word combination pattern of second order co-occurrence recurs across documents through b words then the number of second order co-occurrence paths for the said word combination is the number of unique words c out of b words. This example can be further extended to third, fourth or $n^{th}$ order co-occurrence paths for word combination patterns. A word may co-occur with multiple words via co-occurrence paths of various orders. Higher the order of co-occurrence for a word combination, lesser is the contribution of the word to the meaning or concept described by the documents in which it is used.

Apart from capturing the word combinations within a document LSA also captures higher level associations among words that occur across multiple documents in a collection. From the mathematical point of view, for a term-by-document matrix A, the $AA^T$ and $A^TA$ give only the first order word co-occurrence and first-order document co-occurrence matrices respectively. Using SVD these first order co-occurrence matrices are mapped to their corresponding eigenvector matrices that capture even higher order co-occurrences resulting in the singular matrices U and V respectively. From a semantic perspective LSA derives a latent semantic structure from the documents represented by matrix A. In the reduced k-dimensional LSA space the matrix U connects m words to k concepts. A value in the cell i, j of matrix U is the strength of the word i towards the concept j. Similarly, the matrix V relates n documents to k concepts and the value in the cell i, j of matrix V is the contribution of document i towards concept j. The matrix S gives the importance of each of the concepts. A concept is stronger because there are more documents and more words in the document collection that describes it. If a word combination pattern is recurring across multiple documents then this pattern is captured and represented by a value denoting the word's contribution to one of the concepts represented by the eigenvectors. The magnitude of the corresponding singular value indicates the importance of this pattern within the document. Any document containing this word combination pattern will be projected along this eigenvector and the sentence that best represents this pattern will have the largest index value for this vector [12].

DOI 10.17148/IJARCCE.2015.4750

## V. Geometrical Interpretation of LSA

Viewing LSA from a geometrical perspective gives a better understanding of the working of the model. In this context, the foremost understanding is that a VSM m × n term-by-document matrix is viewed as m points (or vectors) in the n-dimensional document space. Each point has coordinates along the dimensions in which it is projected.

A concrete understanding of how LSA captures the semantic structure of documents is achieved using a document classification example. Consider the training documents D1, D2, D3 and a test document D4. An LSA model is constructed using these training documents to classify the test document. In this example only word occurrences in the documents are considered without giving much of importance to the grammatical formation of documents. There are in all 15 words spread across the documents.

D1:...speed....gravity....speed...rate...space...disease...plant.
.....petrol....gravity....gravity...disease...
D2:...profit..company...petrol...company...launch...market..
.....company...rate...profit...profit..goal..
D3:....rate...cricket...player...speed...market...goal...player..
.....champion....player....goal....champion....speed....space...
.....company
D4:...goal...company.....rate...

One can understand that the words speed, gravity, rate, space, launch, disease and plant describe concepts in the category say "science". The words profit, company, petrol, launch, market are used to represent the category "business". And the words cricket, player, champion, goal, rate infer "sports". It is to be noticed that some words like launch, rate, etc. may relate to more than one category. In such cases their usages along with other words i.e. their co-occurrence patterns in the documents will infer their concept category. On the whole, a document speaks about a concept category based on its word usages. LSA helps to extract such concepts like science, business and sports based on the word co-occurrences. However, it won't give nice human readable names to these categories.

The document collection is represented as a term-by-document matrix A of order 15 × 3 in Table 1. As each term is a point in space, the matrix A is represented graphically as a plot of 15 points in a 3-dimensional coordinate space shown in Figure 1. The x, y, z directions represent positive values of the axes and x', y' and z' represent the negative values of the axes. The axes x, y and z correspond to the documents D1, D2 and D3 in this plot. The coordinates say (x1, y1, z1) of any point (term) with respect to the axes x, y, z (documents) in this original "document space" represents the number of times that term occurred in documents D1, D2 and D3 respectively. In the rest of the figures in this section, terms will be represented as red spots and documents as black spots in the coordinate system.

