A Study on the Performances of Representation Strategies Handled For Text Categorization

Dr. K. Meenakshi Sundaram,
Associate Professor of Computer Science,
Erode Arts and Science College,
Erode-638009, India.
e-mail: lecturerkms@yahoo.com

K. Ramya,
Research Scholar of Computer Science,
Erode Arts and Science College,
Erode-638009, India.
e-mail: anuramgowri.msc@gmail.com

Abstract--Text mining is the process of deriving high quality information from text. High quality of information typically derived through the devising of patterns and trends through means such as statistical pattern learning. Categorization is the process in which ideas and objects are recognized, differentiated and understood. Text categorization is a popular research area that has been lot of research work undergone on it. There are three types of text representation approaches in text categorization namely, Keyword based approach,Phrase based approach and Pattern based approach..Keyword based representationapproachalso referred asTerm based approach. It extracts the bag of words and stores in the vector space. In this approach we calculate frequency of terms. In phrase based approach we have more than one word instead of single word. In pattern based approach, it does not have any vector space but it has the low frequency problem. To overcomethis low frequency problem,we use inner pattern evaluation and shuffling method. This paper attempts a study on the various text classification methods and representation strategies for text categorization.

Keywords--*K-Nearest Neighbour, Decision tree, Support Vector Machine, Keyword based representation, Phrase based representation, and Pattern based representation.*

*****

1.INTRODUCTION

Vast amount of new information and data are generated everyday through economic, academic and social activities, much with significant potential economic and social value, techniques such as text and data mining and analytics are required to exploit this potential.

Text mining also referred as text data mining roughly equivalent to text analytics, refers to the process of deriving high quality information from text. Text mining is to process unstructured(textual) information, extract meaningful numeric indices from the text. Most general terms, text mining will “turn text into numbers” (meaningful indices), which can be incorporated in other analyses such as predictive data mining. The process of analyzing text means extract information from it for particular purposes

Text categorization is the one of the well studied problem in data mining and information retrieval. Categorization is the process in which ideas and objects are recognized, differentiated and understood. Categorization implies that objects are grouped into categories, usually for some specific purpose. A category illuminates a relation between the subjects and objects of knowledge. The data categorization includes the categorization of text, image, object voice, etc. With the rapid development of the web, large numbers of electronic documents are available on the internet. Text categorization becomes key technology to deal with and organize large numbers of documents. Most of text documents belong to one of this group like newspapers, letters, journals, articles, reports, etc

Now a days, text categorization plays with a important role of text mining. Text categorization is a essential topic due to high growth of electronic documents. In previous days knowledge engineering technique manually categorize text documents, based on some logical rules. Some text categorization areas are sports, art, politics, education, medicine, etc. Information retrieval provides the term based approach used for to solve challenges such as rocchio, probabilistic model, rough set model and svm is based on filtering model. Term based approach was suffers from the polysemy and synonymy. Polysemy is refers to one word has the different meanings. And synonymy means more than one word has same meaning. After the term based approach we use the phrase based approach it gave the better performance to compare the term based approach. Phrase carries the semantics. The limitations of the phrase based approach low frequency problem, redundant and noisy patterns. To overcome this problem we used pattern taxonomy model. In presence sequential patterns are used in data mining. PTM also occurs the low frequency and the misinterpretation problem. So we use the inner pattern evaluation.

2. Applications of text classification

Document classification may appear in many applications:

2.1 Email filtering: Systems for filtering a person’s incoming emails to weed out scam or to categorize them into different classes are just now becoming available.

2.2 News filtering and organization: Most of the news services today are electronic in nature in which a large volume of news articles are created every single day by the organizations. In such cases, it is difficult to organize the news articles manually. Therefore, automated methods can be very useful for news categorization in a variety of web portals. This application is also referred to as text filtering.

2.3 Document organization and retrieval: The above application is generally useful for many applications beyond news filtering and organization. A variety of supervised methods may be used for document organization in many domains these include large digital libraries of documents, web collections, scientific literature or even social feeds. Hierarchically organized document collections can be particularly useful for browsing and retrieval.

