International Journal of Research in Advent Technology, Vol.6, No.5, May 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Novel Approach for Document Categorization Based On Latent Semantic Indexing

Dr. Sandeep Nain[1], Ankita[2]
Associate Professor, *Computer Science & Engineering*, Galaxy Global Group of Institutions, Dinarpur, India[1]
M. Tech. student, *Computer Science & Engineering*, Galaxy Global Group of Institutions, Dinarpur, India[2]
Email: sandeepnain77@gmail.com[1], ankitadhimn@gmail.com[2]

Abstract- Various text mining schemes were used for the classification of documents available in unstructured form on net such as Naive Bayes, TF/IDF weights, Latent Semantic Indexing, SVMs etc. To represent the documents in the form of vectors, VSM schemes are widely used. In this paper, we have proposed a novel technique to classify the text documents based on latent semantic indexing (TD-LSI). Latent Semantic Indexing (LSI) is a classifier used to represent texts in the better form as it retains the meaningful information between the terms. Also, singular value decomposition (SVD) method is used to extract textual vectors of LSI. In our experiments, we conducted comparison between our TD-LSI and various other classifiers such as, k- Nearest Neighbours, Support Vector Machine.

Index Terms - LSI, KNN, VSM, categorization, SVD, classifier.

1. INTRODUCTION

The intensive expansion of the web and the enlarged number of users has forced new organizations to place their processed data on the web [1]. Web provides many refined services like mail, social media, shopping, education etc. Besides all this, the constant development in Internet usage is enhancing the problems in controlling the information. The swift dominance of World Wide Web relevance and the want to arrange the data efficiently, to look up the data for knowledge, have emphasized to develop more intellectual and efficient real time techniques to categorize the information on net.

1.1. *Web-Content Management*

A WCMS is data management software [2], generally accomplished as a web application, to create and manage the HTML data. It is also used to handle a large amount of material available on net. The authoring tools of the software allow the users having little or no knowledge of programming or markup languages to build and handle the data with great ease. With the advent of technology, man is attempting for relevant and optimal results from the web through search engines. Content management is done to serve the objective of reduction in time and user's effort in finding the information. The effectiveness and efficiency is enhanced by using keywords and relevant phrases during the search process.

1.2. *Latent Semantic Analysis*

LSI is an indexing and information retrieval technique based on Singular Value Decomposition system to recognize the repetition in the interaction between the words and ideas in unsupervised text compilation. The words that are used in the same situation have same intentions is the basis of LSI therefore it can extract the conceptual idea from text by maintaining relationships between the terms that occur in similar situation.

1.3. *Singular value decomposition*

For document categorization by label or topic [3] Singular Value Decomposition technique is widely used. It is a way to decay a matrix into consecutive estimation [4]. The decaying of the matrix can disclose inner construction of the matrix. This is very effective method for text classification.
In this work, we have used Latent Semantic and Singular value decomposition methods to improve the precision and usefulness of the classification. The main idea behind using the combination of both is to find the association among the words and documents [5]. LSI has wide area of applications as in search engines [6], digital image processing applications [7] etc. SVD states that a rectangular X, a x b matrix can be decomposed into the product of an orthogonal matrix U, a diagonal matrix $\sum$, and the transpose of matrix V according to linear algebra theorem. LSI decomposes X using SVD as follows:

$$X = U \sum V^{T} \quad \text{.....Eq. 1}$$

LSI uses the first k vectors of the matrix U as the transformation matrix to implant the original documents into a k-dimensional region [8].

1.4. *TF*IDF*

TF*IDF; term frequency–inverse document frequency is unfolded from IDF [9, 10]. In information retrieval system TF.IDF is a mathematical sign that shows the

utility of a term in a document in the collection of text [17]. In text, web mining TF*IDF is used as weighting factor to search the content. The formula of TF*IDF used for weighing the terms is given below:

$$W = TF \times \log N/Df \quad \text{Eq. 2}$$

where W represents the weight, N is the number of documents, TF is the term frequency and Df is the document frequency of the terms in the text collection [8]

2. RELATED WORK

Document categorization is not a new area of research but still the research is going on and the new algorithms are being proposed. We have reviewed the research papers to gain some knowledge on the previously done work in order to improve the already developed algorithms and analysed to optimize those techniques.

