Concept Mining in Text Documents using Clustering

K.N.S.S.V. Prasad
CSE Dept.
MANIT Bhopal 462003, MP

S. K. Saritha, PhD
CSE Dept.
MANIT Bhopal 462003, MP

ABSTRACT
Due to daily quick growth of the information, there are considerable needs to extract and discover valuable knowledge from data sources such as World Wide Web. The common methods in text mining are mainly based on statistical analysis of term either phrase or word. These methods consider documents as bags of words and they will not give any importance to meanings of document content. In addition, statistical analysis of term frequency extracts the significance of term within a document only. Whenever any 2 terms might have same frequency in their documents, but only 1 term pays more to meaning of its sentences than other term.The concept-based model that analyses terms on corpus, document and sentence levels instead of ancient analysis of document is introduced. The planned model consists of, concept-based analysis, clustering by using k-means, concept-based similarity measure

Term that contributes to sentence meaning is assigned with 2 dissimilar weights by concept-based statistical analyzer. These 2 weights are united into new weight. Concept-based similarity is used for computing similarity among documents. The concept based similarity method takes full benefit of using concept analysis measures on the corpus, document, and sentence levels in computing the similarity among documents. By using k-means algorithm experiments are done on concept based model on different datasets in text clustering .The experiments are done by comparing the concept-based weight obtained by concept-based model and statistical weight. The results in text clustering show the significant progress of clustering feature using: concept-based term frequency (tf), conceptual term frequency (ctf), concept-based statistical analyzer, and concept-based combined model. In text clustering the results are evaluated using f-measure and entropy.

Keywords
Concept Mining

1. INTRODUCTION
Data Mining is about finding interesting and useful patterns from data. Mining can be done in text, images, videos, and so on. Text Mining [1] is a data mining technique, which attempts to determine new, previously unidentified information by applying methods from the Natural Language Processing. It is different from what are familiar with in web search. Mostly user looks into already existing data which is being written by others. The main problem is pushing all material aside that presently is not related to our needs to discover relevant information. Text mining has different names i.e., Intelligent Analysis, Knowledge-Discovery (KDT) or Data Mining in Text, usually defined as process of mining non-trivial information & interesting and knowledge from unstructured text. As a record of data is warehoused in the procedure of text, text mining is considered to need a high viable potential value. Information may be revealed from different sources of data until now; unstructured texts continue the largest eagerly available sources of knowledge. The task of Knowledge Finding from Text [3] is to mine implicit and explicit concepts and semantic relations among concepts using NLP techniques. The aim of it is mainly to get visions into huge quantities of text data. KDT will play a gradually increasing of significant role in developing applications, those are Text Understanding. Text mining [1] remains identical to data mining, but In data mining the tools[2] are constructed to handle structured data from databanks ,while text mining have a capability to deal with non-structured data For example full-text Documents, emails, html collections etc. As a result of this, text mining is much better solution for the companies.

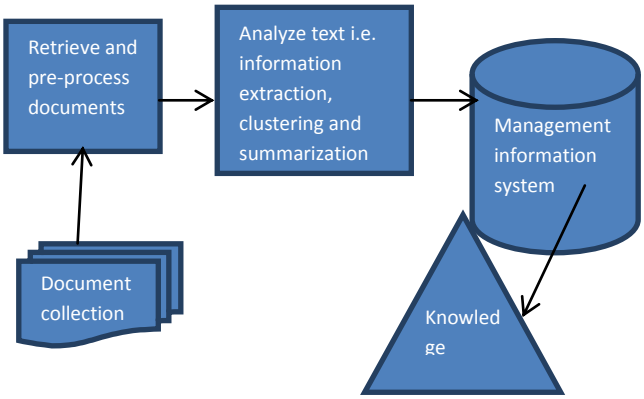


Figure1. An example of text mining process

1.1 Concept Mining [4]
By using concept mining we can extract the concepts embedded in the text document. Concept based search for the useful terms based on their meaning rather than the presence of keyword in text. Concepts are terms or set of terms with some meaning or relation between terms. Sometimes few terms came together very frequently in text that relatedness in terms also lead to one concept. The significance of concept sometimes very different from original meaning of words occurred in the concept.

Ex. - White house.

A concept may have different weight in various corpora/document/sentences.

1.2 Techniques used in Concept Mining:
Term frequency is frequently used as weighting factor in the text mining, depends on frequency of word in corpus, which helps to adjust for the fact that some words appear more frequently in general.