| | D1 | D2 | D3 |
|---|---|---|---|
| speed | 2 | 0 | 2 |
| gravity | 3 | 0 | 0 |
| rate | 1 | 1 | 1 |
| space | 1 | 0 | 1 |
| disease | 2 | 0 | 0 |
| plant | 1 | 0 | 0 |
| petrol | 1 | 1 | 0 |
| profit | 0 | 3 | 0 |
| company | 0 | 3 | 1 |
| launch | 0 | 1 | 0 |
| market | 0 | 1 | 1 |
| cricket | 0 | 0 | 1 |
| player | 0 | 0 | 3 |
| goal | 0 | 1 | 2 |
| champion | 0 | 0 | 2 |

Table. 1 Matrix A of order 15 × 3 for 15 terms across 3 documents

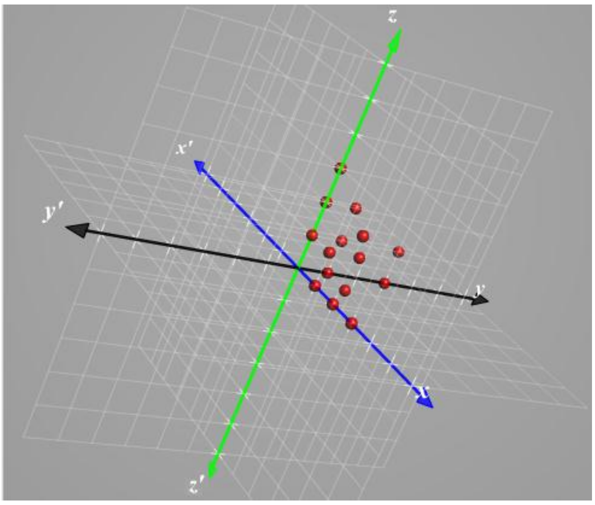

Fig. 1 Terms as 15 points scattered in 3-dimensional document space

The SVD of matrix A results in matrices U, S and $V^T$ that establish an n-dimensional orthogonal space known as the "LSA space" or "Semantic space" where the terms and documents are distributed according to their common usage patterns. This is shown in Figure 2 where the points of matrix A are transformed by SVD to a 3-dimensional semantic space. In this semantic space, the rows of $V^T$ represent concept vectors which are obtained by reorientation of the original document axes. The original position of point (term) does not move but with the reorientation of the document axes, their distances change with respect to the new reoriented concept axes resulting in new projections/co-ordinates. The new projection of a term is a point in this reoriented space whose coordinates are obtained as the product of values in the corresponding row of eigenvectors in U and the corresponding singular value S. The singular values in S represents the degree with which the reorientation takes place with respect to the corresponding original axes. These new coordinates, say (x2,y2,z2) of the term (point) represent the contribution (occurrence) of that term in some concepts say concept1 (x axis), concept2 (y axis) and concept3 (z axis) respectively. The same understanding is extended for documents as well.

