Journal of Theoretical and Applied Information Technology
10th May 2014. Vol. 63 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PATTERN BASED CLASSIFICATION FOR TEXT MINING USING FUZZY SIMILARITY ALGORITHM

[1]V.SHARMILA, [2]I.VASUDEVAN, [3]DR.G.THOLKAPPIA ARASU

[1]ASSOCIATE PROFESSOR, Department of Computer Science and Engineering, K.S.R.College Of Engineering, Tiruchengode, Tamilnadu, India.
[2]ASSISTANT PROFESSOR, Department of Computer Science and Engineering, V.S.B. Engineering College, Karur, Tamilnadu, India.
[3] PROFESSOR, Department of Computer Science and Engineering, AVS Engineering College, Salem, Tamilnadu, India.
E-mail: [1] sachinsv06@gmail.com [2] ivasu78@gmail.com

ABSTRACT

Pattern mining is an important research issues in data mining with few kinds of applications. Many text mining methods have been proposed for mining useful pattern in text documents. It is mainly focuses to approximately identify the different entities such as terms, phrases and pattern. We use the feature evaluation to reduce the dimensionality of high dimensional text vector. Then the system assigns the frequency to each word, all the weight of the document is used for pattern clustering. Pattern clustering is one of the favorable methods for feature extraction in text classification. In this paper we propose a fuzzy estimated and similarity - based self generating algorithm for text classification. It overcomes the low frequency problem, and also calculates the similarity between the different pattern and word in effective manner. Experimental on RCV1 data collection and TREC topics implement that the proposed result achieves better performance

Keywords−*Text Mining, Feature Clustering, Text Classification, Feature Extraction, Pattern.*

1. INTRODUCTION

Data mining is used to discover the large amount of information from various repositories. It has several applications such as financial data analysis, retail industry, telecommunication industry, and market analysis and business management. It can be useful for knowledge discovery and feature extraction from large amount of data. For example Google have been retrieved millions of data from various repositories. Text mining is the process of extracting the some useful information from different structured and unstructured documents. Now a day's most of the information is available in the form of text. The text is represented by following format such as word, phrase, term, pattern, concept, paragraph, sentence, and document. Most of the existing text mining methods are only suitable for term based and pattern based approach. These approaches are suffering from problems polysemy, and synonymy. How too effectively pattern or word discover from various textual document, till now it is an open research issue in pattern mining. There are two kinds of issues such as low frequency problem, misinterpretation problem. Low frequency problem is usually a specific pattern (short pattern). Misinterpretation problem cannot exactly discover the user expectation.

Most of the data mining and knowledge discovery technique have been extracted interesting information or features from various textual documents. These methods include pattern taxonomy, concept-based model, association mining, sequential pattern mining, relevance feature discovery, iterative learning algorithm have been proposed. These approaches have shown some kinds of improvements in text mining.

In this paper we propose a fuzzy similarity based self generating algorithm. It allows approximate text representation, opposes the incomplete or ambiguous data, conceptually distinct due to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

different text categorization. It is used to reduce the data sets, vague and redundant data.

2. PROPOSED METHOD

The proposed model is an extension of the work. The proposed pattern mining model consists of pattern deploying, inner pattern evolution, and fuzzy estimated and similarity based self generating algorithm for text classification. It is one of the extended fuzzy feature clustering algorithms in text classification. A textual document is input to the proposed system. It is collecting from different sectors such as news papers, articles, journals, magazine, university details and IT industry. Here we can extract the word from text document. Same words are classified into one group. Different words are categorized into another one group. After classification we calculate similarity between different words. Then each word has been weights in a desired number of categorizations are formed automatically. Four ways of weighting, paragraph, document, concept and sentence are discussed. By fuzzy algorithm can properly deals with the weighting scheme between different features. It is very easy to avoiding the low frequency and misinterpretation problem in pattern mining. We conduct some real world experiments on data sets RCV1 and TREC. It become show excellent improvements and also run faster, reduces storage requirements.

2.1 Pattern Reduction. There are two ways of doing pattern reduction, pattern selection, pattern extraction. By pattern selection approaches, a new pattern set $p' = \{p'_1, p'_2, \ldots\ldots, p'_k\}$ is carried. Which is a sub set of the initial pattern set p, where is represented by following format $p=\{p_1, p_2, \ldots\ldots, p_k\}$. Information gain due to a particular splits of pattern p into pi, i= 1, 2, . . . , k is then information $gain(p,\{p_1, p_2, \ldots\ldots, p_k\}) = purity(p) - purity(p_1, p_2, \ldots\ldots, p_k)$, it measures uncertainty of pattern in effective manner, then which can assign the weight to each and every pattern in a given set of text groups.

2.2 Pattern Clustering. Pattern or word clustering is an extended approach for feature reduction. Which can used to categorizes the all words into some groups , here we consider as two kinds cluster group such as inter cluster, intra cluster. Inter cluster means all words present in within boundary. Intra cluster means all words or features present in outside of the boundary.

