A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model

V.M.Navaneethakumar [1], Dr.C.Chandrasekar [2]

[1] Assistant Professor, Department of Computer Applications, K.S.R College of Engineering, Tiruchengode,Tamilnadu,India.

[2]Associate Professor, Department of Computer Science Periyar University, Salem,India.

Abstract

Text mining is a growing innovative field that endeavors to collect significant information from natural language processing term. It might be insecurely distinguished as the course of examining texts to extract information that is practical for particular purposes. In this case, the mining model can detain provisions that identify the concepts of the sentence or document, which tends to detect the subject of the document. In an existing work, the concept-based mining model is used only for normal text documents clustering and clustered the text parts of the documents and efficiently discover noteworthy identical concepts among documents, according to the semantics of the sentences. But the downside of the work is that the existing work cannot be linked to web documents clustering and the text classification for the documents is an unreliable one. To make the text clustering more consistent, in our work, we plan to present a Conceptual Rule Mining On Text clusters to evaluate the more related and influential sentences contributing to the document topic. In this paper, the conceptual text clustering extends to web documents, containing various markup language formats associated with the documents (term extraction mode). Based on the markup languages like presentations, procedural and descriptive markup, the web document's text clustering is done efficiently using the concept-based mining model. Experiments are conducted with the web documents extracted from the research repositories to evaluate the efficiency of the proposed consistent web document's text clustering using conceptual rule mining with an existing An Efficient Concept-Based Mining Model for Enhancing Text Clustering.

Keywords: *Concept-based mining model, sentence-based, web document-based, concept-based similarity, text clustering.*

1. Introduction

Text mining mainly depends on geometric examination of a phrase, word or term. Statistical analysis of the word, phrase or a term is done simply with significant outcomes of the word present within the document only. Nevertheless two terms might have equal rating in the appropriate documents. One term gives more to the significance of its sentences than the other term.

The semantics of the text plays a vital role in text mining models. Provisions connected to concepts of sentence and recognize the subject of the document as well. There is a need for notion based text, to examine the terms of the sentence, document, and corpus levels. Concept-based mining model, may discriminate terms with semantic variation and has minimum authority on the sentence implication.

Generally, in text mining techniques, the phrase frequency of a term is calculated to discover the consequence of the term in the document. Nevertheless, two terms could have similar frequency in the given documents, but one term gives more to the denotation of its sentences, compared to the other term. Clustering, one of the conventional data mining strategies is an unsubstantiated knowledge pattern. Here clustering methods endeavor to recognize intrinsic alignments of the text documents, so that a set of clusters is formed in which clusters display high intra-cluster likeness and low inter-cluster likeness. Normally, text document clustering endeavors to separate out the documents into groups where every group characterizes some subject that is different from the topics characterized by the other groups.

Web Documents are a collected work of associated individual pages, each with elective spooling and non-scrolling areas, manuscript, and possibly images and other media entrenched. Instances of Web Documents would be normal windows help files, WWW and so on. A web document consist of several markup languages. A markup language is a contemporary scheme for explaining a document in a way that is entirely different from the text.

A concept-based resemblance evaluation uses the notion examination of the sentence or document. This resemblance evaluation outperforms other resemblance actions that are stranded on analysis models of the document. The correspondence between documents is supported on a mixture of sentence-based and document-based concept analysis. Similarity based on identification of ideas among document pairs, is revealed to comprise a more important outcome on the eminence of clustering due to insensitivity to the similarity’s leads to an erroneous similarity. The concepts are less receptive to noise when it comes to computing document similarity. Normally, concepts are initially mined by the semantic role labeler and examined with the sentence document.

In this paper, a new concept-based mining model is proposed in the web document's text clustering. The proposed model detains the semantic arrangement of each term inside a sentence and web document fairly than the frequency of the term inside a document only based on the various markup languages present.

2. Literature Review

Natural Language Processing (NLP) is a contemporary computational knowledge and a technique of examining and estimating the states the about words in it. NLP is a word that connects the rear into the account of Artificial Intelligence (AI), the universal learning about the cognitive purpose by evaluation procedures, with an accent on the function of knowledge illustrations. Text mining [10] endeavors to determine the novel, and formerly indefinite data by using methods from data mining. Techniques for text clustering comprise decision trees, conceptual clustering [1], clustering based on data summarization [2] and so on. Furthermore, these aspects also precisely involve the consequence of the clustering algorithm considerably [3].

