Document Classification Approach Based on Linear Combination

G.S. Thakur[1], R.C. Jain[2]

ABSTRACT

In this research paper we developed an algebraic approach for document classification. This is the most accurate, efficient, and scalable classification approach that addresses the special challenges of document classification. Algebraic approach is a new document classification method ,comparatively more efficient and accurate. So for, no research was conducted using this concept for Document classification. Our experiments indicate that the accuracy of existing method increase by using algebraic approach. The final goal is achieving high performance and eventually increasing classification accuracy.

Keywords: Classification, LBC, Text mining, document classification.

1. INTRODUCTION

Everyday a vast amount of documents, reports, e-mails, and web pages are generated from different sources, such as enterprises, governments, organizations, and individuals. Approximately 90% of the world's data is held in unstructured formats. This kind of unstructured textual data is usually not stored on relational or transaction database systems, but on web servers, files servers, or even personal workstations.This type of unstructured data has no standard means facility to search, query and analysis. The enterprises spend lots of manpower on organizing these documents into a logical structure for later use. So it is very time-consuming and costly, thus limiting its applicability. Along with the continuously growing volume of information available on the web. They require a systematic and automatic approach for finding, filtering and organizing these documents without human intervention or involvement.Consequently there is an increased interest in developing technologies for automatic text categorization. Text categorization is to automatically assign arbitrary raw documents to predefined categories based on their contents. It is a most widespread application has been for assigning subject categories to documents, to support text retrieval, routing, and filtering. These are the challenges for developing an accurate, efficient, and scalable method for documents classification. To date, the Web has developed a medium of documents for people rather than for data and information that can be processed automatically.

2. RELATED WORK

There are several classification techniques that can be used for text categorization. However, many of these existing schemes do not work better in the text categorization task Researchers have worked on K-Nearest Neighbor(k-NN) Classification, Naïve Bayes Classification, classification decision tree induction algorithm [1,2,3,4,5,6,7,8,9,10,11,12]. Document classification has been studied intensively because of its wide applicability in areas such as web mining, information retrieval. The majority of this information is in text format, for example, emails, news, web pages, reports, etc. Organizing them into a logical structure is a challenging task. More recently, classification is employed

[1,2]Department of Computer Appilications, Samrat Ashok Technological Institute, Vidisha [M.P.]

for browsing a collection of documents or organizing the query results. The standard classification techniques such as k-means, Support Vector Machines, Naive Bayes, Decision Trees, can be applied for document classification.

They usually do not satisfy the special requirements for classification documents: all these approaches suffer from lack of high performance and high accuracy. In addition, many existing document classification algorithms require the user to specify the number of category as an input parameter. In some document sets, category sizes may vary from few to thousands of documents. This variation tremendously reduces the resulting classification accuracy for some of the state-of-the art algorithms. But there are still problems to be tackled such as efficiency and accuracy. Incorrect estimation of the value always leads to poor classification accuracy. Furthermore, many classification algorithms are not robust enough to handle different types of document sets in a real-world environment. Owing to wide significant applicability of text categorization and challenges in the area motivated us to do work in this field. The poor classification accuracy and the weaknesses of the standard classification methods formulate the goal of this research.

3. Methdology

Classification-an important task of data mining is to assign objects to predefined categories or classes – a process called Classification. The input to the classification system consists of a set of example records, called a training set, over several fields or attributes. Attributes are continuous, coming from an ordered domain, or categorical, coming from an unordered domain. One of the attributes, called the classifying attribute, indicates the class or label to which each example belongs. The goal of classification is to induce a model from the training set that can be used to predict the class of a new record.

3.1 Preprocessing

Preprocessing means transform documents into a suitable representation for classification task. In the preprocessing step we perform the following task Remove stop words: - *Stop words are non-informative words* like is, am, are, not, then, the, who, how etc. *Filtering* methods remove words from the dictionary and thus from the documents. Perform word stemming: *Word stemming means remove suffix*. For example walked, walker, walking is replaced by root word walk in documents. In other words *Stemming* methods try to build the basic forms of words, i.e. strip the plural 's' from nouns, the 'ing' from verbs, or other affixes. A stem is a natural group of words with equal (or very similar) meaning. After the stemming process, every word is represented by its stem.So preprocessing is necessary steps because it removes all those words and characters those are not useful in classification.