2.4 Opinion mining: Customer reviews or opinions are often short text documents which can be mined to determine useful information from the review.

2.5 News monitoring: In knowledge-based companies like the stock exchanges, a number of persons is concerned with the scanning of news papers and other information sources for items which are concerned with the national or international economy or with individual companies on the stock market. The results are sent to the person who should be informed.

2.6 Narrow casting: Press agencies strive to give more and more individual service, where each client obtains the large stream of outgoing news items only that are relevant.

3. TEXT CATEGORIZATION PROCESS

Various steps of text categorization process are depicted in Fig 1.

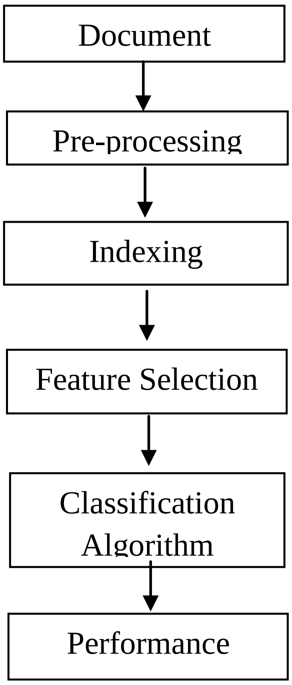

Fig 1:Process of text categorization

3.1 Document Collection

Document collection is refers to collect the different types of text document in the different sources. Like pdf files, word document, web documents, etc[26]

3.2 Pre-processing

Before select the features from the text document we use the following two steps on the data source. Those steps are pruning of infrequent words and the pruning of high frequent words. Perform many task to prune this infrequent and high frequent words.

First tokenize the document based on white spaces. Second step is called stemming which find out the root word for the words. It reduces the dimensionality(number of features) for example: flying, flew=fly. Third step is remove stop words which performs the grammatical functions that include preposition, conjunction, particle, auxiliary verb and so on.Stop words are like the, a, an, with, for, but, and, so on. Common algorithm for removing stop words porters algorithm and KSTEM algorithm.

A text retrieval system often associates a stop words that referred “irrelevant words”. Term frequency calculation is based on number of occurrences of terms in the document. Advantage of this method is it reducing the dimensionality of dataset by removing the irrelevant features. And improve classification accuracy and also reducing the over fitting.

Various steps of pre-processing is depicted in Fig 2,

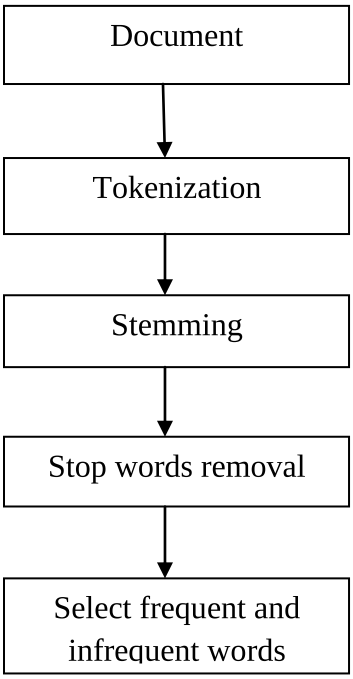

Fig 2: Process of text pre-processing

3.3 Indexing

The documents representation also a one of the pre-processing technique. It reduces the complexity of the documents it makes them easier to handle, and one of main process is all document have to be transforms full text into a document vector. Most commonly used document representation is called vector space model (SMART) [26]. In the vector space model, Documents are represented by vectors of words. Those words are commonly known features. Usually, Collection of documents are represented by word by word document matrix. Bag of words or vector space model representation has some major limitations like high dimensionality, loss of correlation with adjacent words and low semantic relationship that exist among the terms in a document. To overcome these problems, term weighting methods are used.

3.4 Feature Selection

The next step of text categorization process is feature selection [26]. Vector space improves the scalability, efficiency and accuracy of a text classifier. Feature selection actually select the subset of features from the original documents. Feature Selection is performed by keeping the words with highest score. Major problem of the text classification is high dimensionality of the feature space.