Term Frequency Inverse Document Frequency (TFIDF): The TFIDF technique measures the significance of a term in the document contained by one set of documents.

TFIDF calculated as:

$$tfidf(ti, dj) = tf(ti, dj) idf(ti)$$

Where $tf(ti, dj)$ defines as Term Frequency and is defined by

$$tf(ti, dj) = \begin{cases} \frac{N(ti, dj)}{|d|} if N(ti, dj) > 0 \\ otherwise 0 \end{cases}$$

Where 'ti' is term of document dj, N (ti, dj) denotes the frequency ti in dj, and |dj | is total tokens in document dj. IDF (ti) is termed as Inverse Document Frequency and defined as

$$IDF(ti) = \log\left(\frac{Tr}{df(ti)}\right)$$

Where DF (ti) means Document Frequency of term ti and denotes the no. of documents Tr in which ti happens at least once. Terms that occur in a large number of documents tend to be stop words. By using TF × IDF [6], [7] calculation, it is expected that stop words can be discriminated by their TFIDF score. Keywords are measured important if they have high DF value and high TFIDF value. We say that DF Value is high if it will be larger than the given threshold value. Typically, the TFIDF value is an indicator to identify keywords, and the DF value is an indicator to identify interesting keywords.

Conceptual Term Frequency (CTF):

CTF is considered in both document and sentence. The CTF can be defined as no. of existences of concept in document/sentence level.

1.3 Problems Faced in Concept Mining

The mappings of words to concepts are often ambiguous. Each word that is in a given language will relate to several possible concepts.

1.4 Advantages of Concept Mining

Text mining models are very large when compared to concept mining. The model that is going to classify, for sample, news stories using Naive Bayes or svm's algorithm will be very large, in the megabytes (mb), and it takes much time to load and evaluate. Concept mining models can take less time for comparison - hundreds of bytes. For applications like similar meaning detection, concept mining offers new possibilities.

1.5 Concept Mining Applications

1. Indexing and detecting similar documents in huge corpora.
2. Clustering
3. Classification
4. Natural language translation

2. BACK GROUND AND LITERATURE SURVEY

2.1 Clustering techniques:

In this section, we will give a brief discussion about the common clustering algorithms which are used in experimental study of this thesis work.

Hierarchical techniques [15, 16] yield a nested order of partitions, with one, all inclusive clusters at the highest and singleton groups of distinct points at the lowest. Every intermediate level is viewed as combining 2 clusters from consecutive lower level (or ripping a cluster from consecutive higher level). The results of a class-conscious clump algorithmic program are diagrammatically displayed as tree, known as a dendogram. This tree diagrammatically shows the merging method and also intermediate clusters. For document clump, this dendogram offers a class-conscious index, or taxonomy. The hierarchal clustering consists of two basic approaches.

a. Agglomerative
b. Divisive

The agglomerative hierarchical clustering technique as follows:

Simple Agglomerative Clustering Algorithm

1. It calculates the similarity among all the available clusters, i.e., compute similarity matrix whose $ijth$ record gives the similarity among the ith andjth clusters.
2. Combine the 2 clusters, which are most alike.
3. Updating the similarity matrix to look for a transform in the pairwise similarity among the new cluster and the original clusters.
4. Repeat the steps two and three until only a single cluster remains.

K-means is based on the thought that a middle purpose will represent a cluster. Particularly, for K-means we have tendency to use notion of a centroid that is mean or median purpose of a bunch of points. Note that a centroid virtually ne'er corresponds to Associate in nursing actual information. The fundamental K-means agglomeration technique is conferred below.

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Re-compute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids do not change.

K-Nearest Neighbour Clustering (k-NN) [17, 18]

This algorithm is used in clustering and classification. It uses the property of nearest neighbours k, i.e., an object should be put in same cluster as its nearest k neighbours. The algorithm accepts a user specified threshold, e, on the nearest-neighbour distance. For each new document, the similarity is compared to every other document, and highest k documents are chosen. Accordingly, fresh document is grouped with the group where majority of highest k documents are assigned.

Single Pass Clustering [17,19].

Single pass agglomeration technique additionally expects a similarity matrix as its input and outputs clusters. The agglomeration technique takes every object consecutive and assigns it to the nighest antecedent created cluster, or creates a brand new cluster thereupon object as its 1st member. A brand new cluster is made once the similarity to the nighest group is a smaller volume than a nominal threshold. This threshold is that solely outwardly obligatory parameter. Commonly, the similarity among AN object and a group is set by computing the typical similarity of the item to any or all objects in this cluster.