3 . PATTERN BASED MINING MODEL

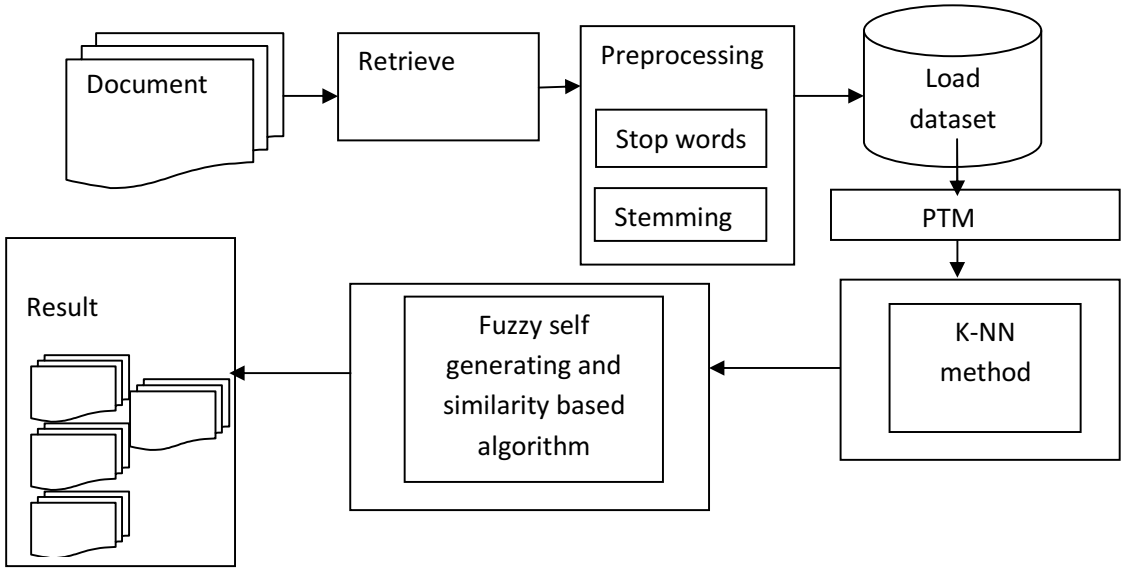

Figure 1: Pattern Based Mining Model

Pattern based mining model consists of following elements such as listed below.

Preprocessing

Finding the concepts

Measuring the word frequency

- Paragraph level

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Document level
- Sentence level

Load dataset

Similarity measure in text by using fuzzy

3.1 Preprocessing

Preprocessing is used to reduce the training time and storage requirements. This is used to remove irrelevant or unwanted words from textual documents. It consists of two types operation such as stemming, stop list; stemming means sequence words have same meaning. For example intelligent, brilliant .these two words have same meaning but word is different. Stop word means irrelevant word or root word. For example {a. an, the, since…}.

3.2 Finding The Concepts

After completion of preprocessing or tokenization step, we are collecting different words from text document. Then which is used to find the similarity between different pattern or word and also classify the relevant and irrelevant document by using pattern taxonomy model.

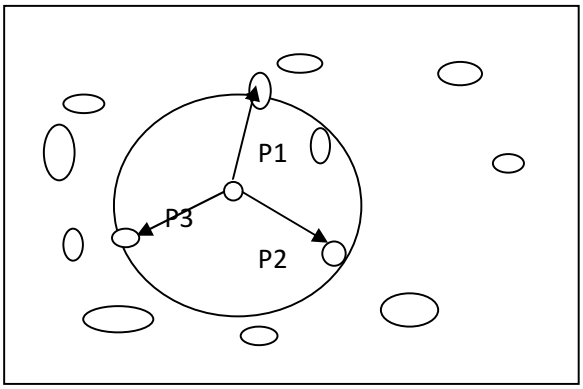

Figure 2: K –NN method

K- Nearest Neighbor classifiers are used to search the pattern boundary with corresponding distance. It discovers the unknown pattern in effective manner. It has one similarity measure such as euclidean distance measure, the euclidean distance between two patterns,

$$\text{dist}(p1, p2) = \sqrt{\sum_{i=1}^{n} (P1i - P2i)^2}$$

This can be effectively measured by distance, and also is computationally efficient.

3.2.1 Procedure

1. Select a pattern from text document.
2. Let W be the first pattern group, W=P1.
3. Sort through the remaining patterns to find the word, W, which is farthest from P1.
4. Let the most distance pattern be the second pattern group, W=P2.
5. Find the distance, T=| P2-P1 |, between the two groups.
6. Dmin(Wj)=min[Di(Wj)],for i=1,2
7. Find Dmax(Ws)=max[Dmin(Wj)] for all j.
8. If Dmax>R/2 then let Ws=Zn+1,
   Otherwise,
   If Dmax<R/2, terminate the procedure.(R is threshold)
9. Growing the no of pattern groups by 1 : Ng=Ng+1
10. Reset the pattern distance ,

$$R = \sum_{i=1}^{n} . \sum_{j=1}^{n} |Pi - Pj| \; / \sum_{k=1}^{n-1} \frac{k(k+1)}{2}$$

11. Return to step 6.

3.3 Measuring the Word Frequency

To measure the frequency at different level such as paragraph level, document level, sentence level.