3.5 Classification

Documents are automatically classified based on predefined categorization. The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. We use many classification techniques to classify these documents[26]. Like Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Rocchio's. Some of techniques are described in section 4.

3.6 Performance Evaluations

The last step of the process is performance evaluation. An important issue of text categorization is how to measures the performance of the classifiers. Many measures have been used, like Precision and recall [26]; fallout, error, accuracy etc.

The classification is performed in free text documents. Feature space reduction, tokenization, indexing are performed in feature selection. Term frequency and inverse document frequency are used in tokenization process. Many categorization methods are there.

4. CLASSIFICATION METHODS

4.1 K-Nearest Neighbour

K-NN classifier is a case-based [7] learning algorithm it is based on Euclidean distance or Cosine similarity measure's. This method has many advantages[13] like effectiveness, non-parametric and easy to implements properties, if the classification time is long and difficult to find optimal value of k. But larger values of k reduces the effect of noise in the classification, K can be selected by various heuristic techniques. To overcomethis drawback, traditional KNN was modify with different K-values for different classes rather than fixed value for all classes. One main drawback of similarity measure used in k-NN is that it uses all features in computing distances. In many document data sets, smaller number of the total vocabulary may be useful in categorizing documents. To overcome this problem is to learn weights for different features. Fang Lu QingyuanBai[6]propose Weight Adjusted k-Nearest Neighbor (WAKNN) classification algorithm is based on the k-NN classification paradigm. With the help of KNN can improve the performance of text classification [13] from training set and also accuracy can improve with combination of KNN [16] with another method.

4.2 Decision tree

Decision trees are designed with the use of hierarchical division of the underlying data space with the use of different text features. The hierarchical division of the data space is designed in order to create class partitions which are more skewed in terms of their class distribution. For a given text instance, determine the partition that it is most likely to belong to, and use it for the purposes of classification. Decision tree is used for text classification. This internal node are referred by term, branches retiring from them are labelled by test on the weight, and leaf node are represent corresponding class labels . Decision tree can categories the text document by running through the query structure from root to until it reaches a certain leaf node, which represents the goal for the categorize of the document. many training data sets are not fit to memory. decision tree construction is inefficient because of swapping the training tuples. To handle this issue mnish Mehta [19] presents method which can handle numeric and categorical data. andPeeraponVateekul proposing [22] as FDT to handle the multi-label document witch reduce cost of induction, and [10] presented decision-tree-based symbolic rule induction system for text categorization which also improves text classification.

4.3 Support Vector Machine (SVM)

Text Classification method has been suggest by [11]. The SVM require both positive and negative training set. But it is not suitable for other classification methods. These positive and negative training set are needed for the SVM to take the decision facade that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. SVM classifier method is exceptional from other with its effectiveness [14] to improve performance of text classification [3] combining the HMM and SVM where HMMs are used to as a feature extractor and then a new feature vector is normalized as the input of SVMs, so the trained SVMs can classify unknown texts successfully, also by combing with Bayes [15] use to reduce number of feature which as reducing number of dimension. SVM is more capable [5] to solve the multi-label class classification

4.4 Neural network classifier

Neural networks are used in wide variety of domains for the purposes of classification. In the context of text data, the main difference for neural network classifier is to adapt these classifiers with the use of word features. Note that neural network classifier are related to SVM classifiers, indeed they both are in the category of discriminative classifiers which are in contrast with the generative classifiers. A neural network classifier is a set of connections of units, where the input units usually symbolize terms, the output units represents the category. For classifying a test document, its term weights are assigned to the input units; the establishment of these units is propagated forward during the network, and the value that the output units takes up as a importance determines the categorization decision. Some of the researches use the single-layer perceptron, appropriate to its simplicity of implementing [5]. The multi-layer perceptron which is more complicated, also widely implemented for classification tasks[17].Models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [4] for documents classification. An well-organized feature selection method [9] is used to reduce the dimensionality as well as improve the performance. New Neural network based document classification method.[1]Was presented, which is helpful for companies to manage obvious documents more effectively.