In[11] the objective of this paper is to perform clustering and concept mining. The author proposed a concept-based mining model. The model consists of sentence, document, Corpus based concept analysis, and concept-based similarity measure. In this model, author tried to say that, it takes raw text document as input and gives clusters as an output .In concept based term analysis, author tried to say that, the main objective of this task is measure concept based term analysis on document and sentence levels. Author used CTF (Conceptual Term Frequency) for analysing concepts at document/sentence level

The similarity among two documents can be calculated by using concept based similarity measure as:

$$\text{Similarity}(d1, d2) = \sum_{i=1}^{m} \max\left(\frac{li}{si1}, \frac{li}{si2}\right) * weighti1 * weighti2$$

Where

Weighti1 = tfweighti1 + ctfweighti1

Weighti2 = tfweighti2 +ctfweighti2

After calculating the similarity measure, concepts in document are sending for clustering. For clustering some techniques are used. They are HAC, SINGLE PASS CLUSTERING, and KNN. And these techniques give an output as set clusters as shown in the figure 2.As compared to traditional investigation of term (word or phrase), the concept based mining results gives fast and better results substantially

In [12] the aim of this paper is to Build/Enrich WorldNet Dictionary and clustering

In NLP Word Net dictionaries are frequently used. Most of the paper uses these dictionaries; but still the enrichment of Word Net dictionaries needs concern. Improvement in this library will definitely help researchers for getting better results. The techniques generally used to improve them are follows automotive strategies of converting text from a source language to another destination language. Most of these techniques follow blindly an unspoken rule: get predefined rules of the language and use it in as dictionary or a simpler parser having grammatical rules. ALOC approach uses to find similar conceptual ideas, and can be useful in Word Net dictionary building as these dictionaries are also stores items by their conceptual meaning. So the ALOC approach is used in this paper, based on conceptual meaning and distance of conceptual same meaning words but described approach was on English language only. And it will be done on other language also.

Limitation of this paper is that, thisapproach was presented on a very restricted domain of samples, which were all in English. Furthermore research should focus on extending the domain of inputs and languages

In [13] the objective of this is to perform concept mining and clustering

3. SUPPORTING THEORIES:

A .TFIDF

The TFIDF measures the significance of term in the document within a bunch of documents.

B. Co-occurring Keywords

A string of one keyword describes a concept. However, a string of more than 1 keyword might describe a new meaning that is beyond the significance of every individual keyword. If this new meaning is important, the set of these keywords will consistently appear in form of co-occurring keywords. Concepts inside a document can encapsulate through high frequency & co-occurring words.

C. Association Rule for Mining Keyword sets

It [8] is used mainly to show relationships between keywords. So Many algorithms have developed for finding association rules. The popular algorithm is "Apriori Algorithm". The algorithm generates n-keyword sets from n− 1- keyword sets, where n > 1. This means that the procedure of generating n-keyword sets depends on the previous step when generating n − 1-keywordsets.

Concepts in a text documents are characterized by graph named as simplicial complex where vertex and edges are keyword and relation between the keywords. This structure named simplex. A simplex with high frequency shows that relation between that keyword is more frequently encountered than others in same text. This relation contains a concept. This simplex relation is figured using Association rule mining. All the simplices of text collectively show the concept structure of text.

A text may contain numeral concepts. This concept may appear in more than one cluster or class so it's not possible to need a separate line between two or more text. In the paper it is described that similar meaning text may have same or nearly same graphical structure of concepts. We can also compare graphical structure of different language to check whether associated text is meaning wise same or not.

Limitation of this paper is that, each document may have several concepts. Therefore, many documents may be intertwined among themselves. Consequently, and cleanly separating documents may not possible always.

In [14] the aim of this paper is web page clustering and concept mining

The author described three important terms that play important role in webpage clustering. They are TF-based, TF-IDF based, TF+filtering. The TF-based, TF-IDF based are explained above [6][7].

TF + filtering: Extract terms with high frequency but filter those words, which appear often on web pages without any special meaning, such as download, link, news etc.

The author summarised webpage clustering with four steps

1) Term extraction (i.e. Feature selection)
2) Term-Document matrix generation
3) SVD –based clustering
4) Merge results if necessary.

Based on clustered results, concept mining consists of three steps: first mining the word sets that frequently appear together on web pages based on Apriori Algorithm [9] then, Generating concepts item sets using the algorithm, finally constructing the concept hierarchy based on their overlap of item sets. Here, concept item set is defined as co-occurring word sets like [10].