The categorization of documents into given topics was given by Kuralenok et al. in [11] that used the latent semantic analysis to disclose meaningful relationships that identify the function of the current closeness of words used to approximate the closeness of credentials. The system was proved to be effective indicating high value of classification with the disadvantage that the topics were not labelled in advance. Considering the initial cost, the method was costly but proved to be cheap at the classification stage.

Anthony et al. in [12] made a survey of systems that used latent semantic indexing technology to categorize the documents as they are independent of secondary structures and native language being categorized. The systems utilized domestic industrial tools for creating and publishing LSI categorization spaces.

Sarah et al. in [13] presented a work that evaluates background knowledge using latent semantic indexing. Classification using LSI's SVD process is based on creating a space by combination of training data and background knowledge. Using different data sets, table of background knowledge related to instruction data, were evaluated.

Chung et al. in [14] have described a novel technique for multi language text classification based on LSI and the performance on the precision, recall was proved to be good in categorizing the multi-lingual text. The centroid of each class had been calculated and compared with its pre-defined threshold value. The documents were labelled as positive if similarity measurement has larger value as the threshold else were labelled negative. April in [15] proposed a model based on LSI that extracted term relationship information with very less dimensions. However, completion time and memory were lessened but the system suffered with noise in the vector space created by unidentified terms.

Al-Anzi et al. in [16] showed that the cosine similarity is important measure to investigate the performance of Arabic language text classification. The paper also conducted comparison between various classification methods such as Naïve Bayes, k- Nearest Neighbors, Neural Network, Random Forest, Support Vector Machine, and classification tree. The results revealed that the classification methods using LSI features outperformed the TF.IDF-based methods. Also k-Nearest Neighbors (based on cosine measure) and support vector machine are the best performing classifiers.

3. METHODOLOGY

The objective of the proposed system is to efficiently categorize the text documents by reducing the computational time as well as improving the accuracy of categorization. The research work focuses on methods using LSI for document categorization.

The steps involved in the TD-LSI algorithm are given as follows:

3.1. *Steps involved in SVD*

Step 1: The transpose of term document matrix X is to be calculated and then find XX^T.

Step 2: The Eigen values of X^TX are to found and arranged in descending order. Then using the square roots of these Eigen values singular values of X are obtained.

Step 3: By placing the singular values in descending order along its diagonal, diagonal matrix $\sum$ is formed and its inverse, $\sum^{-1}$ is calculated.

Step 4: The Eigen vectors of X^TX are calculated by using the ordered Eigen values from step 2.Place these eigenvectors along the columns of V and compute its transpose, V^T.

Step 5: Calculate $U = XV\sum^{-1}$ and $X = U\sum V^T$.

3.2. *Steps involved in TF.IDF*

Step 1: Firstly all the words are gathered in the dictionary.

Step 2: The vectors are created by the documents word counts.
Step 3: The vectors are weighted using TF.IDF. The number of vectors will be equal to the no. of words in the documents in each weighted vector.

3.3. *Steps involved in LSI*

Step 1: Form the term-document matrix X by noting the count of the terms.
Step 2: take the value of $X = U\sum V^T$ from A.
Step 3: Implement a reduced rank approximation by keeping the first columns of U and V and the first columns and rows of $\sum$.
Step 4: Find the new document vector coordinates in this reduced dimensional space. Rows of V hold eigenvector values. These are the coordinates of individual document vectors

After LSI representations have been grouped then a different SOM is used to group the documents which are encoded again by mapping their text, word by word, onto the first stage SOM.

Step 1: Use the SVD of X matrix from A.
Step 2: Form U_k whose rows are the LSI representations of the original term vectors.
Step 3: For inputting data vectors to a SOM, use the rows of the matrix U_k
Step 4: Train the SOM until convergence and re-encode the original documents by mapping their text, word by word.
Step 5: Use the new representations of the documents as input data vectors to a new SOM of fixed topology to cluster the documents

4. EXPERIMENTAL RESULTS

Before conducting the experiments, three parameters were set. The rank k, the small word threshold equal to 1, and the word frequency threshold also equal to 1. For the rank k, a range is selected to find the best performance.

4.1. *Evaluation Measures*

To compare the effectiveness between different categorization techniques, various performance measures are considered like accuracy, precision, recall and F-measure.