4.5 Bayesian classifiers

In Bayesian classifier (also called as generative classifier) attempt to build a probabilistic classifier based on modelling the underlying word features in different classes. The idea is then to classify text based on the posterior probability of the documents belonging to the different classes on the basis of the word presence in the documents. Naïve bias method is kind of module classifier [24] priori probability and class conditional probability. It is used to calculate the probability that document D is belongs to class C. Naive Bias as multivariate Bernoulli and multinomial model. Out of these model multinomial model is more suitable when database is large, but there are identifies two serious problem with multinomial model first it is rough parameter estimated and problem it lies in handling rare categories that contain only few training documents. They [23] propose Poisson model for NB text classification and also give weight enhancing method to improve the performance of rare categories. Modified Naive Bayes is propose [29] to improve performance of text classification, also [18] provides ways to improve naive Bayes classification by searching the dependencies among attribute. Naïve Bayes is easy for implementation and computation.

5.REPRESENTATION TECHNIQUES

5.1 Keyword based representation

Text mining often involves extract the keywords from the document. Keywords are mainly used to text categorization, based on keywords set of documents are sorting into categories. Bag of words is one of the keyword based representation. It referred as term based approach. Bag of words is the simplicity method. It is used in natural language processing and information retrieval. Text is represent as bag(i.emultiset) of its words[2].

Example:

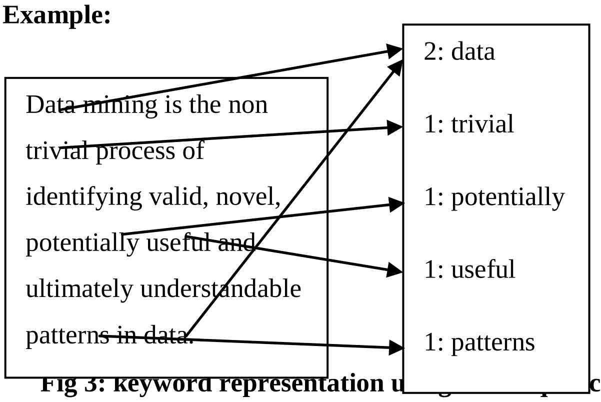

Fig 3: keyword representation u cy

In this example each word in the document is retrieved and stored in the vector space with its frequency. The context of this document is represented by these words known as features. The main drawback of this method is it does not have any relationship between them. Problem of keyword based representation is synonyms and homonyms. Synonyms means more than one word has the same meaning. Homonyms means one word have more than one meanings. Another drawback of this method is overfitting and selecting the limited number of features.

Usage example:

Bag of words technique is used in spam filtering. One bag of word contains the spam related words (like stock, buy, etc) and another one contains the user friends and world place name, Bayesian spam filter easily classifies the spam messages from the email.

5.2 Phrase based representation

Keyword based representation have some problems. To overcome this problem phrase based representation has been proposed. Phrase based method represents the multiple words(phrase). Phrase contains more specific contents compare to keywords. It extracts the multiword from the document based on syntactical structure. This approach extracts co-occurring terms as long phrase from the document.

Phrase contains more specific contents for instances “filter” and “Information filter” group of words create the meaningful phrases to indicating important concept in the text. Advantage of this method is it discovers the hidden semantic sequences from the documents and it gives the accuracy of the classification. It is hypothesis based approach but it not support hypothesis it is main drawback of this approach. This method has the lower consistency of assignment and lower document frequency for long term phrases. N-Multigram model is related to N-gram model. There are five categories of phrase or term extractions:Co-occurring terms, episodes, noun phrases, key phrase and nGram.

5.3 Pattern based approach

Term based approach and phrase based approach did not yield any effective result so we proposed pattern taxonomy model is based on pattern based approach. It have the two stages. First stage extracts the important phrases from the text document. The term weight is occurring in the extracted pattern is calculated to improve the judgements on the new document. Patterns are contains set of terms that frequently appeared in the paragraph. Pattern taxonomy model is more reliable because it uses the positive training documents and all documents are divided into paragraphs. The term weight is evaluated by the term appearance in the discovered patterns.Two key factor affects the pattern based approach that is low frequently and misinterpretation.Low frequency means if minimum support is decreased. Noisy patterns are affect this model and leads to misinterpretation. It discovers unsuitable information for the user. So we use updating patterns for finding useful and relevant information from the text document by using pattern evolving and deploying methods.