3.2 Feature Selection

The major problem in text mining is the very high dimensionality of the feature space. Because every unique term, either text or non-text, adds one dimension to the pace, it can easily reach tens and even hundreds of thousands. Only a small part of the terms are feature terms determining a document class, while the rest are unused terms that make the result unreliable and increase computational time. A common approach in dealing with the problem is feature selection.

Feature selection is the process of removing indiscriminative terms from documents in order to improve classification accuracy and reduce computational complexity. Several feature selection methods can be

applied as the preprocessing step for text classification [5] compares methods.

"Feature selection means remove non-informative terms from documents. Feature selection improves classification effectiveness and reduces computational complexity".

The document categorization problem is often characterized by the high dimensionality of the feature space. In a document collection, each unique term represents one dimension in the feature space. For a typical collection of documents, the number of unique terms that occur in all documents, can be several thousands, it is about 50,000 terms for our collection. It is highly desirable to select a subset of the best terms from the entire feature space to discriminate classes, without losing categorization accuracy.

Feature Selection () Algorithm

After stop word removing and word stemming the document has number of unused terms i.e. those terms that has minimum frequency[5] .So we select only those words that have maximum frequency for this purpose we use a threshold value . If the term frequency [5] is greater or equal to the threshold value then this term will be in document and other terms will remove from the document If the term frequency is less then to the threshold value. This method decreases the number of words in the document.

Algorithms:

1) Input:
2) Documents d_i
3) Threshold value L
4) Output: A list T of valid text term
5) T=NULL;
6) While (!eof(d_i))
7) {
8) Extract the term w from the document d_i
9) N=termfrec(w) ;
10) If($N \geq L$)
11) $T = T \cup w$
12) }

3.3 Document Representation

The Vector Space Model (VSM) represents documents as vectors in *m*-dimensional space. Each document *d* is described by a feature vector di=(0 or 1 | wi €di.).According to this model; a document is represented by a vector in which each component is a value that indicates 0 or 1.Thus, documents can be compared by use of simple vector operations and even queries can be performed by encoding the query terms similar to the documents in a query vector. The query vector can then be compared to each document and a result list can be obtained by ordering the documents according to the computed similarity. The main task of the vector space representation of documents is to find an appropriate encoding of the feature vector. Each element of the vector usually represents a word (or a group of words) of the document collection, i.e. the size of the vector is defined by the number of words (or groups of words) of the complete document collection. With this model we can represent whole class in the form of two-dimension matrix. Almost all classification algorithms use this model for classification of text documents.

4. Linear Combination-based Classification

Linear Combination of Vectors is defined as

If the $\propto_1, \propto_2, \propto_3, \propto_4, \propto_5, \propto_6, \propto_7, \propto_8, \propto_9, \ldots\ldots\ldots\ldots \propto_n$ are the vectors then

$\propto = a_1 \propto_1 + a_2 \propto_2 + a_3 \propto_3 + a_4 \propto_4 + a_5 \propto_5 + a_6 \propto_6 + \ldots\ldots\ldots\ldots + a_n \propto_n$

or $\propto = \Sigma a_i \propto_i$

where 1<=i<=n

and $a_{i \in R}$

Let X be the Unknown document and $a_{i \in R}$, $\alpha_{i \in C1}, \beta_{i \in C2}$ then the Linear Combination of Vectors for document classification as given.

