Asian Journal of Information Technology 15 (6): 996-1004, 2016
ISSN: 1682-3915

Enhancing the Quality of Search Results by a Novel Semantic Model

[1]R. Thilagavathy and [2]R. Sabitha
[1]Department of Computer Science and Engineering, Sathyabama University,
[2]Department of Information Technology, Jeppiaar Engineering College,
Tamil Nadu, 119 Chennai, India

Abstract: In text mining most of the methods are based on the concept of term (i.e., a word or a phrase) analysis. Statistical analysis usually identifies the important terms by means of their frequency within a document. However, more than one term may contain the identical occurrences within the document, however a specific term plays major role towards sentence semantics comparing to the remaining term. Therefore, the fundamental web document clustering method should specify term which identifies meaning of the text. In this case, the semantic-based method identifies expressions that represent the sentence semantics which are very helpful in determining the document's subject. This mining model analyses words or expressions on the individual sentences, documents and core level. The semantic-based model dramatically distinguishes among insignificant terms against the meaning of the sentences and terms which are more close to the sentence semantics. The proposed method can effectively find important similar concepts among documents with respect to the meaning of their sentences. The interrelations among the documents are estimated by a similarity measure which is based on concepts. By using the semantic organization of sentences in the web documents, a considerable improvement in the quality of web document clustering is achieved.

Key words: Text mining, web document clustering, term frequency, concept-based similarity, conceptual term frequency, suffix tree clustering

INTRODUCTION

Today, the internet is used as a major data storage environment and it is struggling with the problem of information overload. At the same time, more number of peoples uses the web as their fundamental source of information. The availability of a large quantity of information with the dynamic and disparate characteristics of the web makes information retrieval as a vague process for the normal user. Many applications have been developed to assist the users, in order to fulfil their information needs easily and quickly (Sambasivam and Theodosopoulos, 2006; Fang *et al*., 2011).

Normally, a person looking for information surrenders a query which consists of a set of keywords to a search engine. Then the search engine does correct matching among the query terms and the key words that describe each web page and exhibit the outcome to the user. Usually, the outcomes of the search engines are long list of URLs which are very difficult to search. Also, users lacking in domain knowledge or unfamiliar with the proper phraseology are frequently submits the incorrect query terms which leads many unrelated pages will be displayed in the result.

The above mentioned situation has motivated the growth of novel methods to help the users to track, find and organise the web documents on hand effectively (Carpineto *et al*., 2009). The main objective of these methods is finding results which are more related to the users needs. Document clustering is one of the methods that play a significant role for obtaining this objective. Later, wide ranges of clustering algorithms were developed for improving the significance of document clustering and to fulfil the needs of applications which are related to information processing or management (Deng *et al*., 2013; Fang *et al*., 2011).

Web mining is the sub-field of data mining which automatically discover and extract information from web documents and services (Sambasivam and Theodosopoulos, 2006). Text mining is also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. Owing to the daily swift growth of the information, there are considerable needs in extracting and discovering valuable knowledge from the vast amount of information found in different data sources today such as world wide web. Data mining in general is the field of

Corresponding Author: R. Thilagavathy, Department of Computer Science and Engineering, Sathyabama University, Tamil Nadu, 119 Chennai, India

extracting useful information and sometimes high-level knowledge from large sets of raw data (Nora and Vazirgiannis, 2010). Text mining is generally considered more difficult than traditional data mining. This is attributed to the fact that traditional databases have fixed and known structure while text documents are unstructured, or as in the case of web documents, semi-structured. Thus, text mining involves a series of steps for data pre-processing and modelling in order to condition the data for structured data mining. Text mining can help in many tasks that otherwise would require large manual effort. Common problems solved by text mining include but not limited to, searching through documents, organizing documents, comparing documents, extracting key information and summarizing documents. 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 (Pradhan *et al*., 2005; Nock and Nielsen, 2006). The need for representations of human knowledge of the world is required in order to understand human language with computers. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. The problem introduced by text mining is that natural language was developed for humans to communicate with one another and to record information. Computers are a long way from understanding natural language. Humans have the ability to understand the meaning of text and humans can easily overcome obstacles that computers cannot easily handle such as spelling variations and contextual meaning. However, although human mind can understand the meaning of unstructured data, human lacks the computers ability to process text in large volumes or at high speeds. Herein lays the key to create a novel technology called concept-based mining (Shehata *et al*., 2006) that combines the human way of understanding with the speed and accuracy of a computer.