With the support of the aforesaid information, the organization of the verdicts can be constructed by a new concept development [4] method is proposed to discover the relations amongst these entities. A probabilistic strategy [5] disperses significant weights to separate features that are measured as random variables, presented by restricted separate out mixtures. To examine the recognition of high-level ideas in multimedia, contented during an incorporated strategy of the diagram, glossary study and visual context, [6] produced algorithms for elevated level semantic construction [9].

A geometric study in clustering has approximately focused generally on data sets [8] by introducing a representative hierarchical clustering model [7] that utilizes probabilistic illustrations for semantic construction. Text classification and Feature clustering [12] is abiding to be one of the mainly explored NLP problems due to the growing quantity of digital libraries and electronic documents. A novel text categorization method [11] integrates the distributional classification of terms and a knowledge sensing technique.

To improve the text classification for web documents, in this work, is proposed concept based mining model for text classification, based on different formats of markup languages. The text classification in web documents is briefly explained under section 3.

3. A Consistent Web Document’s Text Clustering Using Concept Based Mining Model

The proposed text clustering using the concept based mining model is efficiently designed for web documents. Normally, web documents consist of several markup languages formats, associated with the document term extraction. The proposed model used concept-based mining for the semantic structure of each term within a sentence and web document, based on the markup languages used. In the proposed web documents based text clustering using the concept based mining model, the concepts are analyzed in terms of sentence based, web documents based on markup languages.

Each sentence in the document is marked by a semantic task that establishes the terms which give to the sentence semantics, related to their semantic functions in a sentence. Each term which attains a specific function in the sentence, is termed as a concept. Concepts can be defined as words or phrases and are entirely based on the semantic formation of the sentence. When a novel document is initiated into the system, the proposed mining model can perceive a concept match from the web document to all the formerly practiced web documents in the data set by examining the novel web document and mining the matching concepts.

3.1 Concept Based Mining Model

For the proposed concept-based mining model for web document's text clustering, a raw text document is given as the input. Each document has definite sentence restrictions. Each sentence in the document is marked repeatedly and might have one or more marked verb argument formation. The amount of labeled information is totally reliant on the information present in the sentence. The sentence contained many marked verb argument formation comprises many verbs connected with their arguments. The labeled verb argument structures are examined by the concept-based mining model on sentence and web document levels. The process of concept based mining model is shown in fig 3.1.

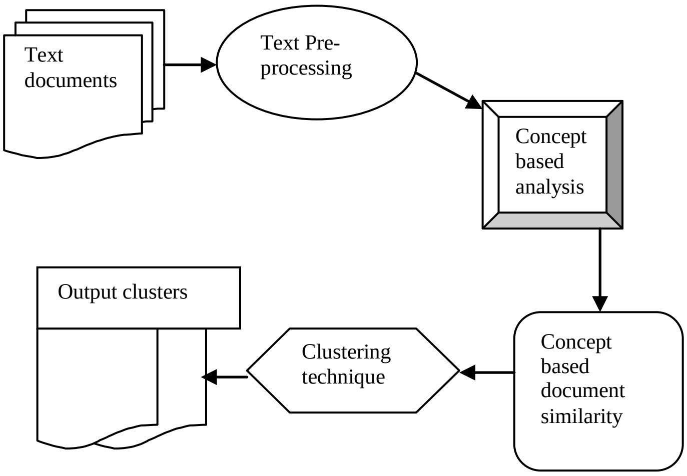

Fig 3.1 Concept Based Mining Model Process

The purpose of following the concept-based analysis task is to realize a precise examination of concepts on the sentence, in the document relatively, than a single-term study on the document only.

3.2 Sentence-Based Concept Analysis

To examine every concept at the sentence level, a novel concept-based frequency assess, called the conceptual term frequency ctf is computed. The ctf is the number of concept c happened in verb argument structures of sentence S. The concept c, which normally emerges in diverse verb argument structures of the similar sentence S, has the prime job of contributing to the significance of S.

3.2.1 Calculating ctf of Concept c in Web Document d

A concept c has many possible ctf values in dissimilar sentences, in the similar document d. Thus, the ctf value [1] of concept c in web document d is considered by:

$$ctf = \frac{\sum_{n=1}^{sn} ctf}{sn} \text{ .... (Eqn 1)}$$

where sn is the entire number of sentences containing concept c in document d. By evaluating the common ctf values of c in its d, measures the significance of concept c to the denotation of sentences in document d. For most of the sentences in the document, ctf values have a main role to the significance of its sentences that guides to determine the subject of the document.