$X=a_1\alpha_1+a_2\alpha_2+a_3\alpha_3+a_4\alpha_4+a_5\alpha_5+a_6\alpha_6+\ldots\ldots\ldots\ldots+a_n\alpha_n$

$X=\sum a_i\alpha_i$

$X= a_1\beta_1+a_2\beta_2+a_3\beta_3+a_4\beta_4+a_5\beta_5+a_6\beta_6+\ldots\ldots .a_n\beta_n$

$X=\sum a_i\beta_i$ where 1<=i<=n

and $a_{i \in R}$, $\alpha_{i \in C1}, \beta_{i \in C2}$

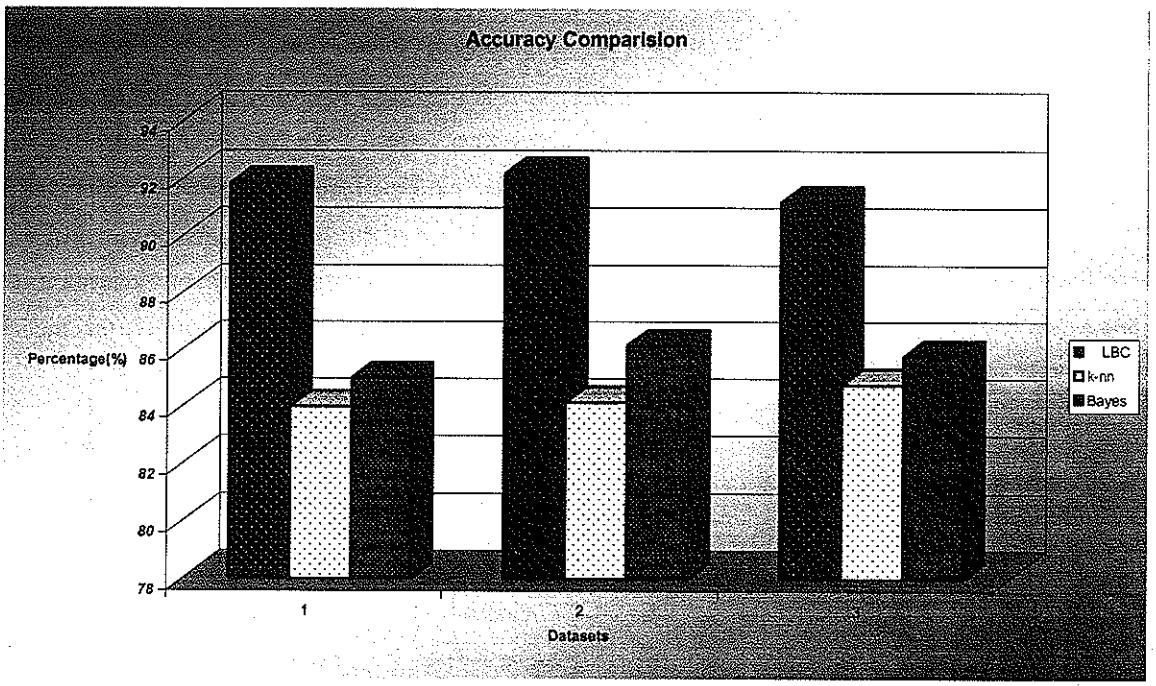
We assign the unknown document X to the class C1 if the solution of the equation $X=\sum a_i\alpha_i$ is consistent otherwise x assign to the class C2.

5. Experimental Result

We have downloaded dataset from UCI KDD Archive **[http:/kdd.icsiuci.edu/]**. This is an online repository of large datasets which encompasses a vide variety of data types, analysis tasks, and applications area. The primary goal of this repository is to enable researchers in knowledge discovery and data mining to scale existing and future data analysis algorithm to very large and complex datasets. In this repository dataset for text categorization is 20 newsgroups. This data set consists of 20000 messages taken from Usenet newsgroup. There is a subset of this newsgroup is mininewsgroup (data file –minnewsgroup.tar.gz). This subset composed of 100 articles from each newsgroup.

The experimental result is shown in the table1.

