A Novel Approach for Document Categorization Based on Latent Semantic Indexing

Mamta Rani[1], Mr. Gagan Dhawan[2]

[1,2] *Computer Sci. &Engg Modern Institute of Engg. & Tech. Kurukshetra University, India*

Abstract- **The intensive expansion of the web and the enlarged number of users has forced new organizations to place their processed data on the web. 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 web clustering algorithms [8].Latent Semantic Indexing is a better textual representation technique as it maintains semantic information between the words. Hence, we used the singular value decomposition (SVD) methods to extract the textual features based on LSI. The LSI also knew LSA. In our experiments, we conducted comparison between some of the well –known classification methods such as Naïve Bayes, k-Nearest Neighbours, Neural Natwork, Random Forest, Support Vector Machine, classification tree. A Novel Approch for document categorization based on LSI in which initially start work on contains Topic and then Topic contains the folders and folders contain categories after that a document will be created.**

Index Terms- **Document Categorization, Tokenizing, preprocessing, Term Finding, VSM (Vector Space Model), Clustering, LSA or LSI, SOM.**

1. INTRODUCTION

Web-Content Management System

A WCMS is data management software, 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. Usually the software provides authoring tools designed to allow users with little or no knowledge of programming languages or mark-up languages to create and manage content with easiness. With the advent of technology man is attempting for relevant and optimal results from the web through search engines. The information can be retrieved in web based documents by searching making the use of keywords, documents categorization and also filtering out the stream. Based on the user requirements, various methods were developed for the automatic grouping of web documents, keeping in mind the reduction of the time and efforts to find the information sought after. Keywords and relevant phrases increase the effectiveness and efficiency of the search process.

Document Classification: To model each topic, use a set of documents that are reclassified to a specific topic. The document newly created are classified and then assigned to existing topics with similar models. The classification methods used are the back propagation neural networks [6], naïve Bayesian [3] or support vector machine [19].. Document clustering: To form a topic group uses a set of unclassified documents to extract the relations among them and organize the similar ones together. The methods used are partition-based clustering [4], agglomerative / divisive hierarchical clustering [7] and self-organizing map (SOM) [5]. Document clustering has been applied in various areas such as browsing large set of documents and expanding search space. The proposed Latent Semantic Analysis (LSA) method is a natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

II. LITERATURE REVIEW

Gonzalo Navarro,[1]“Text Document”, Encyclopedia of Database Technologies and Applications, Idea Group Inc., Pennsylvania, USA.

Latent semantic indexing, sometimes referred to as latent semantic analysis, is a mathematical method developed in the late 1980s to improve the accuracy of information retrieval. It uses a technique called singular value decomposition to scan unstructured data within documents and identify relationships between the concepts contained therein. Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text [8]. They asserted that LSA could serve as a model for the human acquisition of knowledge. From the original application for retrieving information, the use of LSA has evolved to systems that more fully exploit its ability to extract and represent meaning. LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

SVD or two-mode factor analysis

Furnas, G.W.,[20]Landauer, T.K., Gomez, L.M., and Dumais, S.T. Statistical semantics: Analysis of the potential performance of key-word information systems. The latent semantic structureanalysis starts with a matrix of terms by documents. This matrix is then analyzed by singular value decomposition (SVD) to derive our particular latent semantic structure model. Singular value decomposition is closely related to a number of mathematical and statistical techniques in a wide variety of other fields, including eigenvector decomposition, spectral analysis, and factor analysis. We will use the terminology of factor analysis, since that approach has some precedence in the information retrieval literature.

The traditional, one-mode factor analysis begins with a matrix of associations between all pairs of one type of object, e.g., documents [16]. This might be a matrix of human judgments of document to document similarity, or a measure of term overlap computed for each pair of documents from an original term by document matrix. This square symmetric matrix is decomposed by a process called "eigen-analysis", into the product of two matrices of a very special form (containing "eigenvectors" and "eigenvalues"). These special matrices show a breakdown of the original data into linearly independent components or "factors". In general many of these components are very small, and may be ignored, leading to an approximate model that contains many fewer factors.

Liu, J., Niu, X.M., Kong, W.H.[23] Data Set Two text collections, Reuters-215783(www.daviddlewis.com/resources/testcollections) and Industry Sector4 (www-2.cs.cmu.edu/afs/cs.cmu.edu), are used in our experiment. Reuters-21578 (Reuters) is the most widely used text collection for text classification. There are total 21578 documents and 135 categories in this corpus. In our experiments, we only chose the most frequent 25 topics and used ”Lewis” split which results in 6314 training examples and 2451 testing examples. Industry Sector (IS) is a collection of web pages belonging to companies from various economic sectors. There are 105 topics and total 9652 web pages in this dataset. A subset of the 14 categories whose size are bigger than 130 is selected for the experiments.

