Text Categorization using Distributional Features and Semantic Equivalence

Tirupathaiah Kommi
M.Tech in Software Engineering
Aurora's Engineering College
Bhongir
Nalgonda Dist-508 116, India

Srikanth Jatla
Associate Professor and HOD of C.S.E
Aurora's Engineering College
Bhongir
Nalgonda Dist-508 116, India

ABSTRACT

In text mining domain, text categorization is widely used which is nothing but assigning predefined categories to text. The process of assigning values to words based on the occurrences of words known as bag-of-word approach was used by previous researchers in order to find how frequently a word is used in the document. This approach has a drawback as it does not consider other features of words except the count of it. This paper throws light into assigning other values to a word known as distributional features. This approach is novel and the distributional features include the position of first occurrence of word and compactness of its appearances. Our experimental results revealed that text categorization has been improved with the help of distributional features and semantic equivalence. The research has thrown light into another fact that distributional features are very useful when writing style is casual and document is long. The semantic equivalence used to extend equivalence rough set approach.

Keywords: Text mining, machine learning, text categorization, distributional feature, tfidf

1. INTRODUCTION

10 years down the line, in information system field content based document management techniques have gained popularity. This is due to drastic increase in the availability of documents in electronic format. There has been a need of accessing them in a flexible way. The domain is known as text mining. Text categorization is one of the text mining techniques. For text mining various classifiers are used. They include AdaBoost, SVM (Support Vector Machine), kNN (k Nearest Neighbor), Neural Network, Decision Tree, Naïve Bayes and ML (Machine Learning). All these techniques comes under supervised learning. All these techniques are almost based on bag-of-the-word concept where every word assigned a weight assigned on its occurrences in the document. These approaches are proved inefficient as a word can have different meaning in different context.

The novel approach used in this paper is based on the following distributional features.

- Compactness of appearances of a word.
- Position of first appearance of word

The first distributional feature is compactness of appearances of a word. This feature takes into consideration the word's appearance in various parts of the document. In each part word's weight is different. Consider document A and B. A talks about wheat and B talks about grain. However, both documents have the word wheat. The first document is more focused on wheat while second document is more focused on grain. Though word count is almost same, obviously the word wheat in the first document has more importance. Therefore simple bag-of-the- word is not sufficient.

The second distributional feature is position of first appearance of word. This feature is influenced by the fact that any author specified a word which is important in the document mentions in the early Part of the document Consider documents A and B.A talks about grain in which the word grain is in the title of document. B talks about cotton but the word grain is also repeated in B. However, in B the word but the word grain is also repeated in B. However, in B the word grain appears at the end of document as author gives least important to it.

The following contributions are made by this paper.

- with a little additional cost distributional feature are designed to help improve the process of categorization of text.
- The usage of distributional features in addition to traditional word frequency approach is described to improve performance of text categorization.
- The efficiency of distributional features is proportional to the length of documents and factors that affect performance of distributed features are discussed.

The rest of the document is organized into some sections. Section 2 provides review of literature. Section 3 focuses on extracting distributional features. Section 4 discusses the usage of distributional features in text categorization. Section 5 and 6 report results of experiments and concludes them respectively.

2. RELATED WORK

This section reviews literature on text categorization and its previous inventions. The term feature has got two meanings which are related. The first meaning is unit which represents a document while the second meaning is how to assign a weight to the feature. When bag of words is considered as an example, this feature's meaning is a single word. On the contrary tfidf is a feature that gives latter meaning. The review paper [15]

provides many other topics pertaining to text categorization. Apart from first meaning, many researches [6], [14] explored syntactic phrases as well. Language grammars are used to extract syntactic phrases. Overall, the experiments revealed that bag of words and their approach is almost same and no significant improvement.

Statistical phrases have [4], [13] could attract more attention. This concept is known as n-gram. A statistical phrase is a collection of words that occur in document in a statistically interesting way. Here number of words in sequence is represented by n. With the help of a feature selection mechanism this approach reported improved performance in text categorization. Apart from phrases, other features related to linguistic such as word- senses, hypernym, synonym and POS-tag relations of WordNet [7] were tried by researches [14]. The performance brought by linguistic features is disappointing. Another feature used was word cluster for the first meaning [1]. It focused on distribution of word on various categories. Agglomerative approach [1] and Information Bottleneck [2] are the two clustering methods used. Results of experiments revealed that word-cluster approach has improvement over single- word based approaches.