Text mining techniques are mostly based on Vector space model (term frequencies) (Salton, 1975). It captures the importance of the term within a document. But two terms can have the same frequency in the same document so, there is a possibility of redundant and irrelevant results. But the meaning that one term contributes might be more appropriate than the meaning contributed by the other term. Many approaches have been developed to handle these kind of situation in the field of web document clustering (Goyal *et al*., 2013; Iosif and Potamianos, 2010; Pradhan *et al*., 2003). A new concept-based document clustering model is introduced which analyses the terms based on the sentence, document and corpus level. 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. The process of calculating ctf, tf, df measures (Shehata *et al*., 2006) in a corpus is attained by the proposed model which is called semantic-based mining model. By doing so, we cluster the web documents in an efficient way and the quality of the clusters achieved by this model significantly surpasses the traditional single-term and phrase based approaches (Salton, 1975; Hammouda and Kamel, 2004). Semantic based document clustering model helps to achieve the below features:

- To give the exact results requested by the user
- To speed up the searching process
- To filter the refined keywords regarding the given keyword
- To frequently update the changes made in web documents
- Incorporating web document clustering makes the search engine to give exact and accurate information to satisfy the end user

The significant terms used in this study:

- Verb argument structure: (e.g., Ravi hits the ball). “hits” is the verb. “Ravi” and “the ball” are the arguments of the verb “hits”
- Label: a label is assigned to an argument, e.g.,: “Ravi” has subject (or agent) label. “the ball ”has object (or theme) label
- Term: is either an argument or a verb. Term is also either a word or a phrase (which is a sequence of words)
- Concept: in the new proposed mining model, concept is a labeled term

Case role analysis: In general, the semantic structure of a sentence can be characterized by a form of verb argument structure. This underlying structure allows the formation of a compound meaning representation from the meanings of the individual concepts in a sentence (Mitra *et al*., 2002; Gildea and Jurafsky, 2002; Pradhan *et al*., 2003). The verb argument structure allows a link between the arguments in the surface structures of the input text and their related semantic roles (Pradhan *et al*., 2004). Consider the following example: my son wants a bike. This example has the following

syntactic argument frames: (Noun Phrase (NP) wants NP). In this case, some facts could be driven for the particular verb "wants":

- There are two arguments to this verb
- Both arguments are NPs
- The first argument "my son" is preverbal and plays the role of the subject
- The second argument "a bike" is a post-verbal and plays the role of the direct object

MATERIALS AND METHODS

Semantic model: The above discussed methods in the previous research are used for document clustering but these methods are used for clustering the documents which are available on system (Shehata *et al*., 2006; Egozi *et al*., 2001). But, the proposed work is going to make use of web documents instead of plain text documents. As given in Fig. 1, this web document clustering method makes use of concept-based mining model which analyses the important terms at the sentence level, document level as well as at the core level. Also, this model has a "concept-based similarity measure" and that helps to improve cluster quality even more.

The proposed system accepts web documents as the input and processes them to produce the best clusters which improve the quality of search-engine results. Each document is confined with the well-defined sentences. The semantic roll labeller automatically label every sentence in a document as per the "Prop Bank notations" and the output will be the labelled "verb argument structure" (Gildea and Jurafsky, 2002; Pradhan *et al*., 2005). The labelling processs results will be taken and then analysed using the "concept-based model" at the sentence, document as well as at the core level. According to this model, the labelled terms are identified as concept.

Data preprocessing

Step 1: Separation of sentences and label terms The process of mining the relationship among verbs and their arguments in a sentence is capable of analysing the terms contained in a sentence. As a result of identifying this relationship, the importance of every term in a sentence towards the sentence semantics will be known easily. In this proposed web document clustering method, the semantic role labeller assigns labels to each sentence in a document that best determines the terms which are very close to the meaning of the sentence. Generally, the verb argument structure is used to characterize the sentence-semantic structure and it makes a connection among the arguments of the input query with their corresponding semantic-role.