1) Accuracy: The accuracy is the ratio of correctly assumed observations to the number of actual observations.
2) F-Measure: F-Measure is used as the major measure as it is a widely used measure to evaluate the performance of clustering algorithms.
F1 = 2((precision * recall)/(precision + recall)).
3) Precision: For a text search on a set of documents, precision is the ratio of number of exact results to the number of all returned results [18].
4) Recall: Recall is the ratio of number of exact results to the number of results that should have been returned [18].

Table I shows the results for various parameters for three classifiers; k-nearest neighbour, SVM and the TD-LSI, in case of the LSI which was used in this evaluation. It is quite clear from the table that all the performance metrics are better in case of the TD-LSI technique.

TABLE I: Comparison of Performance Metrics for Three Classifiers

	LSI			
Classifier	**Accuracy %**	**Precision %**	**Recall %**	**F measure %**
KNN	83.89	74.28	74.24	74.16
SVM	87.78	82.63	81.28	81.68
TD-LSI	91.11	87.39	87.22	87.08

Figure 1, 2, 3, 4 shows the percentage of accuracy, precision, recall and F-measure for SVM, k-NN and the TD-LSI technique respectively.

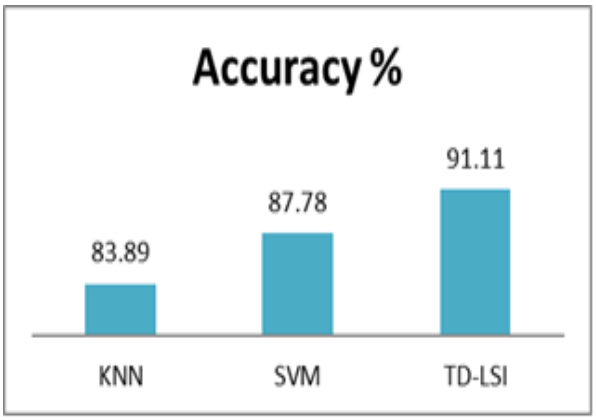

Fig. 1. Accuracy for KNN, SVM and TD-LSI technique.