Landauer and Dumais (1997) [24] report an analysis in which LSA was used to simulate a lexical semantic priming study by Till, Mross and Kintsch (1988).

Simulating semantic priming in which people were Presented visually with one or two sentence passages that ended in an obviously polysemous word. After varying onset delays, participants made lexical decisions about words related to the homographic word or to the overall meaning of the sentence. In paired passages, each homographic word’s meaning was biased in two different ways judged to be related to two corresponding different target words. There were two additional target words not in the passages or obviously related to the polysemous word but judged to be related to the overall meaning or “situation model” that people would derive from the passage. Here is an example of two passages and their associated target words, along with a representative control word used to establish a baseline.

"The townspeople were amazed to find that all the buildings had collapsed except the mint." "Thinking of the amount of garlic in his dinner, the guest asked for a mint. "Target words: money, candy, earthquake, breath Unrelated control word: ground

Latent semantic indexing can be summarized as follows:

- A technology developed in the late 1980s for information retrieval, in response to earlier technologies that could not understand synonymy or polysemy.
- A specific approach that tries to grasp the underlying structure of meaning in language.
- Capable of inducing from these findings the hierarchical categories into which terms and concepts fall.
- Originally useful for working on small sets of static documents.

There are various important steps would be used for Document Categorization based on LSI.

A. *Prepare File List*

Trying to find old files is like trying to read your own mind. Based on old files try to maintain useful files. These files arrange in sequence. Eliminate these exasperating and time-consuming mental exercises.

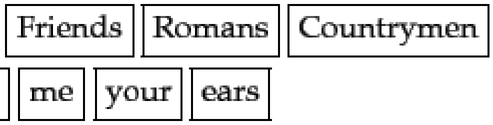
B. Tokenizing

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.

For example:

Input: Friends, Romans, Countrymen, lend me your ears;

Output:



Friends Romans Countrymen lend me your ears

These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A *token* is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

C. *Preproccessing*

The first step in our processing consisted of creating an LSI representation space from the training documents of the Reuters 21578 collection. In preprocessing trying to clean the data text that contains various steps.

- Lowercase Conversion
- Special Symbol Removal
- Stopword Removal
- Stemming/Root words Conversions

And for Root Word Conversion the Porter Algorithm will be used.

D.Term Selection

In term selection select the useful terms.

E.TF-IDF

TF-IDF which stands for Term Frequency-Inverse Document Frequency is a technique for word weighting usually used in information retrieval and text mining. This technique measures statistically the strength of word appearance in a document. TF calculates probability of a term by dividing it by total words in whole documents, while IDF measures a term (word) divide by invers of a document frequency. Document frequency is total documents which the specified term lies there.

tf = f/N..(1)

idf = log(N/df)...............................(2)

tfidf = (f/N) * log(N/df)..................(3)

From equation (1), (2), and (3) above,

tf is term frequency (the number of word occurrence in adocument),

df is document frequency, the number of documents containing the word, N is number of documents. In this research TFIDF is used to measure the strength of word pairs that occurs in whole documents. For example,

the word 'what' will be paired one by one to ot her words, then the paired word which has highest value will be added after the word 'what' will be extended by the predicted word. In the case of extending two words, the 2-highest probability of paired words will be added after the given word.

III. APPLICATIONS OF LSA

The several promising applications of LSA are given below [14]:

- LSA can be used for information retrieval. The LSA woks better than other methods such as standard vector methods when the queries and relevant documents do not share many words.
- Researchers are also carrying out some interesting work of LSA in medical field.
- LSA can be used for information filtering.
- LSA can also be used to return the best matching people instead of document, where people were represented by articles they had written.
- LSA does not depend on literal keyword matching, so it is useful when the text input is noisy, as in OCR (Optical Character Reader), open input, or spelling errors.
- LSA can be used as electronic feedback or e-assessment for e- learning. [11]
- LSA can be used to word sense discrimination within a tutor for English vocabulary learning. [15]

IV. CONCLUSION

This study aims to determine the effect of extending documents terms to the performance of classification. Performance of classification can be measured from its accuracy. In this paper discussed the new method for improving performance of classification by extending documents term. Web content based system has been discussed and a review on latent semantic indexing has been given in this paper. Documents terms are extended using TFIDF weighting method, HMM, and k-means clustering. The result Show that SVM and K-NN Classifiers have almost the same performance.