The "Support Vector Machines (SVMs)" are used for identifying the arguments of verbs in a sentence (Pradhan *et al*., 2005). Also, SVMs classifies the arguments by their semantic roles like Agent, Theme and Goal. SVMs results improved performance over the earlier classifiers.

Step 2: Removal of stop words and stem words once forming the verb argument structure, a data cleaning process is performed to eliminate the stop words and Porter Stemmer algorithm is used for stemming the words. After the removal of stop word and stem words the resulting terms are called as concepts:

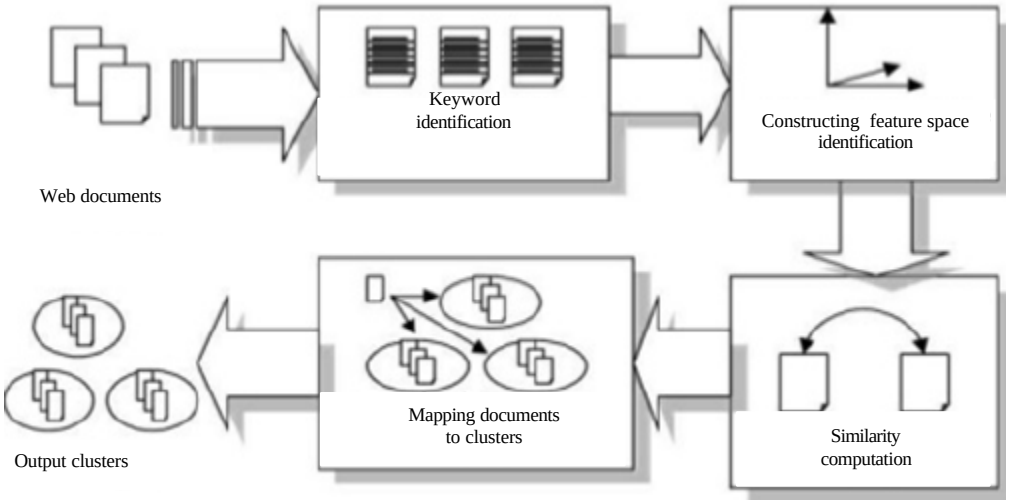

Fig. 1: Semantic-based model for web document clustering

- Step 1: gets rid of plurals and-ed or-ing suffixes
- Step 2: turns terminal y to i when there is another vowel in the stem
- Step 3: maps double suffixes to single ones:-ization,-ational, etc
- Step 4: deals with suffixes,-full,-ness, etc
- Step 5: takes off-ant,-ence, etc
- Step 6: removes a final-e Porter stemmer helpers

Semantics-based analysis: The "semantic s-based analysis" is mainly used to get perfect investigation about concepts on every sentences and documents as well as on the core level instead of analysing the documents only in single-term fashion.

Concept analysis at sentence level: The frequency of the labelled term (or concept) is calculated by using a frequency measure, known as "conceptual term frequency (ctf)" which analyses the concepts at sentence level. Computing 'ctf ' for labelled terms in a sentence. The 'ctf' for a concept is calculated as how many times a labelled term occurred in the verb argument structure corresponding to a sentence. At this point, the frequency measure analyses concepts within a sentence. Computing 'ctf ' for labelled terms in a document. A same labelled term (or concept) may have various,, ctf. Values at different sentences in a document. Therefore, the frequency of a labelled term in a document is computed using:

$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn} \quad (1)$$

Where, sn_number of sentences holds the given concept in a document.

Concept analysis at document level: Each concept is analysed at document level by using the "concept based term frequency (tf)" and it is computed by means of calculating, the number of times a concept is occurred in the given document. At this point, the 'tf' of a concept is analysed at the document level.

Concept analysis at corpus level: The concepts or labelled terms which are making best discrimination among documents are extracted by using the "concept-based document frequency (df)". The value of 'df' clearly tells how many documents that contain a given concept. The 'df' is a kind of global measure because here the concept is analysed at the corpus level.