Of late, a new text representation method is proposed by Sauban and Pharynges [13] which makes use of information of word sequence. The approach they used was to calculate discriminative score for every word and then the document was shown as a curve that shows the change of accumulated scores of words. The curve was named Document Profiling. This curve was turned into two constant number of features. It could achieve improvement overbag of words with less computational cost. Two sources are used for assigning weights for second meaning. They are interdocument and intradocument. The first one uses information between the documents while the second one uses information within document. For tfidf, tf is used to represents intradocument source while idf represents interdocument source. Less number of researches were found that are based on intradocument-based weight. Researches [10], [12] used several variants of tf including inverse frequency and logarithmic frequency. Importances of each sentence calculate weight is used by Ko et al. [9].

Researches, for interdocument-based weight, tried to improve the idf. The approach used was unsupervised view and supervised view. Redundancy approach is proposed by Leopold and Kindermann [12] to measure importance of a word. This approach is able to quantify the skewness of this word's frequency distribution in various documents. In their comparative study Lan et al. [10] used relevance weight. It followed different approach than idf. It divided the documents without given word by documents with given word. Deriving idf directly is not well suited for text categorization as many researchers believed. Many supervised weights approaches came into existence in order to focus on text categorization. A measure similar to Gini Index was used by Karypis [16] in order to calculate the discriminating power of each word .Some feature scoring functions such as Gain Ratio, Information Gain and Chisquare are used by Debloe and Sebastiani [5] to modify the idf. Gain Ratio was the best find which is a variant of Information Gain. A weighting method based on statistical confidence intervals is used by Soucy and Mineau [17]. It performs feature selection implicitly. It revealed improvement over tfidf method on benchmarks. Having reviewed the prior work, our work is based on distributional features that can be considered as a new weighting method for text categorization. The two distributional features are namely compactness of appearances of word and the position of first appearance of word.

3. EXTRACTING DISTRIBUTIONAL FEATURES

3.1 How to model word's distribution?

Modeling a word's distribution in document is done two steps. In the first step, a document is divided into many portions. In the second part, the distribution of word represented by an array in which every element contains number of occurrences of the word in respective portion of document.

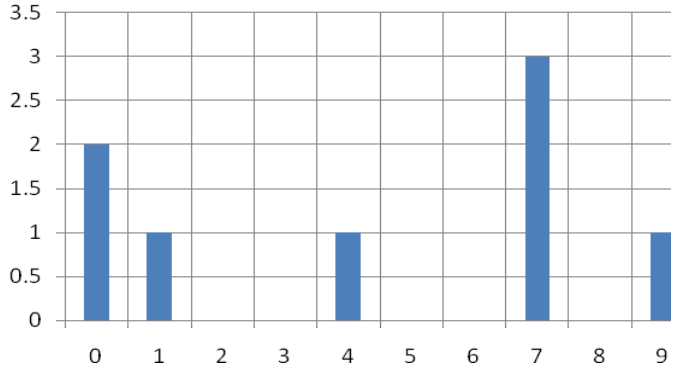

Fig 1: The Index of sentence

Partitioning a document into parts is a problem here. As per Callan [3] three types of passages exist. The merits and demerits of these are discussed by Kim and Kim [8]. The three passages and their advantages and drawbacks are shown in the table below.

Passage Name	Description	Advantages	Disadvantages
Discourse Passage	Based on logic components of document.	Intuitive	Inconsistent passage decoration is provided sometimes.
Semantic Passage	Based on the content of document.	More accurate.	Performance is affected by partition algorithm
Window Passage	Is sequence of words.	Simple to implement.	May break sentence. Length of window is hard to choose

Table 1: shows types of passages.

In this paper the discourse and window passages are explored. Fig. 1 shows the distribution of word corn in a document d with 10 sentences.

3.2 EXTRACTING DISTRIBUTIONAL FEATURES

Three implementations are done for the first distributional feature compactness of the appearances of a word‖ . They are named as
ComPact Part Num, Compact FLDist, and ComPactPosVar.

ComPactPartNum: In this approach the concept of compactness is measured depending on the number of parts where a word appears. If a word appears in many parts of document, it is said to be less compact.

ComPact FLDist: In this approach, to measure the compactness the distance between word's first and last appearances is used. If the distance is more it is less compact.

ComPactPosVar: In this approach to measure the compactness, the difference in positions of all appearances is used. With respect to language of statistics, it is natural way.