3.1 Paragraph Level

To measure the each concept at the paragraph level, the number of occurrences of word or pattern in the document is measured.

3.2 Document Level

To measure the each concept at the document level, the number of occurrences of word or pattern in the document is measured.

$$df = \sum_{i=1}^{n} (Pi)$$

for i=1,2…..n.

Journal of Theoretical and Applied Information Technology
10th May 2014. Vol. 63 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3 Sentence Level

To measure the each concept at the sentence level, the number of occurrences of word or pattern in the document is measured. If we generate the document sentences from the rest of sentences in the training set. we can compute the probability using Baye's Rule.

$$P(s \mid f1,f2,\ldots.fk) = \frac{\prod_{j=1}^{k} P(fj|s) * P(s)}{\prod_{j=1}^{k} P(fj)}$$

Where s is a sentence that belongs to document. The k- features {f1,f2,..fk} are the word location, and sentence length features. The probability P(f1|s) is the ratio of the number of times the features f1 appears in a document sentence to the number of document sentences. P(s) is a constant and statistical independence of the k- features.

3.4 Load the Dataset

We uses the experiments on RCV1 (Reuters Corpus Volume 1) and TREC (Text Retrieval Conference) data sets. This is used to store the database before it will send to either training department or testing department.

Example

Table 1: A Set Of Textual Documents

Documents	Words
d1	w2 w3
d2	w4 w5 w6
d3	w4 w5 w6 w7
d4	w4 w5 w6 w7
d5	w2 w3 w6 w7
d6	w2 w3 w6 w7
d7	w1 w2 w3 w4

Table 2: Frequent Words And Covering Sets

Frequent word	Covering sets
{w4,w5,w6}	{d2,d3,d4}
{w4,w5}	{d2,d3,d4}
{w4,w6}	{d2,d3,d4}
{w4}	{d2,d3,d4,d7}
{w6}	{d2,d3,d4,d5,d6}
{w2,w3}	{d1,d5,d6,d7}
{w2}	{d1,d5,d6,d7}
{w3}	{d1,d5,d6,d7}
{w6}	{d2,d3,d4,d5,d6}

3.4.1 Pattern Taxonomy

Here we assume that set of text documents, which can extract the word from each and every document. Each document has some set words.

Let D be a set document, at containing 7 documents d1, d2, d3, ……., d7 with words 7 words such as "office", "system", " line", "internet", "world", "college", "apartment". The resulting word patterns are shown in table -2.

Table -1 has set of document d.

Where di (w) = {d1, d2, d3, …., d7} and duplicate words are removed from textual documents. Let min – sup = 60%, we can find 7 frequent words in table -1. For example, word {w4, w5} always occurs with word {w6} in a text document. It is called as shorter pattern {w4, w5}, it is also one part of longer pattern {w4, w5, w6} in all of the documents.

Hence, we hope that the shorter one {w4, w5} is a noise word or pattern. So we want to keep the larger pattern {w4, w5, w6, w7}, only.

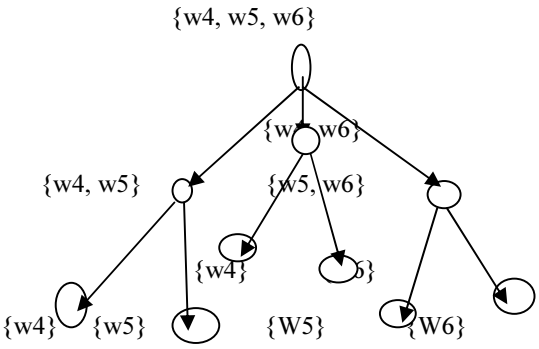

Figure 3: Closed Pattern Representation

Here, each node is represented by frequent word and their covering sets. The edges are relation between different nodes, which is denoted by sub

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pattern, for example word {w6} becomes a sub pattern of {w4, w5, w6}. The pattern taxonomy is used to improve the performance of closed pattern.

Effective pattern discovery is used to calculate the specificities of the particular pattern in textual documents, and also solving the problem of misinterpretation, low frequency problem. Which reduces the noisy patterns; it can improve accuracy of discovering word or term weights. It consists of two processes such as pattern deploying methods (PDM), inner pattern evolving (IPE).