Concept analysis algorithm:

Wddoci is a new web Document
E is an empty List (E is a matched concept list)
Web documents consist of various markup language formats presentational, procedural and descriptive markup
Sdoci is a new sentence in $wddoc_i$
Build concepts list $Cdoc_i$ from $Sdoc_i$
For each concept $c_i \in C_i$
Evaluate ctf_i of c_i in $wddoc_i$//conceptual term frequency
Evaluate tf_i of c_i in $wddoc_i$//Term frequency tf-no of occurrences of given terms in a web document-wd
Evaluate df_i of c_i in $wddoc_i$//Document frequency df-no of web docs. Contains concept c
d_n is seen document where n = (0; 1; . . . ; doc_{i-1})
S_n is a sentence in d_n
Build concepts list C_n from S_n
For each concept $c_j \in C_n$ do
If ($c_i = c_j$) then
Update df_i of c_i
Compute ctf_i weight = avg (ctf_i, ctf_j)
Add new concept matches to L
End if
End for
End for
If wd has presentational markup
Binary codes are used in the text
Instead of searching the concepts in wd, search the binary codes in the wd
Else if wd contains procedural markup Continue with the steps from 6-20
Else wd has Descriptive markup Continue with the steps from 6-20
Output the matched concepts list E

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 that has definite sentence borders. Each sentence is semantically labelled according to (Pradhan *et al*., 2004). Matched concept length of each sentence and their verb argument structures are stored for the concept-based similarity calculations. The concept in the verb argument structures represents the semantic structure of each sentence is processed sequentially. Each concept in the current document is coordinated with the other concepts in the previously processed documents. To match the concepts in previous documents is gifted by keeping a concept list E which holds the entry for previous documents which shares a concept with the existing document. After the document is processed, E contains all the matching concepts between the current document and previous document. So, previously processed documents shares at least one concept with the new document. Finally, E is output as the list of documents with the matching concepts and the required information about them. The concept-based analysis algorithm is capable of matching each concept in a new document with all the previously processed documents in certain time.

Example of calculating the proposed conceptual term frequency (CTF) measure: Consider the following sentence: texas and Australia researchers have created industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles. In this sentence, the semantic role labeller identifies three words (verbs), marked by bold which are the verbs that

represent the semantic structure of the meaning of the sentence. These verbs are "created, made and lead ". Each one of these verbs has its own arguments as follows: [ARG0 Texas and Australia researchers] have [TARGET created] [ARG1 industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles].

[ARG1 industry-ready sheets of materials] [TARGET made] from [ARG2 nanotubes that could lead to the development of artificial muscles]. [ARG1 nanotubes] that [ARGM-MOD could] [TARGET lead] [ARG2 to the development of artificial muscles].

Arguments labels are numbered ARG0, ARG1, ARG2 and so on depending on the valiancy of the verb in sentence. The meaning of each argument label is defined relative to each verb in a lexicon of frames files. Despite this generality, ARG0 is very consistently assigned an Agent-type meaning while ARG1 has a patient or theme meaning almost as consistently. Thus, this sentence consists of the following three verb argument structures. First verb argument structure for the verb created: [ARG0 Texas and Australia researchers] [TARGET created]

[ARG1 industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles]. Second verb argument structure for the verb made: [ARG1 industry-ready sheets of materials] [TARGET made] [ARG2 nanotubes that could lead to the development of artificial muscles].

Third verb argument structure for the verb lead: [ARG1 nanotubes] [ARGM-MOD could] [TARGET lead] [ARG2 to the development of artificial muscles].

All the labelled items are considered are considered as a concept. The identified concept for the above given sentence is:

- Concepts in the first verb argument structure
- Texas Australia researchers
- Created
- Industry ready sheets materials made nanotubes lead development artificial muscles
- Concepts in the second verb argument structure
- Industry ready sheets materials
- Made
- Nanotubes lead development artificial muscles
- Concepts in the third verb argument structure
- Notubes
- Lead
- Development artificial muscles (Table 1)

Concept-based document similarity: The measure which involves concept based similarity is influenced by the following important features: first, the labelled terms which determine each sentences semantic structure, produced by the pre-processing step are considered as concepts. Second, the number of occurrences of a labelled term helps to identify the importance of the term towards the sentence-semantics, in addition to the documents main subject. Third, the concept-based document frequency 'df' is used to distinguish between documents while computing the document similarity. For calculating concept-based similarity between two documents, the following informations are needed:

Table 1: Example of calculating CTF measure

Sentences and Individual concepts	CTF
Taxas Australia researchers	1
Created	1
Industry ready sheets materials nanotubes lead development artificial muscels	1
Materials nanotubes lead development artificial muscels	2
Nanotubes	2
Lead	3
Development artificial muscels	3
Taxas	3
Australia	1
Researchers	1
Industry	1
Ready	1
Sheets	1
Development	3
Artificial	3
Muscels	3

- M = Number of similar concepts in each document, d
- Sn = Number of sentences having similar concept in each document, d
- v = Number of verb-argument structure in each sentences
- Ctf, tf and df of each concept in every document
- L = Length of the concept
- Lv = Length of the verb-argument structure which contains the concept

The similarity among the documents, doc1 and 2 in terms of concept is computed by:

$$Sim_c(doc_1, doc_2) = \sum_{i=1 \text{ to } m} Max(l_{i1}/Lv_{i1}, l_{i2}/Lv_{i2}) \; X \; weight_{i1} X \; weight_{i2} \quad (2)$$

$$Weight_i = (tf\ weight_i + ctf\ weight_i) \; X \log(N/df_i) \quad (3)$$

The sum between the two values of tf $weight_i$ and ctf $weight_i$ presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document. The multiplication between log (N/df_i) value and (tf weighti+ctf $weight_i$) value finds the concepts that can efficiently discriminate among documents of the entire corpus.

Concept based suffix tree clustering: The similarity measures which are computed by using the document, sentence, corpus and combined approach concept analysis are used to compute four similarity matrices in documents. Document clustering techniques used in the proposed model includes semantic suffix tree and k-means algorithm (Sambasivam and Theodosopoulos, 2006; Chim and Deng, 2007) Suffix Tree Clustering (SSTC) is used to cluster web documents with the help of semantic similarities (Wang *et al*., 2008). Suffix tree is constructed through on-depth and on-breadth pass. The matched concepts found in the verb argument of whole document are identified by concept analysis algorithm and the matched concepts are clustered using suffix tree clustering. The concept based suffix tree is an incremental and linear time clustering algorithm. It uses only string matching on the suffix tree structure for finding shared common phrases of documents (Wang *et al*., 2008). This concept based suffix tree will use both semantic similarity and string matching as conditions to create the suffix tree. The concept based Suffix tree clustering has following phases:

- Phase 1: constructing suffix tree for matched concepts
- Phase 2: tree pruning
- Phase 3: identify clusters

Identified clusters are formed as root nodes and then matching nodes are joined to form a base node. Concept based suffix tree algorithm gives exact relevant documents from specific clusters to users. If some of the documents in suffix tree are not able to clustered to form a base node then for that particular documents, we can use k-means algorithm to form a base node.

Concept based suffix tree algorithm:

```
Input<--set of snippets (i.e., matched concepts list)
Output<--set of final cluster
Phase 1: construct tree
For each snippets
Extract the sentences
For each sentence
Construction of suffix tree
}
}
//Phase2: tree pruning
Tree pruning step
Compact tree step
//Phase3: Identify cluster
Keep base cluster
Finding final cluster
Phase 1: constructing suffix tree
Input <-set of String
Output <-- semantic suffix tree
For each word (txt)
If root is empty then{
Create a new node
And update position
} else {
Add a new node into cn
Do until pn = root {
If SemSim = 0 && no match then{
Add a new node into pn
And update suffix link
Update position
} else {update suffix link
And update position }
}
}
}
Phase 2: tree pruning
Assign gn = root
For each Ti
If Ti is subset Tj then Ti is deleted
else{
Assign pn node
For each pn .s child node{
Assign cn
For each cn .s child node{
If cn has suffix link then {
Assign sn
Case gn!=sn {
If sns childs related pn||ccn
Then move sn to replace cn
Else{move sn .s documents to cn
And move csn into cn}
}
Case gn==sn {
Move cn .s documents into pn
And delete cn branch }
}
gn = pn, pn = cn, cn = ccn
}
}
}
}//end of tree pruning step
//compact tree
If pn .s childs = = 1and documents = = null
Then pn and cn are combined
Phase 3: Identify clusters
Assign gn=root;
For each sub-tree{
Assign pn node
Collect into CanCluster
If pn has child node {
//use stack to keep right node
//and work until stack is empty
For each child of pn node
keep cluster&label to CanCluster
If cn does not child node{
Compute ClusSim
Filter the useless cluster
} } }
Transfer CanCluster to final_cluster
And clear CanCluster
}
Return final_cluster
```