Limitation of this paper is that this approach was presented on a very restricted domain of examples, which were all in English and Chinese. Further research should focus on extending the domain of inputs and languages.

4. PROPOSED MODEL

In this model the author have used three clustering algorithms they are K- nearest neighbour algorithm, HAC clustering algorithm and Single pass clustering algorithm. Along with this we have tried with k-means algorithm and it has given results. And we have compared the results with the remaining algorithms. The k-means algorithm has given performance as single pass clustering. From these four algorithms the HAC (ward) was given better results for four datasets

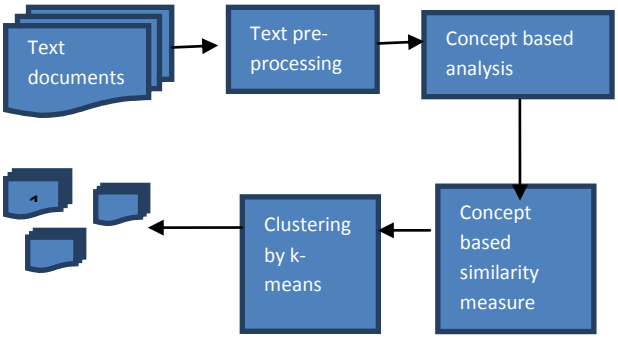


Figure1.Proposed model

Clusters

5. DATASETS AND EXPERIMENTAL RESULTS

This chapter gives the experimental work of applying the concepts extracted by the concept- based model in text clustering. The chapter introduces a comparison among three weighing's w, $weight_{ctf}$ $weight_{tf}$, $weight_{idf}$, and $weight_{combined}$ in text clustering. We discuss the results of concept- based model compared to the traditional techniques. In addition, some observations are reported regarding the text clustering achieved by the concept-based model.

For the single-term weighting, the popular TF-IDF [3] (Term Frequency/Inverse Document Frequency) term weighting is adopted. The TF-IDF weighting is selected due to its wide usage in the document clustering.

The concept-based weighting is one of the main factors that capture the importance of a concept in a document. Thus, to study the effect of the concept-based weighting on the quality of the text clustering the entire sets of the experiments are repeated using the following concept-based weighting schemes:

- The $weight_{tfi}$
- The $weight_{ctfi}$
- The $weight_{combined}$

6. TEXT CLUSTERING

The experimental setup consisted of four datasets. The first data set contains 23,115 ACM abstract articles collected from ACM digital library. The ACM articles are classified according to ACM computing classification system into five main categories: general literature, hardware, computer systems organization, software, and data. The second data set has 12,902 documents from Reuters 21578 dataset. There are 3,299 documents in the test set and 9, 603 documents in training set. Out of the 5 category sets, the topic category set contains 135 categories, but only 90 categories have at least one document in training set. These 90 categories were used in experiment. The third dataset consisted of 361 samples from Brown corpus. Each sample has 2000+ words. The Brown corpus main categories used in experiment were: press: reportage, press: reviews, religion, skills and hobbies, popular lore, belles-letters, learned, fiction: science, fiction: romance, and humour. The fourth dataset consists of 20,000 messages composed from 20 newsgroups. The similarities which are calculated by using concept-based model are used to compute four similarity matrices among documents. Four standard document clustering techniques are chosen for testing effect of concept-based similarity on clustering:

(1) Hierarchical Agglomerative Clustering (HAC), (2) Single Pass Clustering, and (3) k-Nearest Neighbor (k-NN) and (4) k- means clustering

Basically, the aim is to maximize the F-measure, and minimize the Entropy of clusters to achieve high-quality clustering. The ward and the complete linkages were used as cluster distance measures for the HAC method since they tend to produce tight clusters with small diameter. A document-to-cluster similarity threshold of 0.3 was used in the single-pass clustering method. A k of 5 and a cluster similarity threshold of 0.35 were used in k-NN method. The parameters chosen for the different algorithms were the ones that produced best results. For the tf weighting, the percentage of improvement ranges from +6:31 to +67:94 percent increase in the F-measure quality, and -29:56 to -64:67 percent drop in Entropy (lower is better for Entropy). For the ctf weighting, the percentage of improvement ranges from +23:37 percent to (+25.64 above 100 percent) increase in the F-measure quality, and -62:65 to -82:96 percent drop in Entropy. For the ctf and tf weighting, the percentage of improvement ranges from +27:93 percent to (+47:75 above 100 percent) increases in the F-measure quality, and -62:7 to -95:68 percent drop in Entropy. For the ctf, tf, and df combined weighting, the percentage of improvement ranges from +29:04 percent to (+63:14 above 100 percent) increase in the F-measure quality, and -85:35 to -97:25 percent drop in Entropy. It is shown that the HAC clustering with the ward linkage has the best performance. Moreover, the HAC clustering with the complete linkage performance is close to the k-NN clustering performance. It is known that Single-Pass clustering is very sensitive to noise; that is why it has the worst performance. However, when the concept based similarity was introduced using the combined weighting scheme among the tf, ctf, and df, the quality of clusters produced was pushed close to that produced by HAC and k-NN