Pattern deploying is used to discover the pattern in effective manner. That can easily eliminate the overlapping among the pattern; assign the frequency to term weighting method, which is reduce the both searching space and time complexity. It improves the performance of closed pattern as well as normalizing the pattern. After normalization which accurately evaluate the pattern weights, and also extracting the more semantic information from textual documents.

$$\text{Support}(\text{word}, D) = \sum_{k=1}^{n} \frac{|\{P|P\varepsilon CSPi, w\varepsilon p\}|}{\sum_{p\varepsilon CSP}^{n} |P|}$$

Where p is the number of words in p. table -3 explains the set of closed pattern sequential discovered in text document. For example, word *system* occurs in the five documents (d1, d2,d3,d4,d5).
Support(word,D)=1/4+1/4+1/2+2/3+1/3=2

Table 3: Closed Sequential Patterns Discovered (Csp) Where Min_Sup=3

Document	Set of CSP discovered
d1	{(office,**system**),(internet,world)}
d2	{(office,world),(internet,**system**), (world)}
d3	{**(system**), (line)}
d4	**{**(college, **system**), (**system**)}
d5	{(line, **system**, world)}
d6	{(apartment), (college)}

The representation of closed pattern is very difficult to discovered word or pattern in textual documents.

$di = \{(wi_1, mi_1),(wi_2, mi_2), \ldots..,(wi_n, mi_n)\}$

Where wi_j in pair (wi_j, mi_j) denotes a single word.

Table 4: Deploying Pattern From Text Document

Document	Set of CSP discovered
d1	{(office,1),(system,1),(internet,1),(world,1)}
d2	{(office,1),(system,1),(internet,1),(world,2)}
d3	{**(**system,1), (line,1)}
d4	**{(**college, 1), (system,2)}
d5	{(line,1), (system,1),(world,1)}
d6	{(apartment,1), (college,1)}

Table 5: Normalized Deploying Patterns

Document	Set of CSP discovered
d3	{**(**system,1/2), (line,1/2)}
d4	**{(**college, 1/3), (system,2/3)}
d5	{(line,1),(system,1),(world,1)}
d6	{(apartment,1), (college,1)}
d7	{(office,9/20),(system,1/2), (internet,1/2),(world,13/20)}

Here d1 and d3 are replaced by d7 and deploying patterns are normalized. Because table -4 although document d1 and d2 do not have the same set of sequential pattern. So we cannot find the same sequential pattern, word set (d1) = word set (d2) = {office, system, internet, world} as shown in table -5, therefore we compose the pattern with same word set into one.

$d7 = d1 \oplus d2$

d7= {(office,1/4+1/5),(system,1/4+1/4),

(internet,1/4+1/4),(world,1/4+2/5)}

d7= {(office,9/20),(system,1/2),

(internet,1/2),(world,13/20)}

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Inner pattern evolution is used to reduce the side effects of noisy patterns, low frequency problem. Here we are considering threshold value. \

These value is usually used to categorize the given text documents is relevant or irrelevant. There are two kinds of offenders such as complete conflict offender, partial complete conflict offender.

Complete conflict offender means subset of negative documents i.e irrelevant document. Partial conflict offender is a part of negative documents. In general offender means temporary storage devices. It is used to store side effects of negative documents.

For example

Sup_x (<w4, w5, w6>) = 3

Sup_x (<w2, w3>) = 4

Sup_x (<w6>) = 5 and

d = {(w2, 4), (w3, 4), (w4, 4), (w5, 3), (w6, 5)} .

3.5 Fuzzy Similarity Measure In Text

It is mainly used to extract the index word or term from particular textual documents. The similar word describe the different meaning creates ambiguity problem. It reduces dimensionality of feature vectors for text categorization. After it can be automatically form a different group.

These selects the one extracted word or feature from each and every group, and also identify the bag-of-words from text documents. It can handle the overlapping pattern, complex structure of text.

We take three types of word into considerations, a word is considered to belong to a single document group only we refers to a pattern as hard group, using the given below equation.

$$\partial ij = \begin{cases} if\ \ doc\ d\ belongs\ to\ group, \\ 0, \text{otherwise} \end{cases}$$

The mean μ and the standard deviation σ_i are defined as

$$\mu = \frac{1}{M}\sum_{i=1}^{m} xi$$

Where M is the size of the document and xi is the ith value in the documentation. We will use a documentation of 20,000 (M). It is taken from various datasets.

$$\sigma_{i=} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(xi - \mu)^2}$$

The fuzzy similarity of word pattern to categorization is defined by Gaussian membership function.

$$\partial i(x) = \prod_{i=1}^{m} . \exp\left[\frac{-(xi - \mu)^2}{\sigma i}\right]$$

The value of $\partial i(x)$ are bounded in the interval of (0, 1). $1 \leq I \leq m$.we say that x passes the similarity test on categorization

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Input: Text document, G = {n, e} a group where dx, dy ε $n, document\ d.$

Output: Text classification.

