Asian Journal of Information Technology 14 (3): 97-104, 2015
ISSN: 1682-3915

Text Taxonomy Using Datamining Clustering System

[1]A. Ronald Tony and [2]D. Saravanan
[1]Faculty of Computing, Sathyabama University, Chennai, India
[2]Faculty of Operations & IT, IFHE University, IBS Hyderabad, India

Abstract: Usually, in text mining techniques, the term frequency of a term (word or phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in their documents but one term contributes more to the meaning of its sentences than the other term. In this study, a novel concept-based mining model is proposed. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. In the proposed model, three measures for analyzing concepts on the sentence, document and corpus levels are computed. Each sentence is labelled by a semantic role labeller that determines the terms which contribute to the sentence semantics associated with their semantic roles in a sentence. Each term that has a semantic role in the sentence is called a concept. Concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new document is introduced to the system, the proposed mining model can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts. A new concept-based similarity measure which makes use of the concept analysis on the sentence, document and corpus levels is proposed.

Key words: Text mining, clustering, document clustering, similarity measures, text retrieval, intercluster

INTRODUCTION

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. NLP is a term that links back into the history of Artificial Intelligence (AI), the general study of cognitive function by computational processes with an emphasis on the role of knowledge representations. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. Clustering, one of the traditional data mining techniques is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intracluster similarity and low intercluster similarity. Generally, text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups.

Most current document clustering methods are based on the Vector Space Model (VSM) which is a widely used data representation for text classification and clustering (Salton *et al*., 1975). The VSM represents each document as a feature vector of the terms (words or phrases) in the document. Each feature vector contains term weights (usually term frequencies) of the terms in the document. The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure. Methods used for text clustering include decision trees conceptual clustering clustering based on data summarization statistical analysis neural nets inductive logic programming and rule-based systems among others. In text clustering, it is important to note that selecting important features which present the text data properly has a critical effect on the output of the clustering algorithm. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially.

Similarity based on matching of concepts between document pairs is shown to have a more significant effect on the clustering quality due to the similarity's insensitivity to noisy terms that can lead to an incorrect similarity. Through, the rapid growth of multimedia technology, multimedia content can be created, shared and distributed easily (Saravanan and Srinivasan, 2013a, 2012). The concepts are less sensitive to noise when it comes to calculating document similarity. This is due to the fact that these concepts are originally extracted by the semantic role labeller and analyzed with respect to the

Corresponding Author: A. Ronald Tony, Faculty of Computing, Sathyabama University, Chennai, India

sentence, document and corpus levels. Thus, the matching among these concepts is less likely to be found in non-related documents.

Has higher quality than those produced by a single-term analysis similarity only. The results are evaluated using two quality measures, the F-measure and the entropy. Both of these quality measures showed improvement versus the use of the single-term method when the concept-based similarity measure is used to cluster sets of document. The amount of available digital resources is continuously increasing, promoted by a growing interest of users and by the development of new technology for the ubiquitous enjoyment of digital resources (Kumar and Saravanan, 2013).

LITERATURE REVIEW

Data mining methods for knowledge discovery: Data mining and knowledge discovery is a resource collecting relevant common methods and techniques and a forum for unifying the diverse constituent research communities. The journal publishes original technical papers in both the research and practice of data mining and knowledge discovery, surveys and tutorials of important areas and techniques and detailed descriptions of significant applications. Data mining methods: rough sets, Bayesian analysis, fuzzy sets, genetic algorithms, machine learning, neural networks, speech recoginzation (Jurafsky and Martin, 2000) and preprocessing techniques. Numerous illustrative examples and experimental findings are also included (Frakes and Baeza-Yates, 1992).

Recognition of ending of records: The most data reduction techniques in information retrieval use document vectors or term by document matrixes. Using sentences for the comparison of similarities is the recognition of endings of records and inevitability. An algorithm searches for all endings of records like "." and "!". The result of this step are sentences which can be treated as an own unit. The text parser needs the ability to recognize if a punctuation mark is a part of a word or sentence or used as an end of a sentence. The difficulty of this processing step is the distinction if of "." which can be part of a number for example or the end of a sentence. Typing errors make the problem more complex in case of a missing space after the "." (Talavera and Begar, 2001). Hence, the solution of a parser which checks for a blank after the "." is not 100% precise.