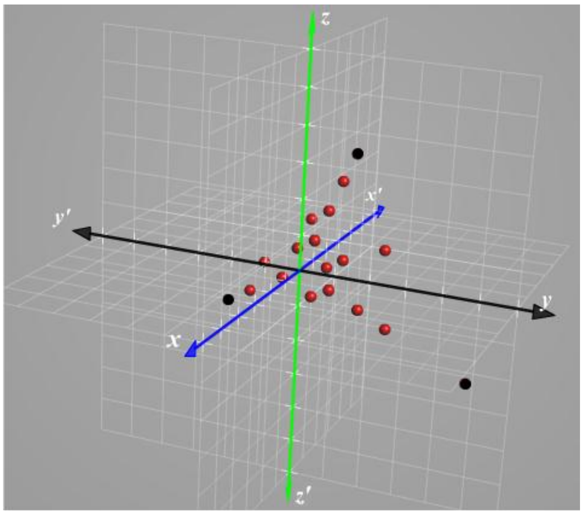

Fig. 2 15 terms and 3 documents in the LSA semantic space

Each eigenvector (the row of $V^T$) represents a concept within documents and the value of its corresponding singular value in S represents the degree of importance of that concept. In this space the first axis x (the one corresponding to the largest singular value) is the most significant concept. Formally, the variance of the points (terms) along this axis is the greatest. The second axis y, corresponding to the second singular value, is the next most significant in the same sense, and so on for each of the singular values. Based on this idea, it is observed from Figure 3 that most of the points lie closer to the xy plane only reflecting the fact that the terms and documents better highlight concept1 (x axis) and concept2 (y axis), rather than concept3 (z axis). So the third dimension (z axis) corresponding to the smallest singular value is removed and the 3-dimensional space is reduced to 2-dimensions.

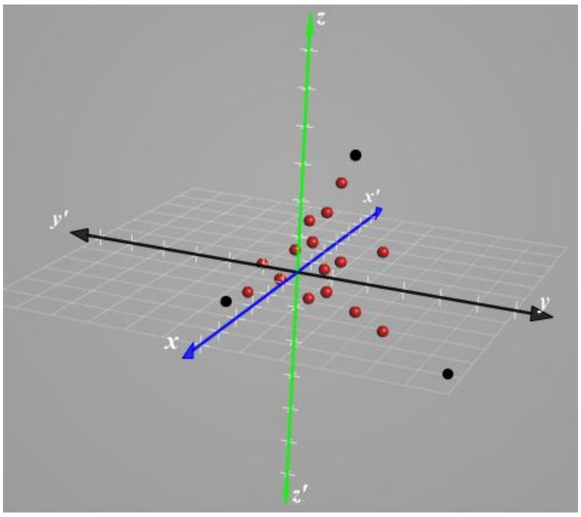

Figure 3: Terms and documents lying closer to XY plane

When the third dimension is removed then every point (term or document) is projected in the remaining 2-dimensions. These new co-ordinates in 2-dimensions are calculated based on the strengths representing the contribution of that term towards concept1 and concept2 (first and second columns of U matrix) which themselves were based on word co-occurrences. Multiplying the strengths in U matrix with the corresponding concept strengths (singular values in S matrix) gives the new coordinates. A term's new coordinates represent the number of occurrences of that term with respect to the concepts (axes x and y). This applies for documents also. Figure 4 is a plot of Figure 3 after dimensionality reduction to 2 dimensions.

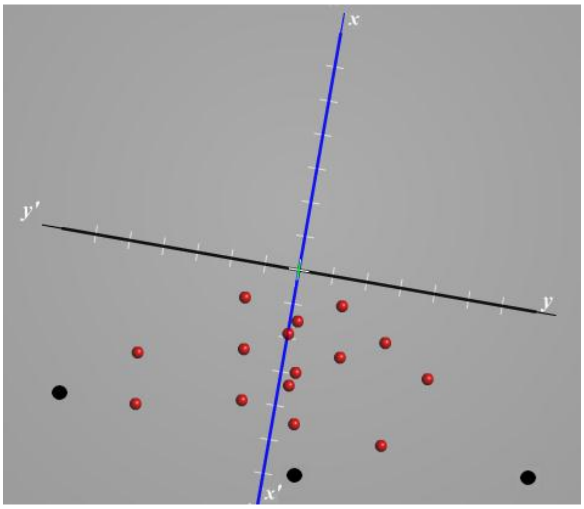

Fig. 4 Terms and documents in reduced semantic space in 2-dimensions

On the whole, in LSA, through a combination of singular value decomposition and dimensionality reduction, the representation of terms that occur in similar contexts become more similar moving closer to the reoriented axes. The LSA space reflects those terms that have been used in the document to give information about the concepts (the axes) to which the terms are closer. Essentially, LSA is a proximity model that spatially groups similar terms and documents together. As the dimensional space is reduced, related documents draw closer to one another. The relative distances between these points in the reduced vector space show the semantic similarity between documents, and is used as the basis for the document classification. A test document (a set of terms) is mapped as a pseudo-document into the semantic space by the process of folding-in. Then the pseudo-document's closeness with all other documents is measured. The category of the document that is the located in its nearest proximity in space is the category of the test document. This is understood by observing Figure 5. The black spot encircled with yellow is the pseudo-document representation of the test document D4. Using the cosine

DOI 10.17148/IJARCCE.2015.4750

similarity to measure closeness, D4 is classified to the category of D2 due to its closeness as is seen in Figure 5.