This illustration shows the complete semantic suffix tree of two documents: D1 = {texas and Australia, created, industry-ready sheets}, D2 = {industry-ready sheets, made, nanotubes} (Fig. 2).

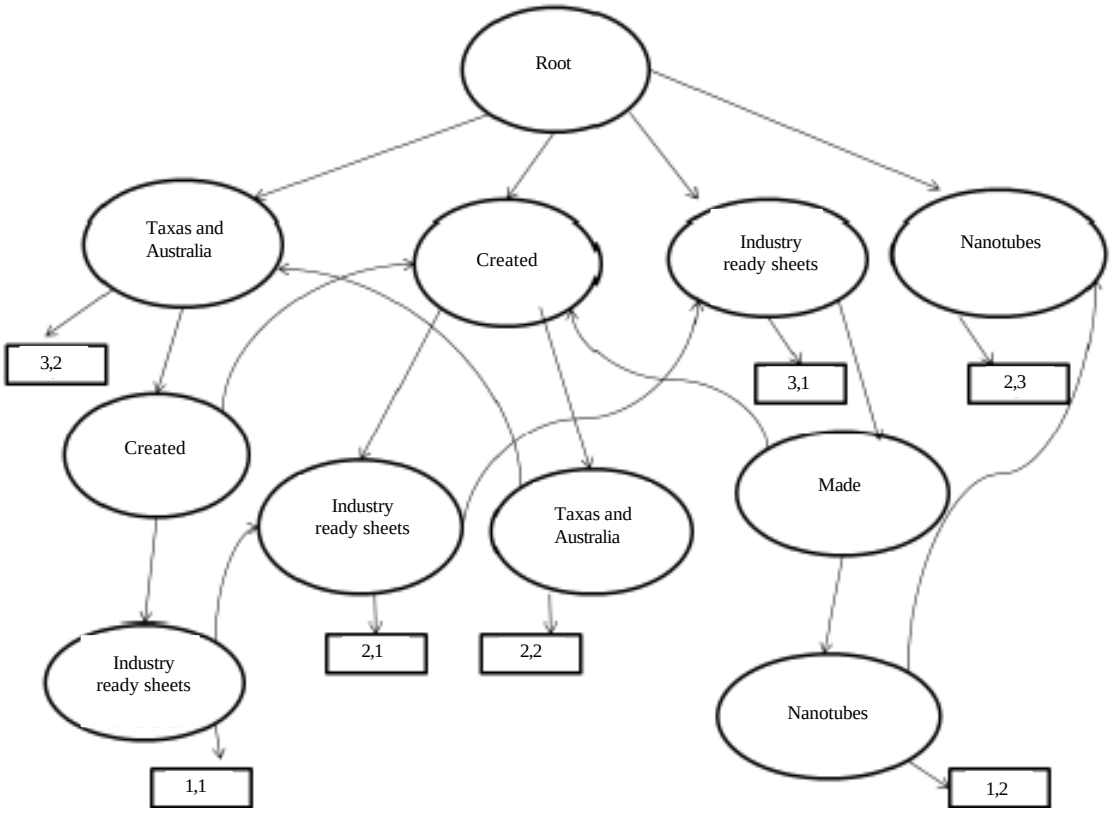

Fig. 2: Concept based suffix tre

RESULTS AND DISCUSSION

To check the potential of our proposed model, we experiments this proposed model on three different datasets those are abstract articles from ACM digital library, articles from IEEE explorer and messages collected from different newsgroups. Totally, we have used 38,500 web documents for analysing the proposed research. From each web document only the text are analysed without including the metadata. For pre-processing the content of web documents we have used porter stemmer algorithm. Suffix Tree Clustering (STC) is one of the widely used algorithm for clustering web documents (Wang *et al*., 2008; Chim and Deng, 2007).