5.3.1 Inner pattern evolution

This method helps to reduce the low frequency problem it only changes the pattern terms with in the pattern because it reshuffle the terms. A threshold is used to classify the documents relevant or irrelevant. These patterns are offenders. There are two types of offenders, complete conflict offender and partial conflict offender. Complete conflict offenders removed from the discovered d-patterns first. Partial conflict offenders reshuffling of their term support is carried out n order to reduce the effects of noise documents[21].

Algorithm 1is used to remove the discovered patterns in the document

```
Input    : a training set $D = D^{+} \cup D^{-}$; $asetofd -$ patterns DP; and an experimental coefficient μ
Output : a set of term support pairs np
np ⟵ Ø;
Threshold = threshold (DP); // eq (5)
Foreach noise negative document nd$\in D^{-}$ do
If weight (nd) ≥ threshold then Δ (nd) ={ p ∈ DP|termset(P) ∩ nd ≠ ∅};
NDP = {β(P)|p ∈  DP};
Shuffling (nd, Δ (nd), NDP, μ, NDP);
Foreach P ∈ NDP do
npnp⊕⟵ p;
End
End
```

Algorithm 1: IP Evolving (D^{+}, D^{-}, DP, μ)

In this method, similarity is measured between the test document and concept is estimated, using the inner product and computer is capable of generating a perfect shuffle. This method gives the better result of discovered pattern which is extracted from the text document.

6. Experimentation & Result

The proposed methodology is experimented with collection of text documents from multiple sources related with different categories.As mentioned, item set-based data mining methods struggle in some topics as too many candidates are generated to be processed.

The following algorithm 2 is used for shuffling the terms of discovered patterns

```
Input    : a noise document nd, its offenders Δ (nd), normal form of d-patterns NDP,    and an experimental co-efficient μ.
Output : updated normal forms of d-patterns NDP

Foreach d-pattern p in Δ (nd) do
If  termset(p) € nd then NDP = NDP- { β (p) }; // remove complete conflict offenders
Else // partial conflict offendrs
    Offering = $(1 - \frac{1}{\mu}) \times \sum_{t \in (\,termset\ (P) \cap nd\,)} support(t);$
    Base = $\sum_{t \in (termset\ (p) - nd\,)} support(t);$
Foreach term t in termset (p) do
        If t∈ nd then support(t) = $(\frac{1}{\mu})$ × support(t); //shrink
        Else // grow supports
            Support (t) = support (t) × ( 1+ offering ÷ base);
  End
End
```

Algorithm 2: Shuffling (nd, Δ (nd), NDP, μ, NDP)

In addition, the result obtained based only on the first 50 TREC topics. It contains the top 20 news groups, MAP, bank profiles, features, etc. Results obtained more practical and reliable since the judgment for these topics is manually made by domain experts, whereas the judgment for the last 50 TREC topics is created based on the metadata tagged in each document.
The proposed approach PTM (SIT), Shuffling Inner Pattern in the pattern taxonomy model. The results of overall comparisons are presented in Table 1, and the summarized results are described in Fig. 5.

Table 1: Comparison of all methods on the first 50 topics

Method	**Top 20**	**b/p**	**MAP**	$F_{\beta} = 1$	**IAP**
PTM(SIT)	0.493	0.429	0.441	0.440	0.466
Phrase Represent	0.447	0.409	0.408	0.421	0.434
Keyword Represent	0.434	0.399	0.401	0.410	0.422

We list in Table 1 since not all methods can complete all tasks in the last 50 TREC topics.