Phrase recognition: The parse recognition is closely related to part of speech tagging. The Phrase Recognition (PR) caters for the locating of groups of words, the phrases. PR is needed to keep relations between word groups which would lose their meaning if disjoined. Phrases are similar to compounds in linguistics but are more complex. Phrases exist of different classes:

- Prepositional phrase (e.g., in love)
- Noun phrase (e.g., the queen of England)
- Verb phrase (e.g., do business)
- Adjectival phrase (e.g., large trousers)
- Adverbial phrase (e.g., very quickly)

APRIORI approach to graph-based clustering of text documents: This study introduces a new technique of document clustering based on frequent senses. The developed system, named GDClust (Graph-Based Document Clustering), works with frequent senses rather than dealing with frequent keywords used in traditional text mining techniques. GDClust presents text documents as hierarchical document-graphs and uses an apriori paradigm to find the frequent sub-graphs which reflect frequent senses. Discovered frequent sub-graphs are then utilized to generate accurate sense-based document clusters. We propose a novel multilevel Gaussian minimum support strategy for candidate sub-graph generation. Sub-graph-extension mining that reduces the number of candidates and overhead imposed by the traditional Apriori-based candidate generation mechanism. GDClust utilizes an English language thesaurus (WordNet) to construct document-graphs and exploits graph-based data mining techniques for sense discovery and clustering. It is an automated system and requires minimal hum interaction for the clustering purpose.

Text clustering in concept based data model: Most of text mining techniques are based on word and/or phrase analysis of the text. The statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. However, to achieve a more accurate analysis, the underlying mining technique should indicate terms that capture the semantics of the text from which the importance of a term in a sentence and in the document can be derived. A new concept-based mining model that relies on the analysis of both the sentence and the document, rather than, the traditional analysis of the document dataset only is introduced. The

proposed mining model consists of a concept-based analysis of terms and a concept-based similarity measure to retrieve the document effectively document clustering is needed (Saravanan and Somasundaram, 2014). The term which contributes to the sentence semantics is analyzed with respect to its importance at the sentence and document levels.

The model can efficiently find significant matching terms, either words or phrases, of the documents according to the semantics of the text. The similarity between documents relies on a new concept-based similarity measure which is applied to the matching terms between documents. Experiments using the proposed concept-based term analysis and similarity measure in text clustering are conducted. Experimental results demonstrate that the newly developed concept-based mining model enhances the clustering quality of sets of documents substantially.

Existing system: Existing methods that are used for text clustering include decision trees, conceptual clustering, clustering based on data summarization, statistical analysis, neural nets, inductive logic programming and rule based systems.

Decision trees analysis: Decision trees are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional display in the decision tree interface. The name of the field of data that is the object of analysis is usually displayed, along with the spread or distribution of the va lues that are contained in that field.

Conceptual clustering: Conceptual clustering is one technique that forms concepts out of data incrementally by subdividing groups into subclasses iteratively, thus, building a hierarchy of concepts.

Text clustering using concept-based mining model: Most of text mining techniques are based on word and/or phrase analysis of the text. The statistical analysis of a term (word or phrase) frequency captures the importance of the term within a document. However, to achieve a more accurate analysis, the underlying mining technique should indicate terms that capture the semantics of the text from which the importance of a term in a sentence and in the document can be derived. A new concept-based mining model that relies on the analysis of both the sentence and the document, rather than, the traditional analysis of the document dataset only is introduced. In mining model consists of a concept-based analysis of terms and a concept-based similarity measure. The term which contributes to the sentence semantics is analyzed with respect to its importance at the sentence and document levels. The model can efficiently find significant matching terms, either words or phrases, of the documents according to the semantics of the text. The similarity between documents relies on a new concept-based similarity measure which is applied to the matching terms between documents.

Vector learning for semantic argument classification: Recent works and patents in iterative unsupervised learning have emphasized a new trend in clustering. It basically consists of penalizing solutions via weights on the instance points, somehow making clustering move toward the hardest points to cluster. The motivations come principally from an analogy with powerful supervised classification methods known as boosting algorithms. However, interest in this analogy has so far been mainly borne out from experimental studies only.