Therefore, we have chosen to use a variation STC algorithm called concept based STC to test the impact of concept-based similarity in web document clustering. The value of ctf, tf and df are calculated for each concept. The combined weight of these three values is calculated as given in Eq. 3.

Our experimental result shows that, the proposed model produce quality clusters comparing to the earlier single-term method. We have chosen two quality measures called F-measure and entropy for identifying the quality of clusters produced. F-measure is identified by combining the Precision (P) and Recall (R) values. Precision and recall of cluster j with respect to class i are calculated as:

$$P = \text{Precision}(i,j) = \frac{M_{ij}}{M_j} \quad (4)$$

$$R = \text{Recall}(i,j) = \frac{M_{ij}}{M_j} \quad (5)$$

Where: Mij represents how many members of class i are there in cluster j, Mj is the total members of cluster j and Mi is the total member of class i. The F-measure of a particular class i is calculated as:

$$F(i) = 2PR/P + R \quad (6)$$

The overall F-measure is calculated as:

$$F\left(C = \left(\sum_i = \left(|i| \times F(i)\right) / \sum_i = |i|\right)\right) \quad (7)$$

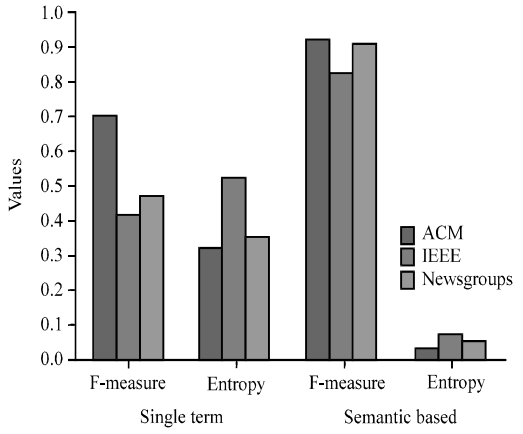

Fig. 3: Performance comparison between single-term and concept-based web document clustering

Table 2: Clustering improvements using single-term method and semantic-based method

	Single-term		Semantic-based	
Data set	F-measure	Entropy	F-measure	Entropy
ACM	0.697	0.317	0.921	0.021
IEEE	0.411	0.523	0.817	0.061
Newsgroups	0.471	0.345	0.905	0.042

Table 3: The average of precision, overlap, no Of nodes and no of branches compared between STC and concept-based STC

Algorithm	Precision	Overlap	No Of nodes	No Of branches
STC	0.69	11.02	2,28,061	38,005
Concept-based STC	0.84	5.40	1,36,563	23,863

The second measure Entropy is used to estimate how similar a cluster is. The higher similarity of a cluster gives the lower entropy value and vice versa. The overall entropy for a group of clusters is calculated as:

$$E_c = \sum_{j=1}^{M}\left(\frac{M_j}{M}\times E_j\right) \quad (8)$$

Where: E_j is the entropy of each cluster calculated by equation. The membership of cluster j belongs to class i is computed by the probability pij (Table 2 and Fig. 3).

Also to evaluate the performance of concept based STC algorithm, we compare with the STC algorithm or precision, coverage, overlap, number of node and number of branch. The results of comparing are as shown in Table 3. The results clustering engine is created from the above dataset.

From Table 3, the precision of concept-based STC is higher than the STC algorithm because the concept-based approach tries to get specic clusters. Users can examine specic clusters to access the relevant document. However, the coverage and overlap measure is less than the STC algorithm. The concept-based STC can reduce the number of nodes and branches of the original STC >40.12 and 37.21%, respectively.

CONCLUSION

This study proposes a semantic model and a concept-based STC algorithm which makes an association among text mining and natural language processing disciplines. The "semantic-based mining model" is being used by the proposed system will results substantial improvements in the quality of the clusters. The determination of sentence semantic structure in documents plays an important role in producing better clustering results. By combining the factors affecting, the weights of concepts on the sentence, document and corpus levels, a concept-based similarity measure 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 strong and accurate way with the help of new concept-based STC and k-means algorithms. The quality of the output clusters obtained by these methods can have considerable improvements than the traditional single term-based approach. In future, there are number of possibilities for extending this study. One among the different way is to applying the same model for text classification which analyses the use of this new technique on other corpora and its effect on classification compared with existing methods.