A k of 5 is used in k-means clustering algorithm. But this algorithm hasn't given the results to up to mark. In particular, the parameter k is known to be hard to choose (as discussed above) when not given by external constraints. Another

drawback of the algorithm is that it cannot be used with arbitrary distance functions or on non-numerical data. For these use cases, many other algorithms have been developed since.It is observed that the standard deviation is improved by using the concept-based mining model. It is illustrated that the ctf weighting scheme is more accurate than the tf weighting when it comes to calculate the concept-based relations between documents. The grouping between the tf, ctf, and df coefficient schemes will accurately live the importance of an concept at the sentence, document, and corpus levels that results in enhance the agglomeration quality well

Clustering Improvement using Conceptual Term Frequency (weight tf)

Dataset	Clustering techniques	F-measure	Entropy	F-measure	Entropy	Improvements
Reuters	Hac(Ward)	0.723	0.251	0.782	0.113	+8.16%F,-54.98%E
	Hac(Complete)	0.623	0.315	0.766	0.121	+22.95%F,-61.58%E
	Single pass	0.411	0.523	0.623	0.279	+51.58%F,-46.65%E
	k-NN	0.511	0.348	0.736	0.174	+44.03%F,-50%E
	k-means	0.696	0.238	0.762	0.136	+9.48%F,-56.90%E
ACM	Hac(Ward)	0.697	0.317	0.688	0.136	+6.31%F,-43.84%E
	Hac(Complete)	0.481	0.362	0.653	0.211	+48.02%F,-48.06%E
	Single pass	0.398	0.608	0.611	0.263	+52.52%F,-48.02%E
	k-NN	0.491	0.402	0.657	0.149	+41.95%F,-41.54%E
	k-means	0.372	0.321	0.617	0.138	+65.86%F, 57%E
Brown	Hac(Ward)	0.581	0.385	0.688	0.136	+18.41%F,-64.67%E
	Hac(Complete)	0.547	0.401	0.653	0.211	+19.37%F,-47.38%E
	Single pass	0.437	0.551	0.611	0.263	+39.81%F,-52.26%E
	k-NN	0.462	0.316	0.657	0.149	+42.20%F,-52.84%E
	k-means	0.444	0.238	0.608	0.133	+36.93%F, 44.12%E
20 newsgroups	Hac(Ward)	0.535	0.316	0.661	0.152	+23.55%F,-51.89%E
	Hac(Complete)	0.471	0.345	0.643	0.243	+36.51%F,-29.56%E
	Single pass	0.312	0.643	0.524	0.371	+67.94%F,-42.30%E
	k-NN	0.462	0.457	0.621	0.256	+34.41%F,-43.98%E
	k-means	0.324	0.236	0.532	0.138	+64.197, 41.52%E

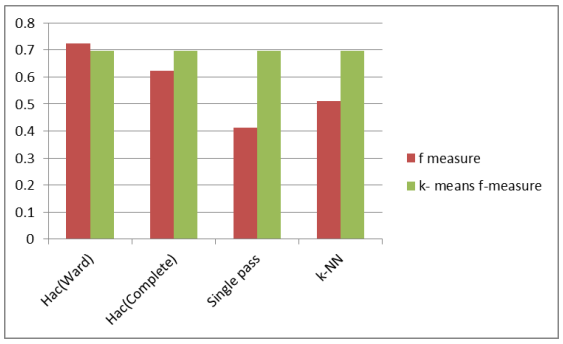
Clustering Improvement using Conceptual Term Frequency (weight ctf)