The following are the formulae used to find compactness of word in all implementations.

$$\text{FirstApp}(t, d) = \min_{i \in \{0 \ldots n-1\}} ci > 0?\ i : n; \quad (1)$$

$$\text{ComPact}_{\text{PartNum}}(t, d) = \sum^{n-1}_{i=0} Ci, > 0?\ 1:0; \quad (2)$$

$$\text{LastApp}(t, d) = \max_{i \in \{0 \ldots n-1\}} Ci > 0?\ i : -1, \quad (3)$$

$$\text{ComPact}_{\text{FLDist}}(t,d) = \text{LastApp}(t,d) - \text{FirstApp}(t,d),$$

$$\text{Count}(t, d) = \sum^{n-1}_{i=0} Ci,$$

$$\text{Centroid}(t, d) = \frac{\sum^{n-1}_{i=0} Ci \times i}{\text{Count}(t, d)}$$

$$\text{ComPact}_{\text{PosVar}}(t,d) = \frac{\sum^{n-1}_{i=0} Ci \times |i - \text{centroid}(t, d)|}{\text{Count}(t, d)}$$

The following formulae are used to apply the above to example given in Fig. 1.

$\text{FirstApp}(\text{corn},d) = \min\{0,1,10,4,10,10,7,10,9\}=0,$
$\text{ComPact}_{\text{PartNum}}(\text{corn},d) = 1+1+0+0+1+0+0+1+0+1 =5,$
$\text{LastApp}(\text{corn},d) = \max\{0,1,-1,-1,4,-1,-1,7,-1,9\}=9,$
$\text{ComPact}_{\text{FLDist}}(\text{corn},d) = 9-0 = 9,$
$\text{Count}(\text{corn}, d) = 2+1+1+3+1 = 8,$
$\text{Centroid}(\text{corn}, d) = (2\times0+1\times1+1\times4+3\times7+1\times9)/8 = 4.375.$