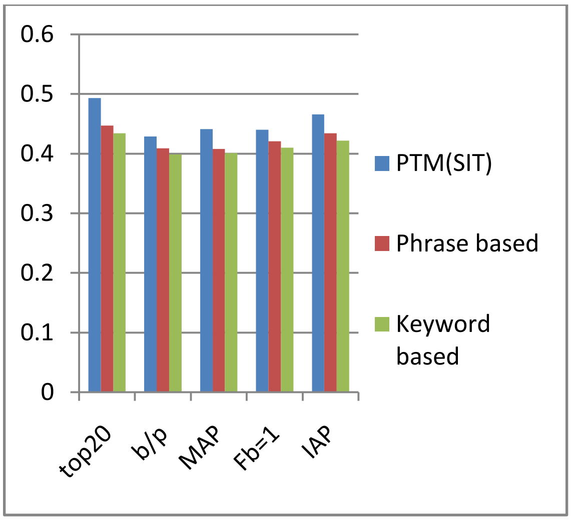

Fig 4 : Comparison of all representation methods on the first 50 topics

The most important information revealed in this table is that our proposed PTM (SIT) outperforms not only thepatternbased methods, but also the termbased methods and phrase based methods. An important issue of text categorization is how to measure the performance of the classifiers. Many measures have been used, like Precision and recall; fallout, error, accuracy etc. Precision and recall are widely used for evaluation measures in text categorization.

4.3.1 Precision

Precision is the measure of the accuracy provided that a specific class has been predicted. Precision is defined as the fraction of the retrieved documents that are relevant, and can be viewed as a measure of the system's soundness, that is:

$$\text{Precision} = \frac{\#\,RelevantRetrievedDocuments}{\#\,RetrivedDocum\ \ ents}$$

4.3.2 Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a dataset. Recall is defined as the fraction of the relevant documents that is actually retrieved, and can be viewed as a measure of the system's completeness, that is:

$$\text{Recall}=\frac{\#\,RelevantRetrievedDocuments}{\#\,RelevantDocuments}$$

4.3.3 Accuracy

Accuracy, which is defined as thepercentage of correctly classifieddocuments,Usually, Accuracy is represented as a real value between 0 and 1.

$$\text{Accuracy} = \frac{\#\,Correctly\ \ Classified\ \ Documents}{\#\,TotalDocuments}$$

7. CONCLUSION

Many data mining techniques are proposed in text representation to classify the text document in last decade.Keyword based representation, Phrase based representation and Pattern based representation techniques are handled in common.

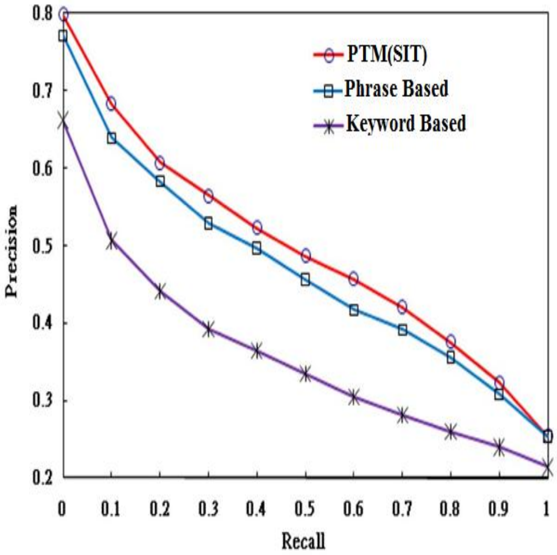

Fig 5: Comparing PTM(SIT) with Phrase and keyword

The main attempt of this paper is applying the inner pattern evaluation method which helps the elimination of thelow frequency problem in text categorization process.

REFERENCES

[1] Amy J.C. Trappey a, Fu-Chiang Hsu a,Charles V. Trappey b, Chia-I. Lin "Development of a patent

document classification and search platform using a back-propagation network", Expert Systems with Applications Vol.31, pp.755–765, 2006

[2] AnishaRadhakrishnan, Mathew Kurian, "Efficient Updating Of Discovered Patterns For Text Mining: A Survey", IJCSNS International Journal Of Computer Science And Network Security, VOL.13(10), October 2013

[3] Chen donghui Liu zhijing, "A new text categorization method based on HMM and SVM", IEEE 2010

[4] Cheng Hua Li , Soon Choel Park "An efficient document classification model using an improved back propagation neural network and singular value decomposition", Expert Systems with Applications, pp.3208–3215, 2009

[5] Dagan, I., Karov, Y., and Roth, D. "Mistake-Driven Learning in Text Categorization." In Proceedings of CoRR. 1997

[6] Fang Lu QingyuanBai, "A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization", IEEE 2010.