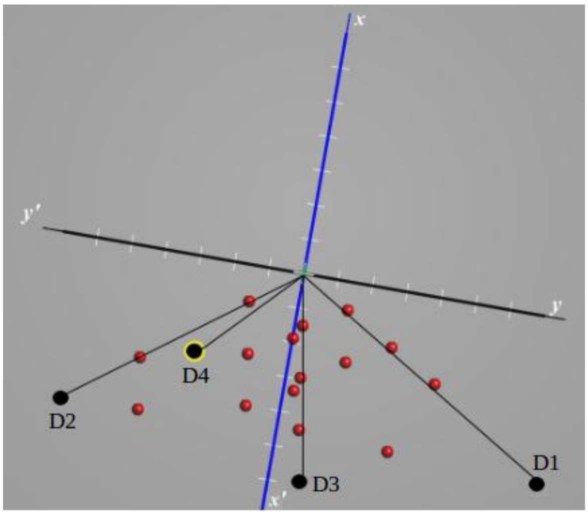

Fig. 5 Pseudo document D4 with other terms and documents

In contrast to many other methods of document classification, LSA is able to categorize semantically related texts as similar even when they do not share a single term. This is because in the reduced space, the closeness of documents is determined by the overall patterns of term usage. So documents are classified as similar regardless of the precise terms that are used to describe them. As a result, terms that did not actually appear in a document may still end up close to it if that is consistent with the major patterns of association in the data.

## VI. Supplementary 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 well 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. There are several extensions of LSA that have shown to perform better for a variety of research tasks in the literature. Many of these have been specifically extended for classification problems.

In the present work LSA is applied on documents along with supplementary information in the model. Essentially, adding extra information to LSA is just like adding new words to the initial term-by-document matrix. So extra rows get added for the information that is intended to be given as a supplement to the LSA process. From the geometrical perspective, the newly added supplementary information are points in the initial document space. By understanding how LSA behaves in the coordinate space, it is evident that it is the correlation between words that decides the amount of reorientation of axes in the semantic space. This reorientation in turn draws the points closer to or farther from the concept axes. Thus, the performance of LSA depends on how the reorientation of axes affects the variance in each concept, i.e. how many number of words and documents are spread across each concept. With respect to document classification, LSA's performance depends on how best the reorientation of axes draws the test document closer to the training documents of the appropriate concept. An example will help in getting a concrete understanding of how supplementing LSA with extra information influences its performance. Consider the following sample collection of three training documents d1, d2, d3 and a test document dT.

**d1**: The boy was walking with his dog
**d2**: The dog went to the park
**d3**: The girl was strolling with her pet
**dT**: The boy was strolling in the park

| | **d1** | **d2** | **d3** |
|---|---|---|---|
| boy | 1 | 0 | 0 |
| walk | 1 | 0 | 0 |
| dog | 1 | 1 | 0 |
| went | 0 | 1 | 0 |
| park | 0 | 1 | 0 |
| girl | 0 | 0 | 1 |
| stroll | 0 | 0 | 1 |
| pet | 0 | 0 | 1 |

Table. 2 Matrix A

| | **dT** |
|---|---|
| boy | 1 |
| walk | 0 |
| dog | 0 |
| went | 0 |
| park | 1 |
| girl | 0 |
| stroll | 1 |
| pet | 0 |