Method:

1. Arrival of document d;
2. Insert the all document in node n;
3. Document xy = 0;
4. **if** Document = similar then
5. **foreach** document =1: n class
6. Document = Document + 1/dm; where dm = $= \sum_{xy}^{n}$ di for i = 0, 1…
7. Return document;
8. **End for loop;**
9. Partition document into 'G' groups using k-NN method;
10. **if** dx, dy are in same groups, then
11. dm+1 = dx ∪ dy; // composition where dxε$du, dy\varepsilon du$
12. d(dm+1, du) = min{d(dx, du), d(dy, du)}; // for all remaining groups in du
13. normalize n;
14. dm+1 = dm+1 ∪ n;
15. **end if ;**
16. **else** partition into G groups using k-NN method;
17. Repeat step 6 into 11 ;
18. Add document dm+1, as a new group, remove existing document;
19. **endif;**
20. Calculate the similarity between two documents refer equation;

Figure 4: Fuzzy Algorithms for Pattern Mining Using Graph

In this algorithm first collects the arrival of new text documents (line 1). After all documents are inserted into node n (line 2). The similar documents are partitioning into some set of groups using k-NN method (line 4 to 9). If the documents are same group then perform the composition operation (line 10 to 11). These documents are comparing with the general document du (line 12). After this documents are sending to normalization (line 13). It is used to analysis the given set of document and also analysis another set of document relationship and their capabilities. Again (line 6 to 11) performs the same operation. All existing documents are removing from database in (line 18).

Finally (line 20) is used to calculate the similarity between two documents, to improve the efficiency of text classification performance. It reduces the searching space and also avoids the low frequency problem.

4. EVALUATIONS

In this section, datasets are used to evaluate the proposed technique. Text preprocessing consists of two types of algorithm such as stemming and stop list algorithm. Some standard measures are used for performance execution and results are compared with PTM, K-NN, FUZZY similarity based methods.

4.1 Dataset

RCV1 is one of the most popular dataset in text mining. This consists of various articles, newspaper, journals for different set of time period.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

It is used to test the proposed approach in effective manner. This is categorized into two type of department. One is training department and another one is testing department. TREC is another one of the popular dataset. It is used to track the different text document. This is divided into two set of groups such as relevant and irrelevant group. Experimental on RCV1 data collection and TREC topics implement that the proposed result achieves better performance. So that we can guarantee the efficiency of the proposed system. Concept based mining model is one the classification technique, which finds semantic relationship between the different word

Table 6: Classification Performances On The First 60 Topics.

Method	Top15	Top30	Top45	Top60
Fuzzy	**2.5678**	**2.2345**	**2.0786**	**1.8765**
K-NN	2.0012	1.8765	1.6476	1.4237
PTM	1.5678	1.3456	1.1223	0.9876
CBM	1.1234	0.9782	0.7544	0.5667

4.2 Experimental Results

In this section we describe the results of our proposed approach fuzzy estimated and similarity based self generating algorithm. The overall results of proposed method is described in table -6, and outputs are showing in figure -5 .we are collecting 120 topics from recent and most popular datasets. All approach cannot achieve the all topic at particular movement of time.

The first 60 topics are more efficient and computationally better performance for our proposed approach, and also execution of results are more accurate. Our proposed method, fuzzy estimated and similarity based self generating algorithm performs better than PTM, K-NN, and concept based model. Our experiments, all methods used 600 pattern or word for each and every topic from the textual document. It has better understanding between the large datasets.

This can reduce the noisy value, missing value, and unknown value. Concept based mining model is one the classification technique, which finds semantic relationship between the different word. Which can be exactly classified the different pattern, to ensure the model more reliable, scalable, and robust.