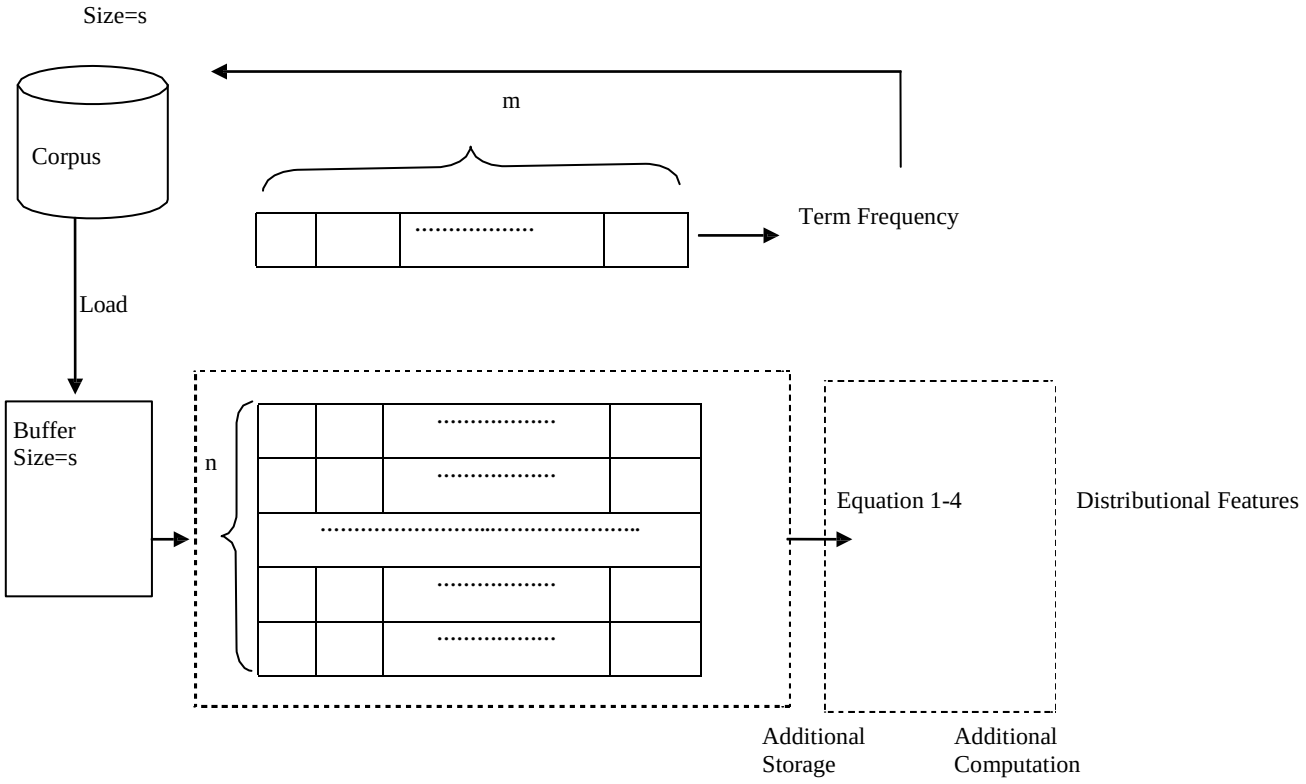

Fig. 2: shows the process of extracting the term frequency and distributional features.

4. USAGE OF DISTRIBUTIONAL FEATURES

Term frequency in tfidf is nothing but a value to measure the significance of a word in a document. As discussed in related work section term frequency is not sufficient. We also consider distributional features. For this reason the standard tfidf equation is modified as follows.

$$\text{tfidf}(t,d) = \text{Importance}(t,d) \times \text{idf}(t). \quad (5)$$

The following formulae are used to calculate TF (Term Frequency), CP(Compactness) ,and FA(First Appearance).

$$TF(t,d) = \frac{count(t,d)}{Size(d)} \quad (6)$$

$$CP_{PN}(t,d) = \frac{ComPact_{FLDist}(t,d)}{len(d)} \quad (7)$$

$$CP_{FLD}(t,d) = \frac{ComPact_{FLDist}(t,d) + 1}{len(d)} \quad (8)$$

$$CP_{PV}(t,d) = \frac{ComPact_{PosVar}(t,d) + 1}{len(d)} \quad (9)$$

$$FA(t,d) = f(FirstApp(t,d), len(d)). \quad (10)$$

4.1 Semantic Rough Set Approach

The idea is to approximate the concept by two descriptive sets called lower and upper approximations. The main philosophy of rough set approach to concept approximation problem is based on minimizing the difference between upper and lower approximations. This leads to many efficient applications of rough sets in machine learning, data mining and also in granular computing. Many clustering methods based on rough sets and other computational intelligence techniques were proposed including support vector machine (SVM), genetic algorithm (GA), modified self-organizing map (SOM). The rough set based clustering methods were applied to many real life applications, e.g., medicine, web user clustering and marketing.

Two most popular approaches to facilitate searching for information on the web are represented by web search engine and web directories. Web search engines allow user to formulate a query, to which it responds using its index to return set of references to relevant web documents (web pages). Web directories are human- made collection of references to web documents organized as hierarchical structure of categories. One approach to manage the large number of results is clustering. The concept arises from document clustering in Information Retrieval domain: find a grouping for a set of documents so that documents belonging to the same cluster are similar and documents belonging to different cluster is dissimilar. Search results clustering thus can be defined as a process of automatically grouping search results into to thematic groups. However, in contrast to traditional document clustering, clustering of search results are done on-the-fly and locally on a limited set of results return from the search engine. Clustering of search results can help user navigate through large set of documents more efficiently. By providing concise, accurate description of clusters, it lets user localizes interesting document faster.

In document clustering, the main emphasis is put on the quality of clusters and the scalability to large number of documents, as it is usually used to process the whole document collection (e.g. for document retrieval on clustered collection). For web search results clustering, apart from delivering good quality clusters, it is also required to produce meaningful, concise description for cluster. Additionally, the algorithm must be extremely fast to process results on-line (as post-processing search results before delivered to the user) and scalability with the increase of user requests.

5. ANALYZING EXPERIMENTS AND RESULTS

kNN and SVM achieved best performance as per previous study [18]. For this reason all experiments specified in this section are based on them.