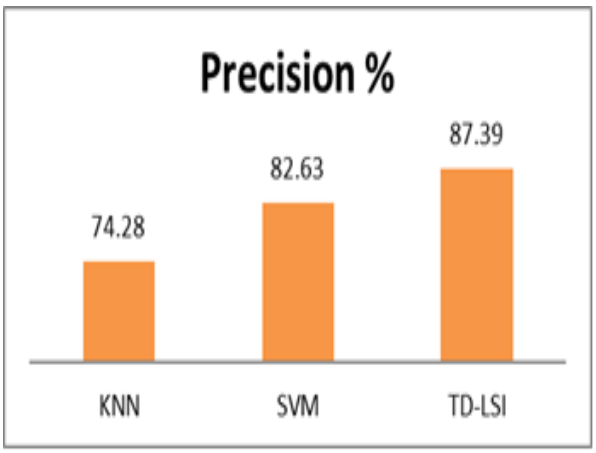

Fig. 2. Precision for KNN, SVM and TD-LSI technique

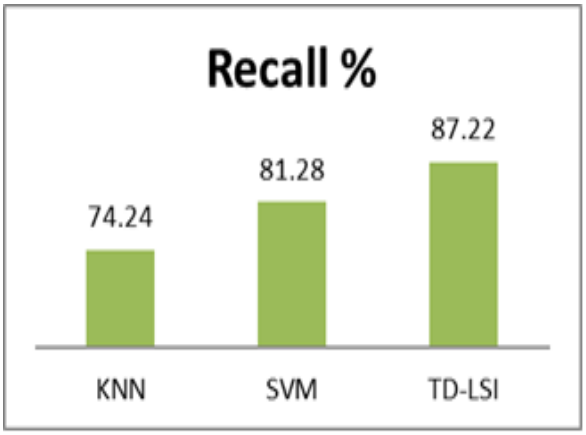

Fig 3: Recall for KNN, SVM and TD-LSI technique

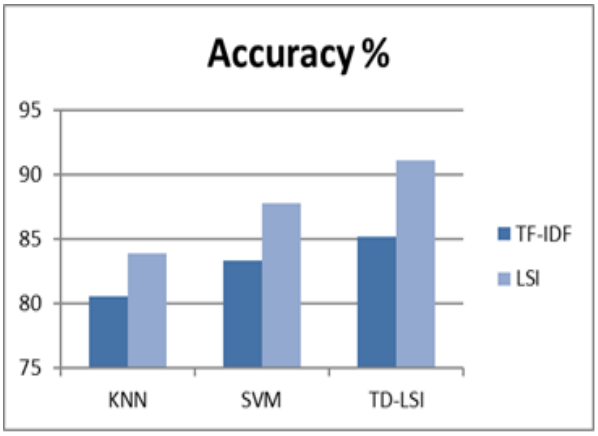

Fig 5: Comparison of Accuracy for KNN, SVM and TD-LSI technique for TF.IDF and LSI

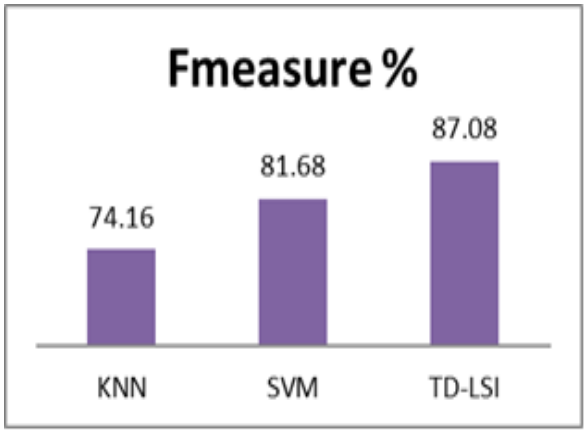

Fig 4: F-Measure for KNN, SVM and TD-LSI technique

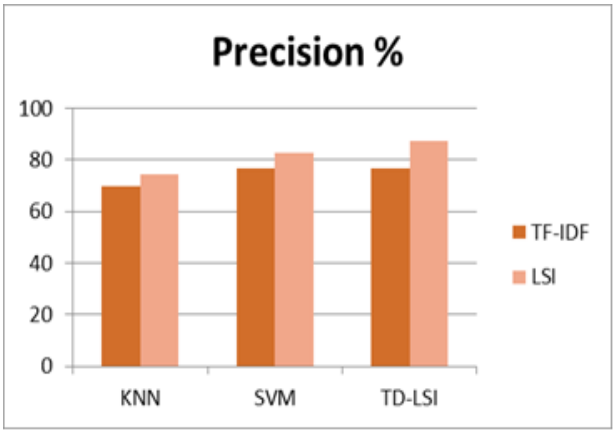

Fig 6: Comparison of Precision for KNN, SVM and TD-LSI technique for TF.IDF and LSI

The performance for TF.IDF for TD-LSI, K-nearest neighbour and SVM classifier is shown in table II

TABLE II: Performance for TF.IDF of Three Classifiers

	TF-IDF			
Classifier	**Accuracy**	**Precision**	**Recall**	**F measure**
KNN	80.56	69.79	71.09	70.16
SVM	83.33	76.55	75.46	75.15
TD-LSI	85.19	76.94	84.73	76.89

The figure 5-8 shows the comparison graph of overall performance metrics for TF.IDF and LSI vectors for the three classifiers.