The table1 given below describes the parametric description used in the proposed work.

Table 1 Description of Constraints

Parameter	*Description*
ctf	Conceptual term Frequency
tf	Term frequency
df	Document frequency
wd	Web document
E	Empty list
S	Sentence
C	Concepts
Sim ()	Similarity between the docs/sentences
$tfweight_i$	weight of concept i in document wd at doc. level
$ctfweight_i$	The weight of the concept i in document wd at the sentence level
cn	Sum of the concepts

3.3 Web Document Based Text Clustering using the Concept Based Mining Model

Web document based clustering is done through a concept based mining model and identify the similarity measure of the document by analyzing each concept at the document level, based on the type of markup language formats and the number of occurrences of document are also being identified and discriminated. The analysis of web document text clustering is done by the proposed web document based text clustering algorithm.

```
Step 1:  $wd_{doci}$ is a new web Document
Step 2:  E is an empty List (E is a matched concept list)
Step 3: Web documents consist of various markup language formats
        Presentational,
        Procedural and
        Descriptive markup
Step 4: $S_{doci}$ is a new sentence in $wd_{doci}$
Step 5: Build concepts list $C_{doci}$ from $S_{doci}$
Step 6: For each concept $c_i \in C_i$ do
Step7:Evaluate $ctf_i$ of $c_i$ in $wd_{doci}$  // **Conceptual term frequency**
Step 8: Evaluate $tf_i$ of $c_i$ in $wd_{doci}$ // **Term frequency** tf – no. of occurrences of given terms in a web document wd
Step 9: Evaluate $df_i$ of $c_i$ in $wd_{doci}$  // **Document frequency** df – no. of web docs. Contains concept c
Step 10: $d_n$ is seen document, where n = (0; 1; . . . ; doci-1)
Step 11: $S_n$ is a sentence in $d_n$
Step 12: Build concepts list $C_n$ from $S_n$
Step 13: For each concept $c_j \in C_n$ do
Step 14: if ($c_i$ == $c_j$) then
Step 15: Update $df_i$ of $c_i$
Step 16:Compute ctf weight = avg($ctf_i$ , $ctf_j$)
Step 17:Add new concept matches to L
Step 18:End if
Step 19: End for
Step 20: End for
Step 21: If wd has presentational markup
Step 21.1: binary codes are used in the text
Step 21.2: Instead of searching the concepts in wd, search the binary codes in the wd
Step 21.3: Else if wd contains procedural markup
Continue with the steps from 6 to 20
Step 21.4: else wd has Descriptive markup
Continue with the steps from 6 to 20
Step 22: Output the matched concepts list E
```

The web document concept-based study algorithm illustrates the practice of evaluating the ctf, tf , and df of the matched concepts in the web documents based on the markup languages. The above pseudo code initiates the process with a new document has well-defined sentence boundaries. The matched concepts and the terms identified through the above pseudo code are stored for the concept-based similarity calculations in Section 3.4.

Each concept (line 6) in the given web document, which characterizes the semantic structures of the sentence, is practiced consecutively. To match the concepts with the prior web documents is done by maintaining a concept list E, which embraces the way for each of the prior documents that contributes to a concept with the recent document. After the processing of a web document, E comprises of all the matching concepts between the present document and any prior document that divides at least one concept with the novel document. After all this process, E is termed as the output with the list of documents contains matching concepts and the essential information regarding them.

3.4 Concept Based Similarity Measure for Web Document Text Clustering

A concept-based similarity measure, supported on matching concepts at the sentence, processed on three vital features. The first process is to examine the labeled terms of the concepts that detain the semantic formation of each sentence in the web document. The second process is to measure the frequency, as well as to the key focus of the document. The final process is to compute the similarity between the concepts by considering the number of documents that contains the analyzed concepts.

The concept-based matching based on two types, one is a precise match and another one is an incomplete match between two concepts occurred in the given document.