Theoretical results show benefits resembling those of boosting algorithms and bring modified (weighted) versions of clustering algorithms such as k-means, fuzzy c-means, Expectation Maximization (EM) and k-harmonic means. Experiments are provided for all these algorithms with a readily available code. They display the advantages that subtle data reweighting may bring to clustering. Usually, in text mining techniques that are mentioned above, the term frequency of a term (word of phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in their documents but one term contributes more to the meaning of its sentences than the other term. But, it is important to note that extracting the relations between verbs and their arguments in the same sentence has the potential for analyzing terms within a sentence where these existing solutions are limiting.

Proposed system: A “sentence-based concept analysis”, “document based concept analysis”, “corpus-based concept analysis” text clustering approach has been proposed that performs “concept-based similarity measure”. The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure.

Data mining technique can be applied in various documents. Multimedia information has become increasingly prevalent and it constitutes a significant component of multimedia contents on the internet (Saravanan and Srinivasan, 2013b). Thus, developing four staged concept-based mining model which target to prove that proposed concept based mining model will improve the text clustering quality. By exploiting the semantic structure of the sentences in documents, a better text clustering result should be achieved. Implemenst new concept term frequency measurement model followed by analysis like sentence-based concept analysis which should be done to analyze the semantic structure of each sentence, semantic structure of the sentence captured in above analysis will be associated with conceptual term frequency measure to capture the sentence concept. Next step of analysis is document-based concept analysis that analyzes each concept at the document level using the concept-based term frequency followed by analysis of concepts on the corpus level using the document frequency global measure. After all analysis concept-based similarity measure that will allow measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document and the discrimination among documents in a corpus have been carried out in this proposed work.

Advantages over existing system: In existing system very long documents make similarity measures difficult. In Vector Space Model, the computational standpoint it is very slow, requiring a lot of processing time. Each time we add a new term into the term space we need to recalculate all vectors. These effects are reduced in our Concept-based Mining Model.

False negative matches: Documents with similar content but different vocabularies may result in a poor inner product. This problem avoided because our proposed work based on semantic structure of each term within a sentence.

False positive matches: Improper wording, prefix/suffix removal or parsing can results in spurious hits (falling, fall+ing; therapist, the+rapist, the+rap+ist; Marching, March+ing; GARCIA, GAR+CIA). This is just a pre-processing limitation are avoided because of our structured semantic approach.

Limitations of proposed system: The assumption that the attributes are independent of each other is often too strong because correlation may exist. Not suitable for large data base data. For large document clustering the database values need to be frequently modified.

PROPOSED METHODOLOGY

Here, we used for separate the sentences for identify the meanings. It takes all the documents one by one for analysis each concept at the sentence level. With the help of that meaning the labels are constructed. Because the stop words are removed by using the labeled terms. The meaningful words are constructed here with the help of stemming the words (Fig. 1).

Text preprocessing

Concept-based analysis algorithm: It is used for calculating concept in sentence (ctf), calculating number of occurrences of a concept in the original document (tf), calculating number of documents containing concept (df):

1. ddoci is a new Document
2. L is an empty List (L is a matched concept list)
3. sdoci is a new sentence in ddoci
4. Build concepts list Cdoci from sdoci
5. for each concept ci 2 Ci do
6. compute ctfi of ci in ddoci
7. compute tfi of ci in ddoci
8. compute dfi of ci in ddoci
9. dk is seen document
10. sk is a sentence in dk
11. Build concepts list Ck from sk
12. for each concept cj 2 Ck do
13. if (ci ═ cj) then
14. update dfi of ci
15. compute ctfweight ¼ avgoctfi; ctfjÞ
16. add new concept matches to L
17. end if
18. end for
19. end for
20. output the matched concepts list L