		kNN			SVM	
Gain(%)	Reuters	Newgroup	WebKB	Reuters	Newgroup	WebKB
	miF1 maF1	miF1 maF1	miF1 maF1	miF1 maF1	miF1 maF1	miF1 maF1
TF	0.822 0.550	0.859 0.859	0.788 0.729	0.883 0.554	0.901 0.899	0.901 0.892
CPpv	-2.7	3.1** 3.0**	6.4** 10.4**	-0.4 -2.8	0.8** 0.8**	2.6** 2.9**
CP(best)	0.8 3.1*	3.9** 3.9**	6.4** 10.4**	0.2 1.7	1.1** 1.1**	2.9** 3.3**
FA_{GI}	-2.5++ -4.7++	4.1** 4.1**	7.0** 12.4**	-0.1 -1.7	2.4** 2.4**	3.8** 4.4**
FA(best)	0.0 -0.5	5.6** 5.5**	7.7** 12.9**	-0.1 -1.7	2.4** 2.4**	4.4** 4.9**
$TF+CP_{PV}$	0.7 3.0	3.0** 3.0**	4.0** 6.6**	0.4* -1.7	0.9** 0.9**	2.4** 2.7**
TF+CP(best)	0.9 3.0	3.8** 3.8**	5.1** 8.4**	0.4* 1.9	1.0** 1.0**	2.5** 2.7**
$TF+FA_{GI}$	0.1 0.9	5.0** 4.9**	5.3** 8.9**	0.3 -1.9	2.1** 2.1**	3.2** 3.8**
TF+FA(best)	1.6** 3.2	5.8** 5.8**	7.0** 11.1**	0.4 -1.9	2.3** 2.3**	3.9** 4.4**
$CP_{PV} + FA_{GI}$	-0.1 -0.2	5.5** 5.5**	6.9** 11.6**	0.2 -1.2	2.3** 2.3**	3.8** 4.4**
CP+FA(best)	1.6** 3.2	5.8** 5.8**	7.6** 12.4**	0.5 1.7	2.4** 2.4**	4.1** 4.7**
$TF+CP_{PV}+FA_{GI}$	0.8 1.9	5.4** 5.3**	6.1** 10.1**	0.5** -2.1	1.8** 1.9**	3.3** 3.8**
TF+CP+FA(best)	1.7** 5.3**	5.9** 5.9**	7.0** 11.1**	0.6** 1.8*	2.3** 2.3**	3.7** 4.3**

Table2.Simplified Results of the Distributional Features of Three Data Sets(Discourse Passage)

5.1 Data Sets

Two samples of data sets are used in experiments. The first collection of documents are taken from the Reuters [18] which contains 21,578 articles, the second collection documents Contains (from 20 Newsgroup corpus [11]) 19,997 articles.

5.2Performanc Measure and Experimental Configuration

To estimate the presentation on these three corpora, the usual accuracy, remind and F1 evaluate is used. The precision (p_i),

recall, F1 measure ($F1_i$), the contingency table of category Ci are computed as follows:

$$P_i = \frac{TP_i}{TP_i + FP_i} \qquad r_i = \frac{TP_i}{TP_i + FN_i} \qquad F1_i = \frac{2 \times p_i \times r_i}{(p_i + r_i)}$$

These evaluates can be combined all kinds in two ways. One is to average each type of precision, recall, and F1 to get the total accuracy, recall and F1. This process is called macroaveraging.Other is found on the global Contingency table (Table 3), which is called microaveraging. Macroaveraging is further involved by the classifier presentation on unusual kind, as microaveraging is further exaggerated by the presentation on familiar kinds.

Table 3 The Global Contingency Table

Category Set C= C1,C2,…..,$C_{\|C\|}$		Expert Judgement	
Classifier Judgement		Yes	No
	Yes	$TP=\sum_{i=1}^{\|C\|} TP_i$	$FP = \sum_{i=1}^{\|C\|} FP_i$
	No	$FN=\sum_{i=1}^{\|C\|} FN_i$	$TN = \sum_{i=1}^{\|C\|} TN_i$

WebKB, as it is unilabel data se Reuters-21578 and 20 Newsgroup, the section is utilized as the conversation passage. For WebKB, because it is web page corpus, it is hard to get the conversation passage directly. Now, a webpage segmentation algorithm called VIPS is utilize to separate web pages into various blocks, various examiners describe precision on this data set, which is same as miF1.The communication passage and windows passages with various sizes are make use of further distributional characteristics.

6. CONCLUSION

The prior work on text categorization basically depended on the frequency of appearance of a word in the document to find its importance and category. The consideration of were count of word in the document is not sufficient as a word might be used differently in different document and the theme of that might be different. This paper proposed two distributional features. They are compactness of appearances of word in various parts of document and the first appearance of the word in the document. These two distributional features coupled with tfidf-style equation resulted in improved performance in text categorization. Another fact known is distributional features are more effective if the input documents for text categorization are especially long. The main philosophy of rough set approach to concept approximation problem is based on minimizing the difference between upper and lower approximations. Testing this approach with a blog data set is an interesting area for future improvement of it.