Dataset	Clustering techniques	F-measure	Entropy	F-measure	Entropy	Improvements
Reuters	Hac(Ward)	0.723	0.251	0.892	0.062	+23.37%F,-75.29%E
	Hac(Complete)	0.623	0.315	0.817	0.068	+31.13%F,-78.41%E
	Single pass	0.411	0.523	0.783	0.125	+90.51%F,-76.09%E
	k-NN	0.511	0.348	0.861	0.066	+68.49%F,-81.03%E
	k-means	0.423	0.467	0.797	0.136	+88.41%F,-70.8%E
ACM	Hac(Ward)	0.697	0.317	0.883	0.054	+26.68%F,-82.96%E
	Hac(Complete)	0.481	0.362	0.843	0.173	+75.25%F,-52.20%E
	Single pass	0.398	0.608	0.764	0.201	+91.95%F,-66.94%E
	k-NN	0.491	0.402	0.845	0.143	+72.09%F,-64.42%E
	k-means	0.475	0.562	0.823	0.183	+73.26%F,-67.43%E
Brown	Hac(Ward)	0.581	0.385	0.876	0.083	+50.77%F,-78.44%E
	Hac(Complete)	0.547	0.401	0.852	0.127	+55.75%F,-68.32%E
	Single pass	0.437	0.551	0.726	0.133	+66.13%F,-75.86%E
	k-NN	0.462	0.316	0.843	0.118	+82.46%F,-62.65%E
	k-means	0.452	0.454	0.738	0.133	+63.26%F,-70.70%E
20 newsgroups	Hac(Ward)	0.535	0.316	0.852	0.064	+59.25%F,-79.74%E
	Hac(Complete)	0.471	0.345	0.823	0.082	+74.73%F,-76.23%E
	Single pass	0.312	0.643	0.704	0.153	+100%F,-76.20%E
	k-NN	0.462	0.457	0.815	0.121	+76.40%F,-73.52%E
	k-means	0.368	0.572	0.762	0.138	+100%F,-75.87%E

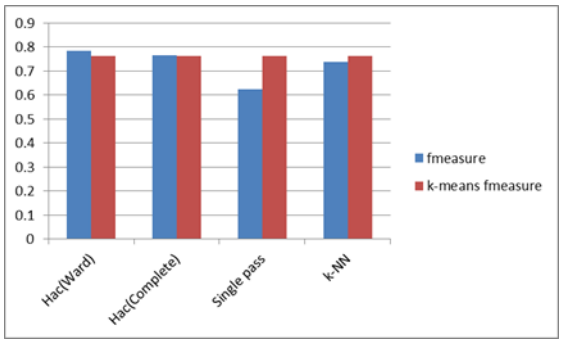
Clustering Improvement using Conceptual Term Frequency (weight combined)

Dataset	Clustering techniques	F-measure	Entropy	F-measure	Entropy	Improvements
Reuters	Hac(Ward)	0.723	0.251	0.925	0.012	+27.93%F,-95.21%E
	Hac(Complete)	0.623	0.315	0.907	0.025	+45.58%F,-92.06%E
	Single pass	0.411	0.523	0.816	0.067	+98.54%F,-87.18%E
	k-NN	0.511	0.348	0.917	0.015	+79.45%F,-95.68%E
	k-means	0.442	0.483	0.823	0.064	+86.20%F,-86.75%E
ACM	Hac(Ward)	0.697	0.317	0.918	0.043	+31.70%F,-86.43%E
	Hac(Complete)	0.481	0.362	0.895	0.135	+86.07%F,-62.7%E
	Single pass	0.398	0.608	0.791	0.152	+98.74%F,-75.00%E
	k-NN	0.491	0.402	0.891	0.111	+81.46%F,-72.38%E
	k-means	0.421	0.572	0.807	0.138	+91.68%F,-75.87%E
Brown	Hac(Ward)	0.581	0.385	0.906	0.018	+55.93%F,-95.32%E
	Hac(Complete)	0.547	0.401	0.901	0.021	+64.71%F,-94.76%E
	Single pass	0.437	0.551	0.804	0.045	+83.98%F,-91.83%E
	k-NN	0.462	0.316	0.902	0.023	+95.23%F,-92.72%E
	k-means	0.442	0.472	0.806	0.039	+82.35%F,-91.73%E
20 newsgroups	Hac(Ward)	0.535	0.316	0.901	0.035	+68.41%F,-88.92%E
	Hac(Complete)	0.471	0.345	0.892	0.074	+89.38%F,-78.55%E
	Single pass	0.312	0.643	0.773	0.087	+100%F,-86.46%E
	k-NN	0.462	0.457	0.865	0.065	+34.41%F,-43.98%E
	k-means	0.330	0.551	0.784	0.072	+100%F,-86.93%E