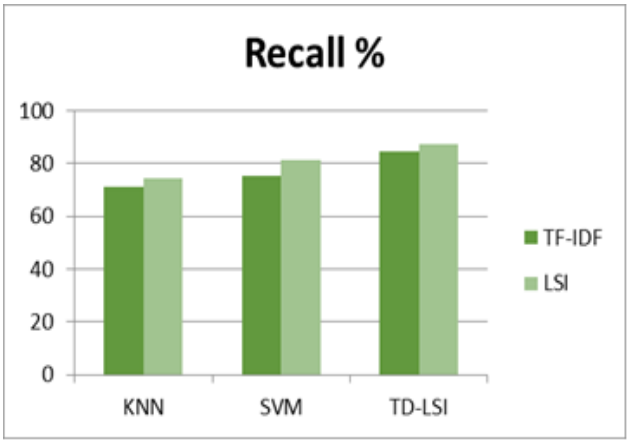

Fig 7: Comparison of Recall for KNN, SVM and TD-LSI technique for TF.IDF and LSI

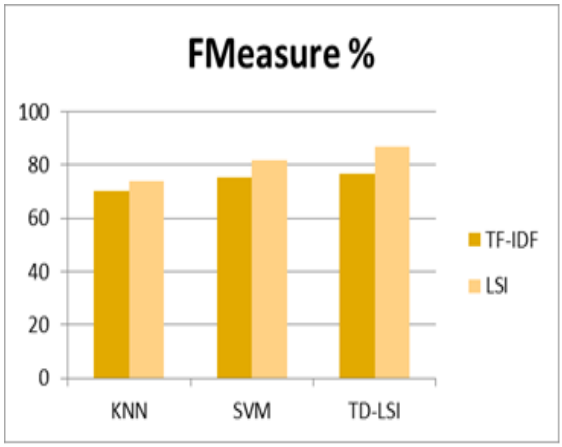

Fig 8: Comparison of F-measure for KNN, SVM and TD-LSI technique for TF.IDF and LSI

5. CONCLUSIONS

In this paper, a novel scheme for categorizing the texts has been represented using LSI. The performance metrics so obtained in LSI is better than those obtained in TF.IDF. It also provides a comparative analysis between two already existing classifiers along with the TD-LSI one. The results of the experiments showed us that the results are better in case of LSI which shows the superior classification and hence enhanced efficiency. Enhanced efficiency means that more documents are classified correctly with low cost of computation and with lesser dimensions.

As a future work, we recommend inspecting the performance of multi-level classification of texts and indent to compare the TD-LSI scheme with other feature reduction methods and weighting schemes on real data sets.

REFERENCES

[1] Lijuan, C.; Hofmann, T. (2004): Hierarchical document categorization with support vector machines. 13th ACM international conference on Information and knowledge management, pp. 78-87.

[2] Martinez-Caro, J. M.; Aledo-Hernandez, A. J.; Guillen-Perez, A.; Sanchez-Iborra, R.; Cano, M. D. (2018): A Comparative Study of Web Content Management Systems. Information, 9(2), pp. 27.

[3] Symeonidis, P., Kehayov, I., & Manolopoulos, Y. (2012, September). Text classification by aggregation of SVD eigenvectors. In East European Conference on Advances in Databases and Information Systems (pp. 385-398). Springer, Berlin, Heidelberg.

[4] Nugumanova, A., & Baiburin, Y. (2014, October). Using SVD for text classification. In Proceedings of International Academic Conferences (No. 0702094). International Institute of Social and Economic Sciences.

[5] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[6] Carpineto, C., Osiński, S., Romano, G., & Weiss, D. (2009). A survey of web clustering engines. ACM Computing Surveys (CSUR), 41(3), pp. 17.

[7] Andrews, H., & Patterson, C. (1976). Singular value decompositions and digital image processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(1), pp. 26-53.

[8] Zhang, W., Yoshida, T., & Tang, X. (2008, October). TFIDF, LSI and multi-word in information retrieval and text categorization. In Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on (pp. 108-113). IEEE.

[9] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), pp. 11-21.

[10] Spärck Jones, K. (2004). IDF term weighting and IR research lessons. Journal of documentation, 60(5), pp. 521-523. Kuralenok, I., & Nekrest'yanov, I. (2000). Automatic document classification based on latent semantic analysis. Programming and Computer Software, 26(4), pp. 199-206.

[11] Price, A. Z. R. J. (2003, April). Document categorization using latent semantic indexing. In Proceedings 2003 Symposium on Document Image Understanding Technology, UMD (p. 87).

[12] Zelikovitz, S., & Marquez, F. (2005). Evaluation of Background Knowledge for Latent Semantic Indexing Classification. In FLAIRS Conference (pp. 868-870).

[13] Lee, C. H., Yang, H. C., & Ma, S. M. (2006, August). A novel multilingual text categorization system using latent semantic indexing. In Innovative Computing, Information and Control, 2006. ICICIC'06. First International Conference on (Vol. 2, pp. 503-506). IEEE.

[14] Kontostathis, A. (2007, January). Essential dimensions of latent semantic indexing (lsi). In System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on (pp. 73-73). IEEE.

[15] Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University-Computer and Information Sciences, 29(2), 189-195.