[7] GongdeGuo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, "KNN Model-Based Approach in Classification", Proc. ODBASE, pp.986 – 996, 2003

[8] Harish B.S, Guru D.S, Manjunath S, "Representation and Classification of Text Documents:ABrief Review" IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010.

[9] Hwee TOU Ng Wei Boon GohKok Leong Low, "Feature Selection, Perception Learning,and a Usability Case Study for Text Categorization", SIGIR 97 Philadelphia PA,

[10] Johnson D.E,Oles F.J, Zhang T, Goetz T, "A decision-tree-based symbolic rule induction system for text Categorization", by IBM SYSTEMS JOURNAL, VOL 41(3), 2002

[11] Joachims, T."Text categorization with support vector machines: learning with manyrelevant features". In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz,DE), pp.137–142 1998.

[12] Khan A, Baharudin B, Lee L.H, Khan K, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances Information Technology, vol. 1, 2010.

[13] Kwangcheol Shin, Ajith Abraham, and Sang Yong Han, "Improving kNN Text Categorizationby Removing Outliers from Training Set", Springer-Verlag Berlin Heidelberg 2006.

[14] Liu Y.Y.X, "A re-examination of Text categorization Methods" IGIR-99, 1999.

[15] Loubes, J. M. and van de Geer, S "Support vector machines and the Bayes rule in classification", Data mining knowledge and discovery pp.259-275.2002

[16] Methods Ali DaneshBehzadMoshiri "Improve text classification accuracy based on classifier fusion methods". 10th International Conference on Information Fusion, pp.1-6 2007.

[17] MIgual E .Ruiz, PadminiSrinivasn, "Automatic Text Categorization Using Neural networks", Advaces in Classification Research, Vol.VIII.

[18] Michael J. Pazzani "Searching for dependencies in Bayesian classifiers" Proceedings of the Fifth Int. workshop on AI and, Statistics. Pearl, 1988.

[19] Mnish Mehta, Rakeshagrwal" SLIQ: A Fast Scalable Classifier for Data Mining"1996.

[20] MuhammedMiah, "Improved k-NN Algorithm for Text Classification", Department of Computer Science and Engineering University of Texas at Arlington, TX, USA.

[21] Ningzhong, yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE transactions on knowledge and data engineering, vol.24(1), Jan 2012

[22] PeeraponVateekul and MiroslavKubat, "Fast Induction of Multiple Decision Trees in Text Categorization From Large Scale,Imbalanced, and Multi-label Data", IEEE International Conference on Data MiningWorkshops 2009

[23] Sang- Bum Kim, et al, "Some Effective Techniques for Naive Bayes Text Classification "IEEE Transactions on Knowledge and Data Engineering, Vol. 18, November 2006.

[24] SHI Yong-feng, ZHAO, "Comparison of text categorization algorithm", Wuhan university Journal of natural sciences. 2004.

[25] Sholom M. Weiss, ChidanandApte, Fred J. Damerau, David E. Johnson, Frank J. Oles, ThiloGoetz, and Thomas Hampp, IBM T.J. Watson Research Center "Maximizing Text-Mining Performance" pp.1094-7167/99 IEEE INTELLIGENT SYSTEMS. 1999

[26] VandanaKorde, NamrataMahender C, "Text Classification And Classifiers: A Survey"International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3(2), March 2012

[27] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "Text classification based on multi word with support vector machine", knowledge based systems21(2008) pp.879-886

[28] Yiming Yang "An Evolution of statistical Approaches to Text Categorization" Information Retrieval Vol.1, pp.69-90,1999.

[29] YirongShen and Jing Jiang" Improving the Performance of Naive Bayes for TextClassification"CS224N Spring 2003