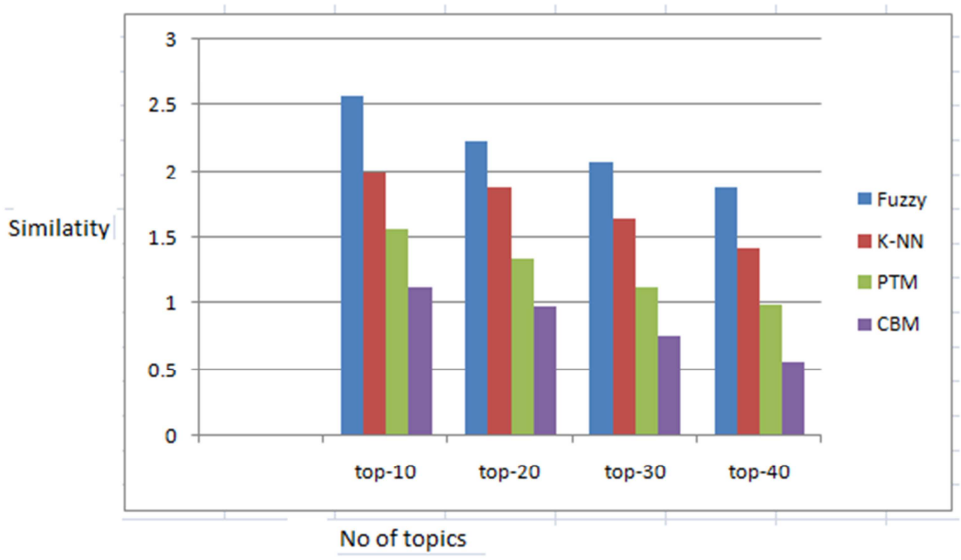

Figure 5: Comparisons of Fuzzy and other classification methods

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

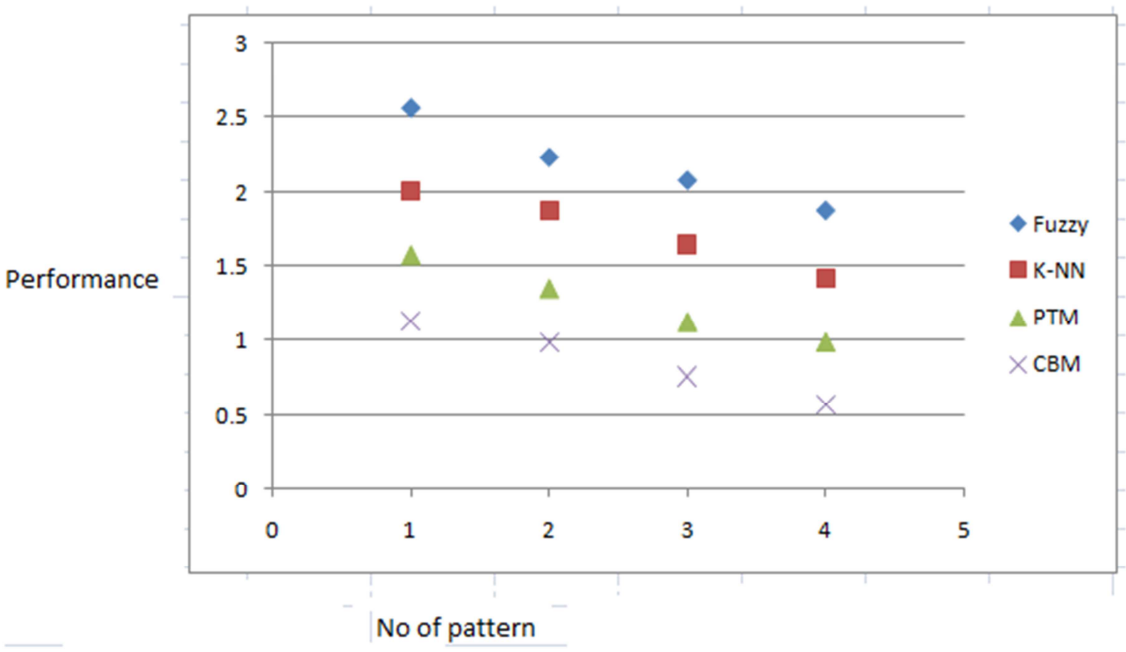

Figure 6: Performance Of Fuzzy Self Generating And Similarity Based Model

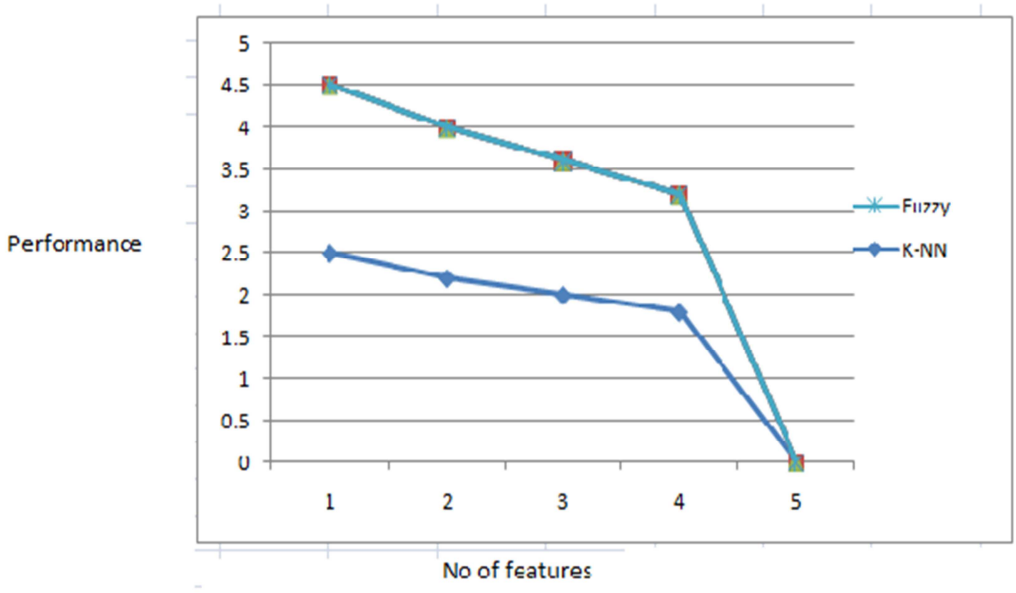

Figure 7: Comparison of Fuzzy and K –NN methods

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

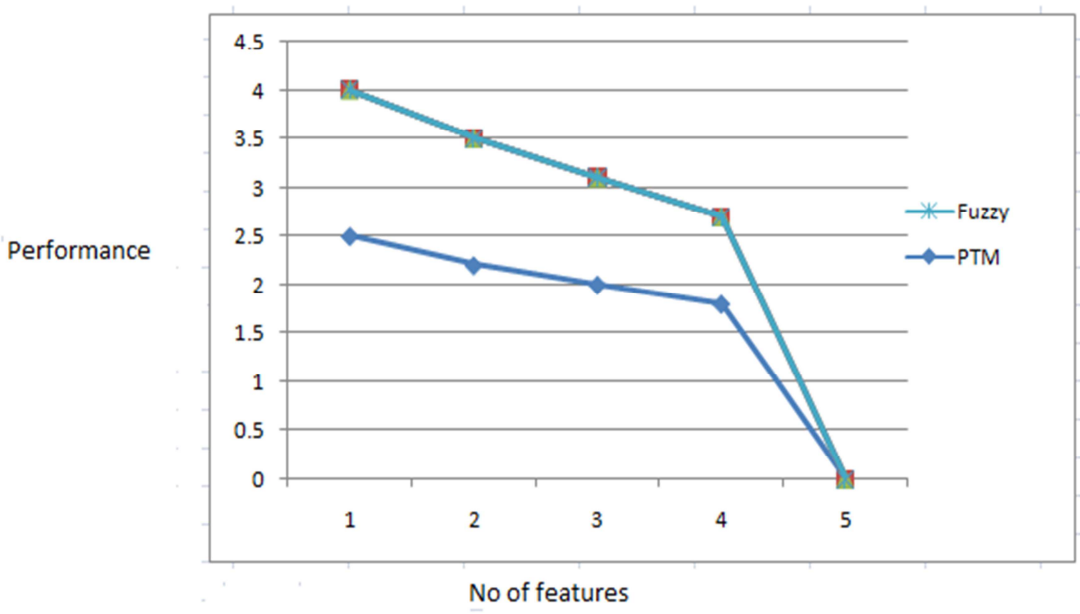

Figure 8: Comparison of Fuzzy and PTM methods

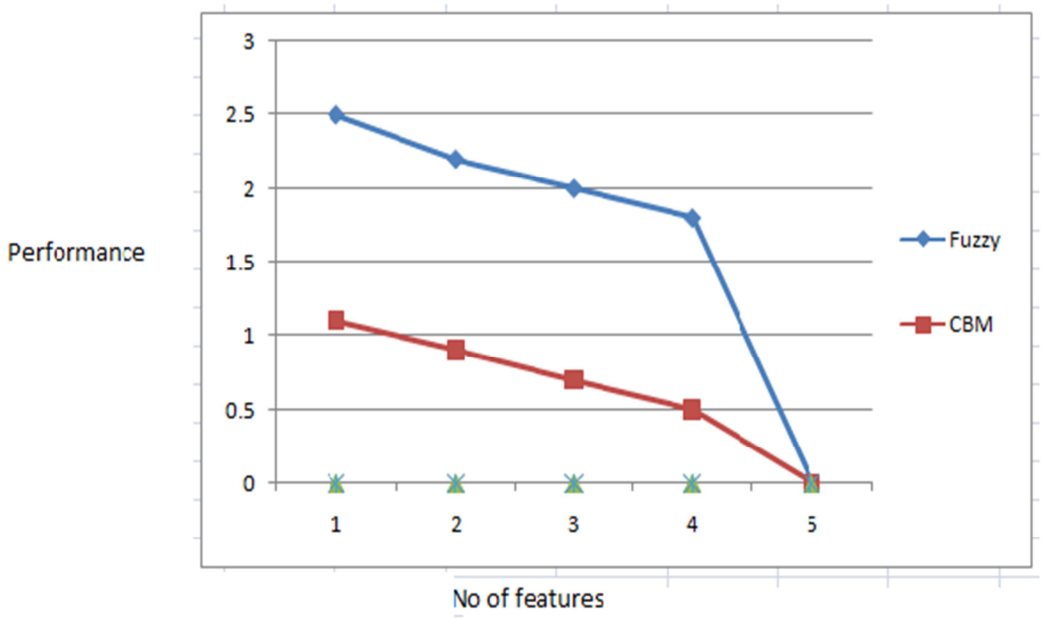

Figure 9: Comparison of Fuzzy and CBM method

4.3 Discussion

The number of patterns is used for both training and testing phase. These training methods have shown the performance in figure - 6. The average number of word or patterns is approximately calculated by each and every text document. It reduces low frequency problem in effective manner. We can achieve our target result with help of our proposed model. PTM is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

one of the efficient techniques for pattern pruning. But

performance is very less compare with fuzzy self generating and similarity based model. K-NN method is another one type of text classification

5. A CASE STUDY OF FEATURE CLASSIFICATION

In this section we describe the case study of pattern of feature classification. We form one text data set by selecting three classes such as newspaper, articles, and research papers.