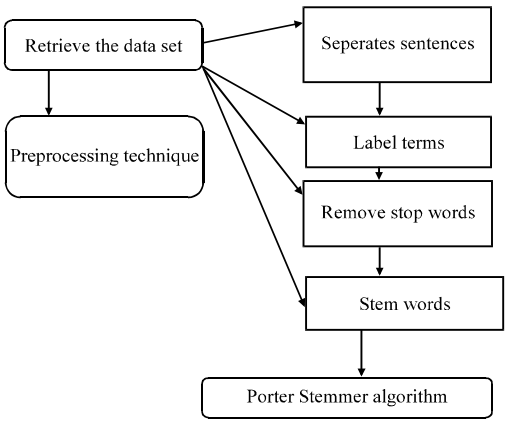

Fig. 1: Text preprocessing

The concept-based analysis algorithm describes the process of calculating the ctf, tf and df of the matched concepts in the documents. The procedure begins with processing a new document (at line 1) which has welldefined sentence boundaries. Each sentence is semantically labeled according. The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations in section each concept (in the for loop, at line 5) in the verb argument structures which represents the semantic structures of the sentence is processed sequentially. Each concept in the current document is matched with the other concepts in the previously processed documents. To match the concepts in previous documents is accomplished by keeping a concept list L which holds the entry for each of the previous documents that shares a concept with the current document. After the document is processed, L contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, L is output as the list of documents with the matching concepts and the necessary information about them. The concept-based analysis algorithm is capable of matching each concept in a new document (d) with all the previously processed documents in O(m) time where m is the number of concepts in d.

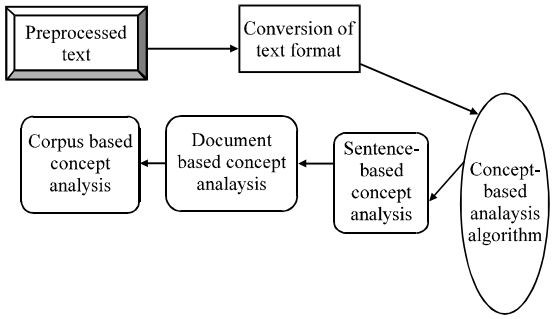

Fig. 2: Implementation of concept based analysis

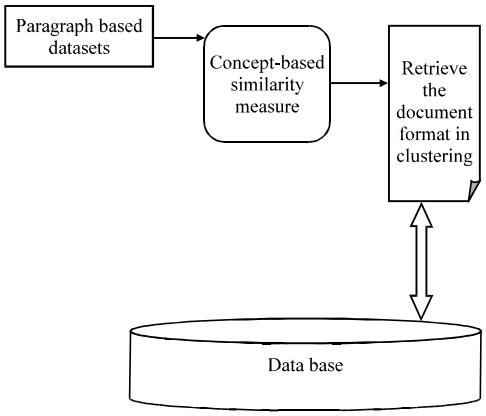

Fig. 3: Implementation of Clustering algorithm

Implementation of concept-based analysis: This concept-based module is used for form the sentences and the paragraphs based on the concept (Fig. 2). The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document and corpus levels rather than a single-term analysis on the document only. It is comprised of sentence-based concept analysis, document-based concept analysis, corpus-based analysis and concept-based analysis.

In sentence based concept at the sentence level, Conceptual term frequency is proposed for for analyze each concept at the sentence level. In document based concept analysis term frequency is used for analyze each concept at the document level. In corpus-based concept analysis, document frequency is used for extract concepts that can discriminate between documents. All the processes are done here to form the concept oriented structure.

Implementation of clustering algorithm: The clustering algorithm is used for form all the concepts in the form of clusters. By using one of the clustering algorithm we can form all the document as related concept. Because it is the main thing to form the various concepts and produce the concept related output. It is used to measuring the importance of the each concept based on the semantics of the sentence, the topic of the document and at last it produces the output (Fig. 3).

EXPERIMENTAL OUTCOMES

Here, we are consider the some of the text document and text data, it perform the text preprocessing, text classification, text clustering. The out come of the above process is shown below in Fig. 4-11.