Table. 3 Test document dT

| | **d1** | **d2** | **d3** |
|---|---|---|---|
| boy | 0.5 | 0.5 | 0 |
| walk | 0.5 | 0.5 | 0 |
| dog | 1 | 1 | 0 |
| went | 0.5 | 0.5 | 0 |
| park | 0.5 | 0.5 | 0 |
| girl | 0 | 0 | 1 |
| stroll | 0 | 0 | 1 |
| pet | 0 | 0 | 1 |

Table. 4 LSA Matrix A2 after retaining 2 dimensions

| | **d1** | **d2** | **d3** |
|---|---|---|---|
| **d1** | - | 2 | 0 |
| **d2** | 2 | - | 0 |
| **d3** | 0 | 0 | - |

Table. 5 Matrix A2$^T$A2 with document-document similarity

The term-by-document matrix A with term-frequencies is created after stop-word removal and stemming as shown in Table 2. Similarly the test document dT is represented in Table 3. Upon performing SVD on matrix A followed by dimensionality reduction retaining 2-dimensions, the

DOI 10.17148/IJARCCE.2015.4750

reconstructed LSA matrix A2 is shown in Table 4. To compare if two documents are conceptually the same, the dot product between the two columns of matrix A2 is calculated which reflects the extent to which they have a similar profile of words inferring the same meaning [1]. This is obtained by multiplying the transpose of matrix A2 with matrix A2. The resulting matrix containing the document-document dot products is shown in Table 5. Here the similarity of a document with itself is replaced by a '-' symbol. The 0 value between d1 and d3 indicate that they are not similar. This is due to the fact that the original documents d1 and d3 do not share any common words between them, so no connectivity chain is established between words of d1 and d3 and hence it is impossible for LSA to capture any sort of co-occurrence pattern between words of d1 and d3. Using this LSA space for classifying a test document dT, the cosine similarities 0.78, 0.78 and 0.63 are obtained with respect to d1, d2 and d3 respectively. So the categories of both d1 and d2 are assigned to the test document giving an ambiguity.

| | d1 | d2 | d3 |
|---|---|---|---|
| boy | 1 | 0 | 0 |
| walk | 1 | 0 | 0 |
| dog | 1 | 1 | 0 |
| went | 0 | 1 | 0 |
| park | 0 | 1 | 0 |
| girl | 0 | 0 | 1 |
| stroll | 0 | 0 | 1 |
| pet | 0 | 0 | 1 |
| X | 1 | 0 | 1 |

Table. 6 Matrix B with X as supplement

| | dT |
|---|---|
| boy | 1 |
| walk | 0 |
| dog | 0 |
| went | 0 |
| park | 1 |
| girl | 0 |
| stroll | 1 |
| pet | 0 |
| X | 0 |

Table. 7 Test document with X

| | d1 | d2 | d3 |
|---|---|---|---|
| boy | 0.65 | 0.44 | 0.19 |
| walk | 0.65 | 0.44 | 0.19 |
| dog | 1.09 | 0.89 | -0.05 |
| went | 0.44 | 0.46 | -0.24 |
| park | 0.44 | 0.46 | -0.24 |
| girl | 0.19 | -0.24 | 0.89 |
| stroll | 0.19 | -0.24 | 0.89 |
| pet | 0.19 | -0.24 | 0.89 |
| X | 0.84 | 0.19 | 1.09 |

Table. 8 LSA Matrix B2 retaining 2 dimensions

| | d1 | d2 | d3 |
|---|---|---|---|
| **d1** | - | 1.96 | 1.43 |
| **d2** | 1.96 | - | -0.53 |
| **d3** | 1.43 | -0.53 | - |

Table. 9 Matrix $B2^TB2$ with document-document similarity

Now suppose that some extra information, say X, is added as a supplement to matrix A. This is done by adding X as a row to matrix A resulting in matrix B shown in Table 6. One can understand that documents d1 and d3 are conceptually the same as they both convey “the act of a person walking”, though they don't share any words in common. In order to include this human understanding into the matrix representation, the row X holds a value 1 for documents d1 and d3 and 0 for d2. The test document dT is represented as in Table 7. Performing SVD on matrix B followed by dimensionality reduction retaining 2-dimensions gives matrix B2 as shown in Table 8. The document-document similarity matrix obtained from B2 is shown in Table 9.