Precise match means that both concepts contain similar words. Incomplete match means that one concept contains all the words that show in the further concept. The concept-based similarity measure [1] between two web documents, wd1 and wd2 is calculated by:

$$Sim_c(wd1, wd2) = \sum_{i=1}^{m} \max(\frac{l_{i1}}{L_{vi}}, \frac{l_{i2}}{L_{vi2}}) \times weight_{i1} \times weight_{i2} \quad \text{.......... (Eqn 2)}$$

$$weight_i = (tfweight_i + ctfweight_i) \times \log(\frac{N}{df_i}) \quad \text{............ (Eqn 3)}$$

Where tfweight$_i$ - value identifies the weight of concept i in document wd at the doc. Level

$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tfij)^2}} \quad \text{........ (Eqn 4)}$$

ctfweight$_i$ - value identifies the weight of the concept i in document wd at the sentence level

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn} (ctf_{ij})^2}} \quad \text{....... (Eqn 5)}$$

$\log(\frac{N}{df_i})$ - weight of the concept i, when i shows in a small amount of documents.

cn is the sum of the concepts which has tf value in web document wd.

tfweight$_{i\ +}$ ctfweight$_i$ - precise evaluation of the role of each concept to the topics stated in a web document.

The above features are considered for identifying the similarity between the concepts occurs in the sentence level and web document level based on the markup languages by the proposed concept-based similarity measure which measures the importance of each concept by the ctf, tf and the df measure. The concept-based evaluation develops the information mined from the concept-based analysis algorithm for web documents text clustering to evaluate the similarity among the documents.

4. Experimental Evaluation

To check the efficiency of concept matching in formatting an exact determination of the comparison between web documents, widespread of experiments are conducted and web documents extracted from the research repositories using the concept-based term analysis of web document based text clustering. The proposed web documents based text clustering is efficiently done by concept based mining model. The similarity measure of the sentence and the terms are identified in a sentence and document level. The experimental evaluation tests aimed at comparing the existing efficient concept based mining model for enhancing text clustering with the proposed web document based text clustering using the concept based mining model. At first, it analyzes the web document formats whether procedural, presentational or descriptive markup. After analyzing the formats, the concepts and sentence based similarity are identified and clustering is done for the texts or terms which are mined from the web document level. The performance of the proposed web documents based text clustering using the concept based mining model is measured in terms of

i) Sentence similarity ratio
ii) Clustering efficiency
iii) Sentence Contributory rate

Sentence Similarity: The similarities between words in dissimilar sentences have enormous control on the similarity among two sentences. Words and their sequences in the sentences are two vital aspects to estimate sentence similarity.

Clustering efficiency: The efficiency of clustering [1] is done by using two measures, F-measure and entropy. The F-measure integrates the Precision P and Recall R. Measures of a cluster j with respect to a class i are termed as,

$$P = \frac{M_{ij}}{M_j} \quad \text{........(Eqn 6)}$$

$$R = \frac{M_{ij}}{M_i}. \quad \text{.........(Eqn7)}$$

where M$_{ij}$ is the no. of members of class i in cluster j, M$_j$ is the no. of members of cluster j, and M$_i$ is the no. of members of class i. The F-measure of a class i is defined as:

$$F(i) = \frac{2PR}{P+R} \quad \text{...... (Eqn 8)}$$

By evaluating class i, the cluster with the main F-measure is measured to be the cluster that draws to class i, and that F-measure turns into the score for class i. The entropy presents an assessment of class for clusters at single level of a hierarchical clustering. Entropy deals with how consistent a cluster is. The higher the consistency of a cluster, the lower the entropy is, and vice versa. The entropy of each cluster is evaluated as

$$E_C = \sum_{j=1}^{n} (\frac{M_j}{M} \times E_j) \quad \text{....... (Eqn 9)}$$

Where M$_j$ is the amount of cluster j, and M is the sum of data objects.

Sentence Contributory rate: The relative sentences identified between the web documents which contains the matched concepts in the web document.

5. Results and Discussion

Compared to an existing document clustering process, in this work, It is seen that hoe web documents are clustered in text using the concept based mining model. It described the process by clustering the concept of the documents in the web document level. The performance of the proposed web document based text clustering is done with the concept based mining model with an enough data set. The table and graph given below show described the performance of the proposed web document based text clustering using the concept based mining model.

Table 5.1 No. of Sentences vs. Sentence Similarity Ratio

No. of Sentences	*Sentence Similarity Ratio*	
	Proposed Web Document Text Clustering	*Existing doc. Clustering*
10	12	8
20	15	11
30	20	14
40	28	17
50	36	21

The table 5.1 described the process of identifying the sentence similarity ratio between the documents. The outcome of the proposed web documents based text clustering using the concept based mining model is compared with an existing document clustering through mining model.