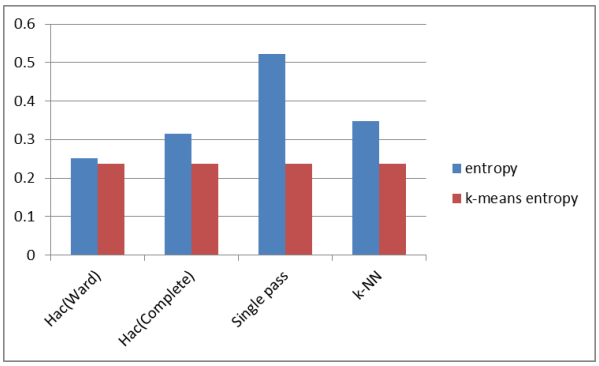
F-measure for single term analysis in Reuters dataset by using (weight tf)



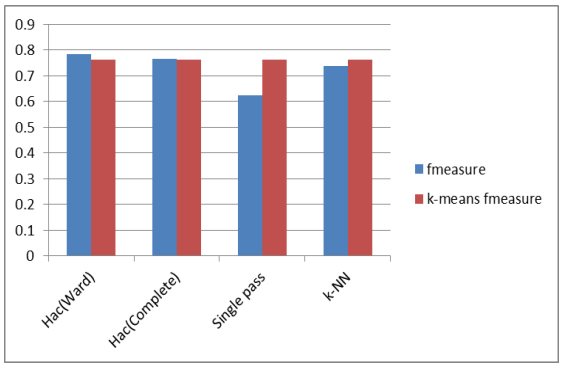
F-measure for concept based analysis in Reuters dataset by using (weight tf)



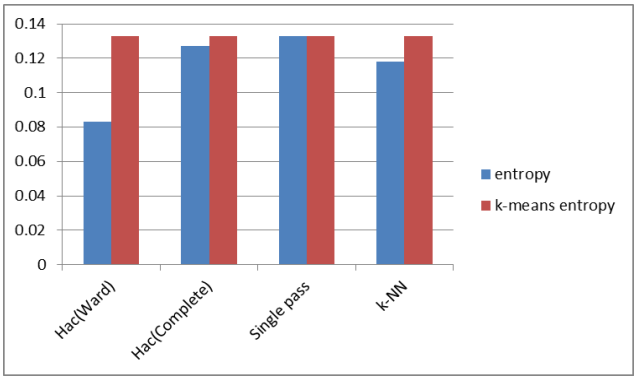
Entropy for single term analysis in Reuters dataset by using (weight tf)

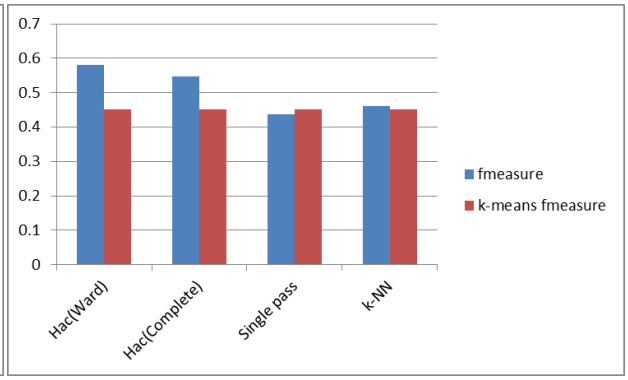


Entropy for concept based analysis in Reutersdataset by using (weight tf)