A keen observation of the values of the matrix in Table 9 reflects how the addition of the supplementary information X influences the semantic space of B2. The similarity between documents d1 and d3 is now 1.43, which was earlier 0 in Table 6 conveying that d1 and d3 are indeed conceptually similar. What X does is that it establishes a connectivity path between words of documents d1 and d3 due to which LSA is now able to capture higher-order co-occurrences within the document structure. Using this supplemented LSA space for classifying the test document dT, the cosine similarities 0.98, 0.66 and 0.60 with respect to d1, d2 and d3 respectively are obtained and dT is precisely classified to the category of the first document.

It is observed that because of supplementing LSA with extra information X, the supplemented model captures word correlations better, thereby strengthening the relationships between documents within a concept. The performance of document classification also is affected by the presence of such extra supplements or any of its combinations.

## VII. CONCLUSION

The work presented here is to determine how supplementing LSA with extra information influences the model's capability of capturing the semantic structure of documents. Supplementary information is added into LSA by adding extra rows to the initial term-by-document matrix from where LSA's processing starts. An analysis of LSA is carried from a coordinate geometrical perspective which gives an understanding of how LSA's behaviour is influenced when extra information is provided. It is shown that the modified LSA model captures reasonably stronger correlations than LSA in the semantic space. It is concluded that supplementing LSA with extra information indeed increases its performance and therefore the modified LSA can be used as an efficient model to analyse word correlations.

## REFERENCES

[1] S. Deerwester, S. Dumais, G. Furnas and T. K. Landauer, “Indexing by latent semantic analysis”, *American Society for Information Science*, 391–407, 1990.

[2] P. Wiemer-Hastings and I. Zipitria, “Rules for Syntax, Vectors for Semantics”, *Annual Conference of the Cognitive Science Society*, 1112–1117, 2001.

[3] R. Serafin, B. D. Eugenio and M. Glass, “Latent semantic analysis for dialogue act classification”, *North American Chapter of the ACL on Human Language Technology*, 94–96, 2003.

[4] D. Kanejiya, A. Kumar and S. Prasad, “Automatic Evaluation of Students Answers using Syntactically Enhanced LSA”, *Workshop on Building Educational Applications using NLP*, 53–60, 2003.

[5] T. Rishel, A. L. Perkins and S. Yenduri, “Augmentation of a Term-Document Matrix with Part-of-Speech Tags to Improve Accuracy of LSA”, *International Conference on Applied Computer Science*, 573–578, 2006.

[6] B. D. Eugenio and R. Serafin, “Dialogue Act Classification, Higher Order Dialogue Structure and Instance-Based Learning”, *Dialogue and Discourse*, 1–24, 2010.

[7] K. Krishnamurthi, V. R. Panuganti and V. V. Bulusu, “Influence of domain information on Latent Semantic Analysis of Hindi text”, *International Journal of Computer Science Information and Engineering Technologies*, 2(4), 1–4, 2014.

[8] K. Krishnamurthi, V. R. Panuganti and V. V. Bulusu, “Capturing the semantic structure of documents using summaries in Supplemented Latent Semantic Analysis”, *WSEAS Transactions on Computers*, 314–323, 2015.

[9] K. Baker, “*Singular Value Decomposition Tutorial”,* http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf, 2005.

[10] T. K. Landauer and P. W. Foltz, “An Introduction to Latent Semantic Analysis”, *Discourse Processes*, 259–284, 1998.

[11] M. Berry and S. Dumais, “Using linear algebra for intelligent information retrieval, *SIAM Review*, 573–595, 1995.

[12] Y. Gong and X. Liu, “Creating Generic Text Summaries”. *International Conference on Document Analysis and Recognition*, 903–907, 2001.

**DOI 10.17148/IJARCCE.2015.4750**