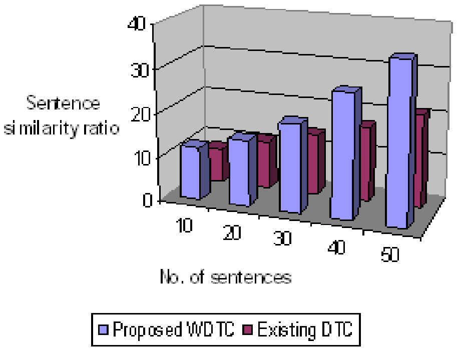

Fig 5.1 No. of Sentences vs. Sentence Similarity Ratio

Fig 5.1 describes the sentence similarity ratio for the number of sentences given in the web document. The various numbers of data sets are used in the experimentation to validate the proposed web documents based text clustering (WDTC) using the concept based mining model. The result of the proposed WDTC is compared with an existing DTC based on concept based mining model, measured in terms of sentence similarity ratio. The proposed concept based mining model used web documents clustering, based on the markup language format used. When the number of sentences in the given web document increases, the similarity of sentences in the given web document is high in the proposed WDTC, contrast to an existing DTC. The performance graph of the proposed WDTC in sentence similarity ratio is shown in the fig 5.1.

Table 5.2 Constraints vs. Clustering Efficiency

Constraints	*Clustering Efficiency*	
	Proposed WDTC	*Existing DTC*
F-measure	1	0.75
Entropy	0.30	0.89

The table 5.2 depicts the process of efficiency of clusters between the documents. The outcome of the proposed web documents based text clustering using the concept based mining model is compared with an existing document clustering through mining model.

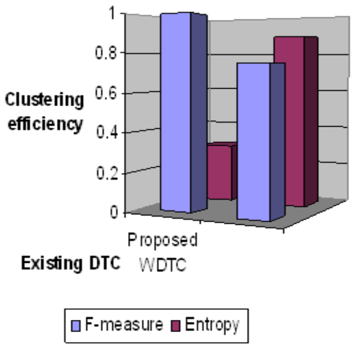

Fig 5.2 Constraints vs. Clustering Efficiency

Fig 5.2 depicts the clustering efficiency for the number of sentences given in the web document. The various numbers of data sets taken out from web research repositories are used in the experimentation to estimate the proposed web documents based text clustering (WDTC), using the concept based mining model. The result of the proposed WDTC is compared with an existing DTC based on concept based mining model, measured in terms of clustering efficiency. Lower the entropy measure and higher the F-measure improves the clustering efficiency. The proposed concept-based mining model used web documents clustering, based on the markup language format. When the number of sentences in the given web document increases, the similarity of sentences in the given web document is high, in the proposed WDTC, The contrast to an existing DTC. The performance graph of the proposed WDTC in clustering efficiency is shown in the fig 5.2.

Table 5.3 No. of Sentences vs. Sentence Contributory Rate

No. of Sentences	*Sentence Contributory Rate*	
	Proposed WDTC	*Existing DTC*
10	14	7
20	20	12
30	32	18
40	41	24
50	54	30

The above table (table 5.3) described the process of sentence contributory rate after identifying the sentence similarity between the documents. The outcome of the proposed web documents based text clustering using the concept based mining model is compared to an existing document clustering through mining model.

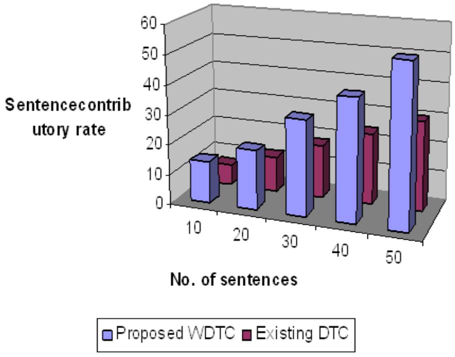

Fig 5.3 No. of Sentences vs. Sentence Contributory Rate

Fig 5.3 described the sentence contribution rate, for the number of sentences given in the web document. The various numbers of data sets taken out from web research repositories are used in the experimentation to estimate the proposed web documents based text clustering (WDTC), using the concept based mining model. Result of the proposed WDTC is compared with an existing DTC based on concept based mining model, measured in terms of clustering efficiency. The text clustering is efficient in the proposed WDTC so, the sentence contributes to the given concept level. The proposed concept based mining model used web documents clustering, based on the markup language format. When the number of sentences in the given web document increases, the similarity of sentences in the given web document is high in the proposed WDTC is contrast to an existing DTC. The performance graph of the proposed WDTC in clustering efficiency is shown in the fig 5.3.