Table -7 Classification results of Fuzzy on the dataset RCV1

Classes	Newspaper	articles	Research papers
D1	360	389	3
D2	8	330	0
D3	6	5	288

Table -7 shows the results of using fuzzy self generating and similarity based approach for RCV1 datasets. It is used to retrieve the original text document. This approach is used to partitioning the text document simultaneously. It forms the one group or class. After these methods are produce the consistent text classification results.

In summary, in this paper we propose a fuzzy estimated and similarity- based self generating algorithm for text classification. It is overcome the low frequency problem, also calculate the similarity between the different pattern and word in effective, accurate manner. Words are categorized into two forms of cluster groups. One is inter cluster group (similar), another one is intra cluster group (dissimilar). Experimental on RCV1 data collection and TREC topics implement that the proposed result achieves better performance.

6. CONCLUSIONS

This paper presented data mining techniques as well as different text classification approach. These techniques include pattern taxonomy model, K- nearest neighbor method, concept based model in the last decade. Pattern clustering is one of the favorable methods for feature extraction in text classification. The main aim of pattern clustering is to group the all original feature into clusters. In this paper we propose a approach. These approaches are very accurately classifying text document, achieve 80% of accuracy. But we are expecting our proposed method is reaching the target above 90%. The concept based model (CBM) has also improving the better performance.

fuzzy estimated and similarity- based self generating algorithm for text classification. It is overcome the low frequency problem, also calculate the similarity between the different text documents in effective, accurate manner. Words are categorized into two forms of cluster groups. One is inter cluster group (similar), another one is intra cluster group (dissimilar). Experimental on RCV1 data collection and TREC topics implement that the proposed result achieves better performance.

This study is still having improvements on it and especially on the researching results. In future, we can implement the fuzzy self generating and similarity based algorithm using java technology. Currently, we have performed well up to the feasibility study of the proposed system.

Future work will also include evaluating fuzzy estimated and similarity based self generating algorithm with various datasets, so that we can guarantee the efficiency of the proposed system. There is also yet more scope for future research in the field of concept-based text classification and another direction is to apply the same approach to web document classification.

7 GLOSSARY OF TERMS

Learning. Learning is used to learn the particular information. It consists of two types of learning methods such as supervised learning (known label) and unsupervised learning (unknown label).
Classification. Classification is one of the supervised learning models. It predicts the target value or attribute. This is mainly focus on extrinsic structure or relations and also well known background knowledge. Classification predicts categorical data.
Clustering. Clustering is one of the un supervised learning model. It focuses the intrinsic structure, relations, and interconnectedness of the data. It is a collection of objects that are similar to one another within the same cluster and are dissimilar cluster to the objects in other cluster.
Fuzzy. Fuzzy allows approximate values, and also opposes the incomplete or ambiguous data. It is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

language independency. Fuzzy represented by binary values such as one or zero.

Document. Document represents the input of unstructured text documents, in the form of E-mail, text documents, and so on.

REFERENCES :

[1] Ning Zhong, Yuefeng Li, and Sheng – Tang Wu, "Effective Pattern Discovery for Text Mining", *IEEE Transactions On Knowledge And Data Engineering,* Vol. 24, No. 1, January 2012.

[2] Mohamed S. Kamel, ".An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Transactions On Knowledge And Data Engineering,* Vol. 22, No. 10, October 2010.

[3] Lijun Zhang,"Locally Discriminative Coclustering", *IEEE Transactions On Knowledge And Data Engineering,* Vol. 24, No. 8, August 2012

[4] Taiping Zhang," Document Clustering in Correlation Similarity Measure Space*", IEEE Transactions On Knowledge And Data Engineering,* Vol. 24, No. 6, August 2012.

[5] Khurram Shehzad," EDISC: sA Class-Tailored Discretization Technique forRule-Based Classification*," IEEE Transactions On Knowledge And Data Engineering,* Vol. 24, No. 8, August 2012.

[6] Natthakan lam-ON, "A Link based Cluster Ensemble Approach for Categorical Data Clustering", *IEEE Trans. Data Engineering*, Vol. 24, NO. 3, MARCH 2012.

[7] Brijesh Verma,"Cluster – Oriented Ensemble Classifier: Impact Of Multicluster Characterization on Ensemble Classifier Learning", *IEEE Trans. Data Engineering*, Vol. 24, NO. 4, APRIL 2012.

[8] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining", *Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06),* pp. 1157-1161, 2006.

[9] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. "A two-stage text mining model for information filtering", *In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, pages 1023–1032, New York, NY, USA*, 2008. ACM.

[10] Y. Li, A. Algarni, and N. Zhong. "Mining positive and negative patterns for relevance feature discovery*", In KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining,* pages 753–762, New York, NY, USA, 2010. ACM.

[11] X. Ling, Q. Mei, C. Zhai, and B. Schatz. "Mining multifaceted overviews of arbitrary topics in a text collection", *In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining,* pages 497–505, New York, NY, USA, 2008. ACM.

[12] P.Mitra, C. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 3, pp. 301-312, Mar. 2002**.**