REFERENCES

Carpineto, C., S. Osinski, G. Romano and D. Weiss, 2009. A survey of web clustering engines. ACM Comput. Sur., Vol. 41. 10.1145/1541880.1541884

Chim, H. and X. Deng, 2007. A new suffix tree similarity measure for document clustering. Proceedings of the 16th international Conference on World Wide Web, May 8-12, 2007, Banff, AB, Canada, pp: 121-130.

Deng, T., L. Zhao, H. Wang, Q. Liu and L. Feng, 2013. ReFinder: A context-based information refinding system. IEEE Trans. Knowledge Data Eng., 25: 2119-2132.

Egozi, O., S. Markovitch and E. Gabrilovich, 2011. Concept-based information retrieval using explicit semantic analysis. ACM Trans. Inform. Syst., Vol. 29. 10.1145/1961209.1961211

Fang, Y., L. Si and A.P. Mathur, 2011. Discriminative probabilistic models for expert search in heterogeneous information sources. Inform. Retrieval, 14: 158-177.

Gildea, D. and D. Jurafsky, 2002. Automatic labeling of semantic roles. Comput. Linguistics, 28: 245-288.

Goyal, P., L. Behera and T.M. McGinnity, 2013. A context-based word indexing model for document summarization. IEEE Trans. Knowledge Data Eng., 25: 1693-1705.

Hammouda, K.M. and M.S. Kamel, 2004. Efficient phrase-based indexing for web document clustering. IEEE Trans. Knowledge Data Eng., 16: 1279-1296.

Iosif, E. and A. Potamianos, 2010. Unsupervised semantic similarity computation between terms using web documents. IEEE Trans. Knowledge Data Eng., 22: 1637-1647.

Mitra, P., C.A. Murthy and S.K. Pal, 2002. Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell., 24: 301-312.

Nock, R. and F. Nielsen, 2006. On weighting clustering. IEEE Trans. Pattern Anal. Mach. Intell., 28: 1223-1235.

Nora, O. and M. Vazirgiannis, 2010. A Review of Web Document Clustering Approaches. In: Data Mining and Knowledge Discovery Handbook, Maimon, O. and L. Rokach (Eds.). 2nd Edn., Springer Science and Business Media, New York.

Pradhan, S., K. Hacioglu, V. Krugler, W. Ward and J.H. Martin et al., 2005. Support vector learning for semantic argument classification. Mach. Learn., 60: 11-39.

Pradhan, S., K. Hacioglu, W. Ward, J.H. Martin and D. Jurafsky, 2003. Semantic role parsing: Adding semantic structure to unstructured text. Proceedings of the 3rd IEEE International Conference on Data Mining, November 19-22, 2003, Boulder, CO, USA., pp: 629-632.

Pradhan, S., W. Ward, K. Hacioglu, J. Martin and D. Jurafsky, 2004. Shallow Semantic parsing using support vector machines. Proceedings of the Human Language Technology Conference/North American chapter of the Association of Computational Linguistics (HLT/NAACL), May 2-7, 2004, Boston, MA., pp: 233-240.

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

Sambasivam, S. and N. Theodosopoulos, 2006. Advanced data clustering methods of mining Web documents. Inform. Sci. Inform. Technol., 3: 564-579.

Shehata, S., F. Karray and M. Kamel, 2006. Enhancing text clustering using concept-based mining model. Proceedings of the IEEE 6th International Conference on Data Mining, December 18-22, 2006, Hong Kong, China, pp: 1043-1048.

Wang, J., Y.Mo, B. Huang, J. Wen and L. He, 2008. Web Search Results Clustering Based on a Novel Suffix Tree Structure. In: Autonomic and Trusted Computing, Rong, C., M.G. Jaatun, F.E. Sandnes, L.T. Yang and J. Ma (Eds.). Springer, New York, ISBN: 978-3-540-69295-9, pp: 540-554.