By integrating the features moving the weights of concepts on the sentence, a document, and a concept-based similarity determination enables the precise computation of pairwise documents, is created. This allows creating concept similarity computation, between web documents in a strong and precise way. The eminence of text clustering attained by this model, considerably exceeds the conventional particular term- based approaches.

6. Conclusion

The proposed work on web documents based text clustering using concept based mining model acts as the opening between natural language processes and text mining restraints. A new concept based mining model, collected with the appropriate components, is planned to estimate the text clustering eminence. By utilizing the semantic structure of the sentences in documents, a better text clustering result is achieved. The first component is the sentence-based concept analysis, which analyzes the semantic construction of each sentence to detain the web document sentence concepts using the proposed web documents based text clustering based on concept-based mining model. Then, the web document-based concept analysis, examines each concept at the document level based on the concept-based term frequency tf and with the document frequency df universal evaluation. The concept-based similarity assessment allowed the measuring of the significance of each concept in esteem to the semantics of the sentence, the subject of the document, and the unfairness among documents obtained. The experimental results showed that the proposed web documents based text clustering, using concept based mining, model outperforms in terms of clustering efficiency, and a sentence similarity, contrast to an existing document clustering, through mining model.

References

[1] Shady Shehata, Fakhri Karray, et. Al., “An Efficient Concept-Based Mining Model for Enhancing Text Clustering”, IEEE Transactions on Knowledge and Data Engineering, VOL. 22, NO. 10, October 2010.

[2] H. Jin, M.-L. Wong, and K.S. Leung, “Scalable Model-Based Clustering for Large Databases Based on Data Summarization,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11,pp. 1710-1719, Nov. 2005.

[3] R. Nock and F. Nielsen, “On Weighting Clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1223-1235, Aug. 2006.

[4] Bo-Wei Chen ,Jia-Ching Wang et. Al., „A Novel Video Summarization Based on Mining the Story-Structure and Semantic Relations Among Concept Entities“, IEEE Transactions on Multimedia, Feb 2009

[5] Bouguila, N., “A Model-Based Approach for Discrete Data Clustering and Feature Weighting Using MAP and Stochastic Complexity”, IEEE Transactions on Knowledge and Data Engineering, Dec 2009

[6] Mylonas, P. , Spyrou, E. et. Al., ‘Using Visual Context and Region Semantics for High-Level Concept Detection’, IEEE Transactions on Multimedia, Feb 2009

[7] Talavera, L. , Bejar, J. ,” Generality-based conceptual clustering with probabilistic concepts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Feb 2001

[8] Yee Leung , Jiang-Hong Ma et. Al., ‘A new method for mining regression classes in large data sets’, IEEE Transactions on Pattern Analysis and Machine Intelligence, Jan 2001.

[9] Yijuan Lu , Lei Zhang et. Al., ‘Constructing Concept Lexica with Small Semantic Gaps’, IEEE Transactions on Multimedia, June 2010.

[10] YanjunLi, Congnan Luo et. Al., “Text Clustering with Feature Selection by Using Statistical Data”, IEEE Transactions on Knowledge and Data Engineering, Volume**:** 20 , Issue: 5 ,May 2008.

[11] Al-Mubaid, H., Umair, S.A. “A New Text Categorization Technique Using Distributional Clustering and Learning Logic”, IEEE Transactions on Knowledge and Data Engineering, Volume**:** 18 , Issue: 9 . Sept 2006.

[12] Jung-Yi Jiang , Ren-Jia Liou et. Al., ‘A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification’, IEEE Transactions on Knowledge and Data Engineering, Volume**:** 23, Issue:3 Page**(s):** 335 - 349 , March 2011.

Bio Data for the Authors

Mr. V.M. Navaneethakumar obtained M.C.A, from K.S.Rangasamy College of Technology, Tiruchengode Tamil Nadu, India, in 2004, and M.Phil., Computer Science from Periyar University, Salem, Tamilnadu, India in 2008. He is working as Assistant Professor, in Department of Computer Applications, K.S.R College of Engineering, Tiruchengode, Tamilnadu, India

Dr. C. Chandrasekar completed his Ph.D in Periyar University, Salem at 2006. He worked as Head, Department of Computer Applications at K.S.R. College of Engineering, Tiruchengode, Tamil Nadu, India. Currently he is working as Associate Professor in the Department of Computer Science at Periyar University, Salem, Tamilnadu. His research interest includes Mobile computing, Networks, Image processing and Data mining. He is a senior member of ISTE and CSI.