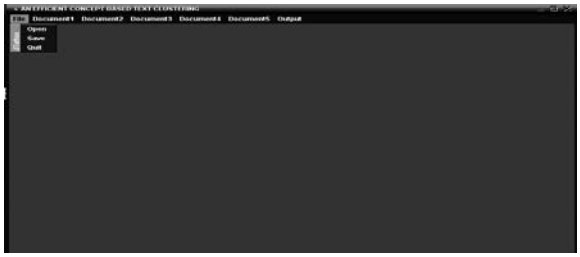

Fig. 4: Document clustering process

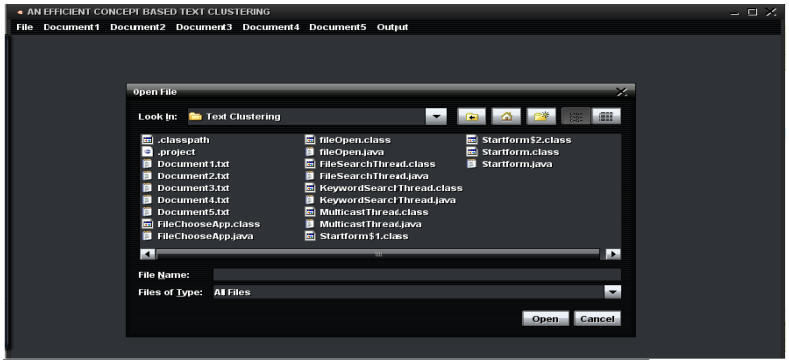

Fig. 5: Text uploading for text processing

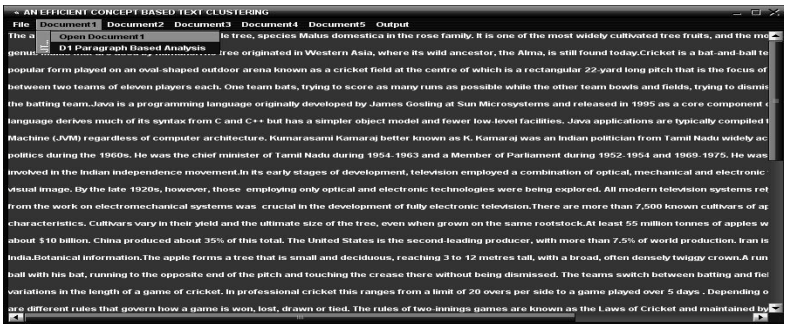

Fig. 6: Text extraction process

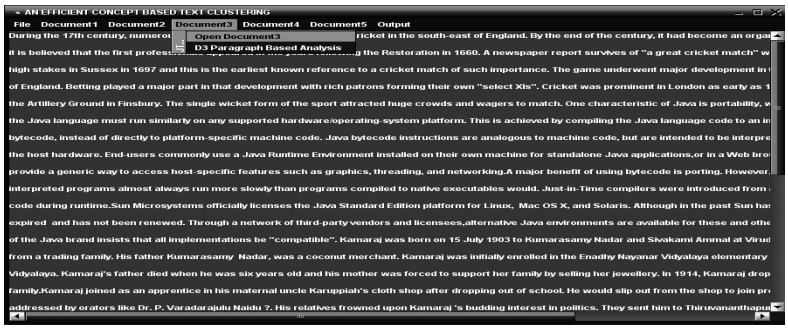

Fig. 7: Text preprocessing

Fig. 8: Text clustering process

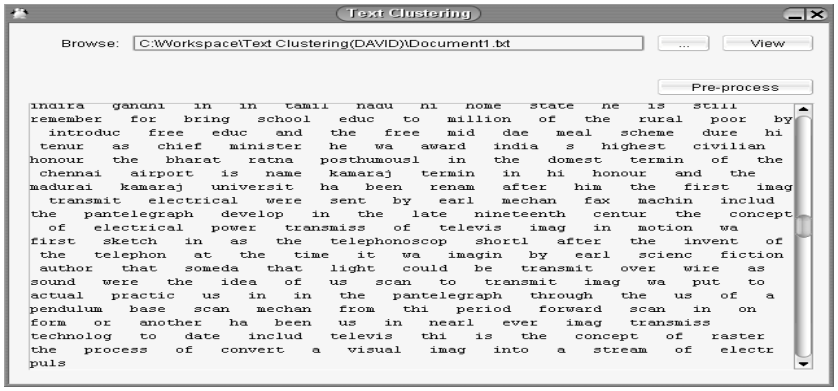

Fig. 9: Concept related output

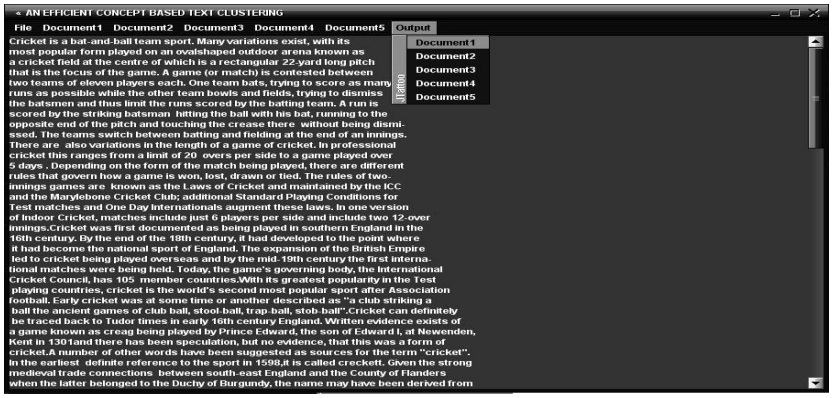

Fig. 10: Paragraph based analysis

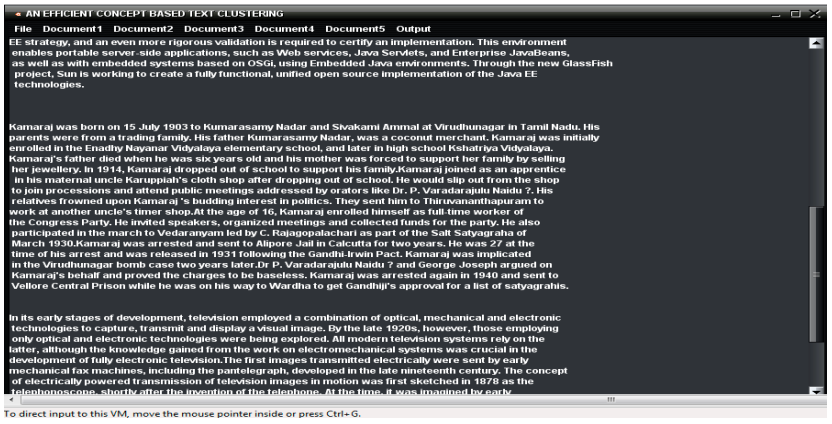

Fig. 11: Text classification and web mining

CONCLUSION

The gap between natural language processing and text mining disciplines. New Concept based Mining Model composed of four components is proposed to improve the text clustering quality. By exploiting 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 structure of each sentence to capture the sentence concepts using the proposed conceptual term frequency ctf measure. Then, the second component, document-based concept analysis, analyzes each concept at the document level using the concept-based term

frequency tf. The third component analyzes concepts on the corpus level using the document frequency df global measure. The fourth component is the concept-based similarity measure which allows measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document and the discrimination among documents in a corpus.

The factors affecting the weights of concepts on the sentence, document and corpus levels, a concept-based similarity measure that is capable of the accurate calculation of pair wise documents is devised. This allows performing concept matching and concept-based similarity calculations among documents in a very robust and accurate way. The quality of text clustering achieved by this model significantly surpasses the traditional single term based approaches.