7. REFERENCES

[1] L.D.Bakerand A.K.McCallum, Distributional Clustering of Words for Text Classification, Proc. ACM SIGIR ’98, pp. 96-103, 1998.

[2] R. Bekkerman, R El-Yaniv, N. Tishb, and Y.Winter Distributional Word Clusters versus Words for Text Categorization, J. Machine Learning Research, vol. 3, pp. 1182-1208, 03.

[3] J.P. Callan, Passage Retrieval Evidence in Document Retrieval, Proc. ACM SIGIR ’94, pp. 302-310, 1994.

[4] M.F. Caropreso, S. Matwin, and F.Sebastiani, A Learner-Independent Evaluation of the Usefulness of Statistica Phrases for Automated Text Categorization,Text Databases and Document Management Theory and Practice, A.G. Chin, ed., pp. 78-102, Idea Group Publishing, 2001.

[5] F.Debole and F.Sebastiani, Supervised Term Weighting for Automated Text Categorization, Proc. 18th ACM Symp. Applied Computing (SAC ’03), pp. 784-788, 2003.

[6] S.T. Dumais, J.C. Platt, D. Heckerman, and M. Sahami, Inductive Learning Algorithms and Representations for Text Categorization, Proc. Seventh Int’l Conf. Information and Knowledge Management (CIKM ’98), pp. 148-155, 1998.

[7] C. Fellbaum, WordNet: An Electronic Lexical Database. MIT Press, 1998.

[8] J. Kim and M.H. Kim, An Evaluation of Passage-Based Text Categorization, J. Intelligent Information Systems, vol. 23, no. 1, pp. 47-65, 2004.

[9] Y. Ko, J.Park and J.Seo, Improving Text Categorization Using the Importance of Sentences, Information Processing and Management, vol. 40, no. 1, pp. 65-79, 2004

[10] M. Lan, S.Y.Sung, H.B. Low, and C.L. Tan, A Comparative Study on Term Weighting Schemes for Text Categorization, Proc. Int’l Joint Conf. Neural Networks (IJCNN ’05), pp.546-551, 2005.

[11] K. Lang, Newsweeder: Learning to Filter Netnews Proc. 12th Int’l Conf. Machine Learning (ICML ’95), pp. 331-339, 1995.

[12] E. Leopold and J. Kingermann, Text Categorization with Support Vector Machines: How to Represent Text in Input Space? Machine Learning, vol. 46, nos. 1-3, pp. 423-444, 2002.

[13] R.E. Schapire and Y.Singer, Boostexter: A Boosting-Based System for Text Categorization, Machine Learning, vol. 39, nos. 2/3, pp.135-168, 2000.

[14] F.Sebastiani, Machine Learning in Automated Text categorization, ACM Computing Surveys, vol. 34, no 1, pp. 1-47, 2002

[15] S. Shankar and G.Karypis, A Feature Weight Adjustment Algorithm for Document Classification,Proc. SIGKDD’00 Workshop Text Mining, 2000.

[16] P. Soucy and G.W. Mineau, Beyond tfidf Weighting for Text Categorization in the Vector Space Model, Proc.19th Int'l J Artificial Intelligence (IJCAI '05), pp.1130-1135,2005

[17] X.-B. Xue and Z.-H. Zhou, Distributional Features for Text Categorization, Proc.17th European Conf. Machine Learning (ICML '06), pp. 497-508, 2006.

[18] Y. Yang and J.O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 412-420, 1997.

8. AUTHORS PROFILE

Mr.Tirupathaiah Kommi received his B.Tech in information technology from JNTU Hyderabad. Currently, he is an M.Tech student specializing in software engineering, in the department of computer science engineering, from Aurora's Engineering College, Bhongir, Andhra Pradesh, India. His areas of interests are data mining and knowledge discovery, especially text categorization, software engineering, testing and mobile computing.

Mr. Srikanth Jatla, working as an associate professor, head of the department of computer Science & engineering at Aurora's Engineering College with a teaching experience of 12 years. He is a B.E and M.Tech in computer science and is pursuing his PhD in Data Stream Mining at JNTU, Hyderabad. His area of interest includes data structures, principles of programming languages, algorithm analysis and compiler design.