REFERENCES

[1] Gonzalo Navarro, “Text Document”, Encyclopedia of Database Technologies and Applications, Idea Group Inc., Pennsylvania, USA.

[2] Richard T. Freeman, Hujun Yin, “Web Content Management by SelfOrganization ,” IEEE transactions on Neural Network, vol. 16, No. 5, pp. 1256-1268, Sept. 2005.

[3] BoYu, Zong-ben Xu, and Cheng-hua Li, ”Latent semantic analysis for text categorization using neural network, ” Knowledge-Based Systems 21 pp. 900–904, 2008.

[4] A. McCallum and K.Nigam, “A comparison of event models for naïve bayes text classification,” in Proc. AAAI / ICML-98 workshop on Learning for Text Categorization , 1998

[5] D. Cutting, D. Karger, J. Pederson and J. Tukey, “Scatter/gather: A cluster-based approach to browsing large document collections,” in Proc. 15th Int. ACM/SIGIR Conf. Research and Development in Information Retrieval., Copenhagen, Denmark, 1992, pp. 318–329.

[6] K. Lagus, S. Kaski, and T. Kohonen, “Mining massive document collections by the websom method,” Inf.

[7] Sukhjinder Singh, KamaljitKaur. A Review on Diagnosis of Diabetes in Data Mining. International Journal of Science and Research (IJSR). 2013.

[8] VeenaVijayan V, AswathyRavikumar. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014.

[9] P.Yasodha, N.R.Ananthanarayanan.Comparative Study of Diabetic Patient Data’s Using Classification Algorithm in WEKA Tool . International Journal of Computer Applications Technology and Research Volume 3– Issue 9, 554 - 558, 2014 .

[10] AiswaryaIyer, S. JeyalathaandRonakSumbaly. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015. Sci.., vol. 163, no. 1–3, pp. 135–156, 2004.

[11] Nicholas Evangelopoulos, Xiaoni Zhang, Victor R. Prybutok, “Latent Semantic Analysis: five methodological recommendations”, European Journal of Information Systems (2012) 21, 70–86.

[12] G. Navarro, R. Baeza-Yates, E. Sutinen, and J. Tarhio.“Indexing methods for approximate string matching”. IEEE Data Engineering Bulletin, 24(4):19–27, 2001.

[13] Berry, M. W., Dumais, S. T. and O'Brien, G.W. (1995) Using linear algebra for intelligent

information retrieval. SIAM: Review, 37, 573-595.

[14] Britton, B. K. &Sorrells, R. C. (1998/this issue). Thinking about knowledge learned from instruction and experience: Two tests of a connectionist model. Discourse Processes , 25, 131-177.

[15] Kanjilal, P.P., Dey, P.K., Banerjee, D.N.: Reduced-size neural networks through singular value decomposition and subset selection. Electronics Letters. 29, 1516– 1518 (1993)

[16] Landauer, T. K., Foltz, P. W., &Laham, D. (1998). Latent Semantic Analysis passes the test: knowledge representation and multiple-choice testing. Unpublished manuscript.

[17] Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998/this issue). Learning from text: Matching readers and text by Latent Semantic Analysis. Discourse Processes, 25, 309-336.

[18] Zeno, S. M., Ivens, S. H., Millard, R. T., &Duvvuri, R. (1995). The educator's word frequency guide. Brewster, NY: Touchstone Applied Science Associates.

[19] Studer R, Benjamins V R, Fensel D. Knowledge engineering: principles and methods [J].Data and Knowledge Engineering, 1998, 25(122):161-197.

[20] Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T. Statistical semantics: Analysis of the potential performance of key-word information systems.

[21] Anna Formica. Ontology-based concept similarity in Formal Concept Analysis [J]. Information Sciences,2006,V6(176):2624-2641.

[22] Rodrguez MA, Egenhofer MJ. Determining semantic similarity among entity classes from different ontologies[J].IEEE Transactions on Knowledge and Data Engineering,2003,15 (2):442-456.

[23] Macedche A, Motik B.A mapping framework for distributed ontologies[J] .Web Intelligence and AgentSystem,2003,(1):235-248.

[24] Liu, J., Niu, X.M., Kong, W. H. Data Set Two text collections, Reuters -215783 (www.daviddlewis.com/resources/testcollections)

[25] Landauer and Dumais (1997) report an analysis in which LSA was used to simulate a lexical semantic priming study by Till, Mross and Kintsch (1988).