RECOMMENDATION

Future, enhancement to our research can be many in number like it can be used to our research to web document clustering, some text classification can be applied to our research, our model can be combined with corpora used to investigate efficiency on classification on compare other traditional method.

REFERENCES

Frakes, W.B. and R. Baeza-Yates, 1992. Information Retrieval Data Structures and Algorithms. Prentice-Hall Inc., Englewood Cliffs, New Jercy.

Jurafsky, D. and J.H. Martin, 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, New Jersey.

Kumar, A.R. and D. Saravanan, 2013. Content based image retrieval using color histogram. Int. J. Comput. Sci. Inform. Technol., 4: 242-245.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.

Saravanan, D. and S. Srininvasan, 2012. Video image retrieval using data mining techniques. J. Comput. Applic., 5: 39-42.

Saravanan, D. and S. Srinivasan, 2013a. Video information retrieval using: Chemeleon clustering. Int. J. Emerg. Trends Technol. Comput. Sci., 2: 166-170.

Saravanan, D. and S. Srinivasan, 2013b. Matrix based indexing technique for video data. J. Comput. Sci., 9: 534-542.

Saravanan, D. and V. Somasundaram, 2014. Matrix based sequential indexing technique for video data mining. J. Theor. Applied Inform. Technol., 67: 725-731.

Talavera, L. and J. Begar, 2001. Generality-based conceptual clustering with probabilistic concepts. Trans. Pattern. Anal. Mach. Intell., 23: 196-206.