Table1



6. Conclusion

In this paper we have developed a new Document classification method that is called LCB. We conducted extensive comparative experiments on standard text collections (the 20-Newsgroups). We experimentally predict unknown document class and the results high accurate and efficient. The results show that LCB better than other widely used techniques.The real datasets used in experiments is the standard Text dataset (Text dataset is available from the UCI machine –learning repository) the experiments have also been performed on the synthetic dataset.

References

[1] Ana Cardoso-Cachopo Arlindo L. Oliveira, *"Empirical Evaluation of Centroid-based Models for Single-label Text Categorization"*,INESC-ID Technical Report 2006.

[2] Ken Williams, *"A Framework for Text Categorization"*, Web Engineering Group The University of Sydney Bldg J03, Sydney NSW 2006.

[3] Makoto, *"Hierarchical Bayesian Clustering for Automatic Text Classification"*, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.

[4] Andreas Hotho, *"A Brief Survey of Text Mining"* KDE Group University of Kassel May 13, 2005.

[5] Huan Liu and Lei Yu, *"Toward Integrating Feature Selection Algorithms for Classification and Clustering"*, Knowledge and Data Engineering, IEEE Transactions, Vol. 17, pages 491- 502, April 2005.

[6] Huang. J and Ling. C, *"Using AUC and Accuracy in Evaluating Learning Algorithms"*, IEEE Trans. on Data and Knowledge Engineering,17 3 pp 3:299-310, 2005.

[7] Kathrin Eichler, *"Automatic Classification of Swedish Email Messages"*,17th August 2005.

[8] Kiritchenko. S, Matwin. S, Nock. R and Famili. A, *Learning and Evaluation in the Presence of Class Hierarchies"*, Application to Text Categorization. Submitted, 2005.

[9] Batista, Gustavo E. A. P. A, Ronaldo C. Prati, Maria, *"A study of the behavior of several methods for balancing machine learning training data"*, In: SIGKDD Explorations. Carolina Monard 2004 .

[10] Cai. L and Hofmann. T, *"Hierarchical Document Categorization with Support Vector Machines"*, In Proceedings of the ACM Conference on Information and Knowledge Management, pages 78-87, 2004.

[11] David D. Lewis and Yiming Yang etl, *"RCV1: A New Benchmark Collection for Text Categorization"*, Research Journal of Machine Learning Research, 5 2004 pp 361-397, 5, 2004.

[12] Dekel. O, Keshet. J., and Singer, Y, *"Large Margin Hierarchical Classification"*, In Proceedings of the International Conference on Machine Learning ICML, pages 209-216, 2004.

[13] Guo, Gongde,Yaxin Bi, Kieran Greer, *"An k-NN Model based Approach and Its Application in Text Categorization"*, In: The 5th International Conference on Computational Linguistics and Intelligent Text Processing, LNCS 2945, 2004.

[14] Lewis. D, Yang .Y, Rose. T and Li. F, RCV1, *"A New Benchmark Collection for Text Categorization"*, Research, Journal of Machine Learning Research, pp 5:361-397, 2004.

Author's Biography

G.S. Thakur passed B.Sc.with honors from Dr. Hari Singh Gour University Sagar (M.P.) MCA with honors from Pt. Ravi Sankar University Sagar (C.G.) and is pursuing Ph.D from Barkhatullah Viswavidyalaya Bhopal (M.P.) in Computer Science.At present he is a lecturer in Department of Computer Application, Samrat Ashok Technological Institute, Vidisha (M.P.), India. His area of interests are in Data Mining, Text mining, Web Mining, Machine learning, information retrieval, Data structure, Object Oriented system, Operating System, Data base management system. He has so for published 15 papers in International, national journals and conferences.

Dr. R.C.Jain received B.Sc. ,M.Sc.,M.Tech. and Ph.D. degree. At present he works as a Professor & Head of Department of Computer Application, Samrat Ashok Technological Institute, Vidisha (M.P.), India. His area of interests are in Simulation and Modeling, Computer Graphics,Computer Optimization, Data Mining, Text mining, Web Mining,Operating System, Data base management system. In these areas he has published about 200 papers in International journals, national journals and conferences.