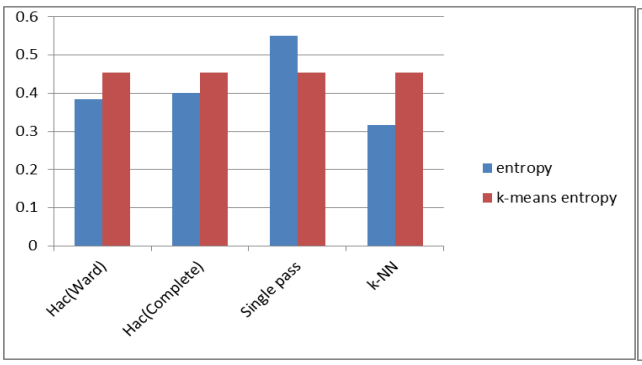


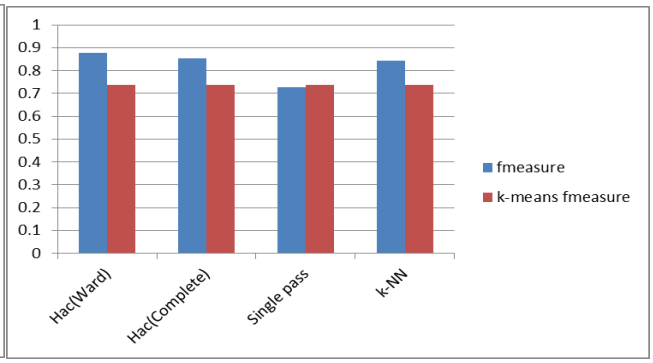
F-measure for single term analysis in Brown dataset Entropy for concept based analysis in Brown dataset by using (weight ctf) by using (weight ctf)



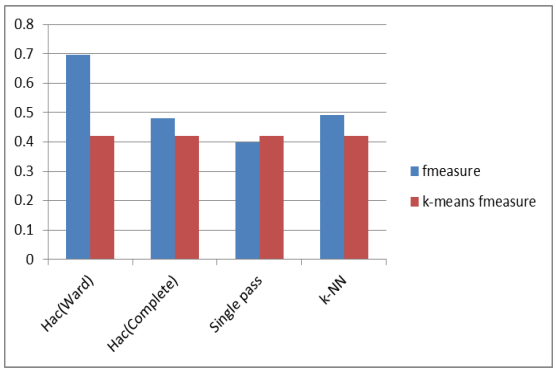


Entropy for single term analysis in Brown dataset F-measure for concept based analysis in Brown by using (weight ctf) dataset by using (weight ctf)

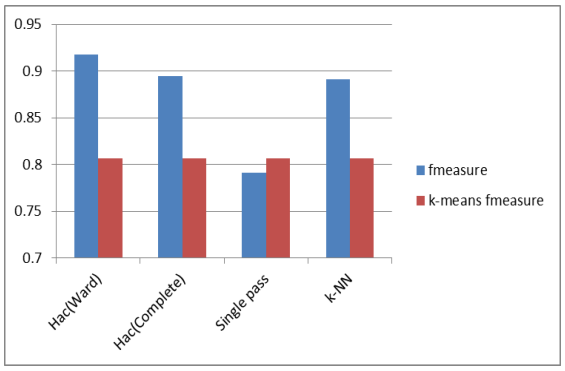




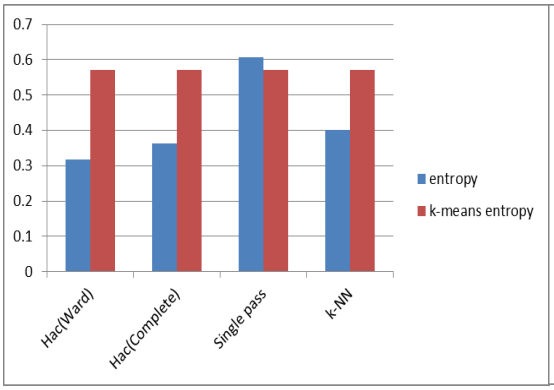
F-measure for single term analysis in ACM dataset



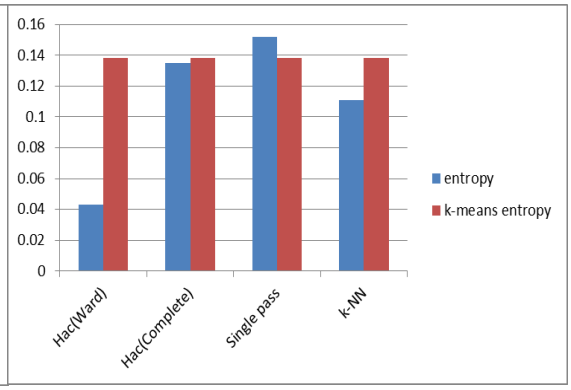
F-measure for concept based analysis in ACM by using weight combined dataset by using weight combined



Entropy for single term analysis in ACM dataset



Entropy for concept based analysis in ACM using(weight combined)dataset by using (weight combined)



7. CONCLUSION

Most of recent text mining systems consider only the presence or the absence of keywords in text. Statistical analysis of word frequencies is not sufficient for representing the meaning of text. Concept-based mining is the area targeting the semantics of text rather than word frequencies. The main role of this work lies in the concept-based model which extracts and signifies the semantics in text based on concepts. The concept-based model discovers structured knowledge to be utilized in several applications. The new concept-based model is introduced to improve the text clustering qualities. By using the semantic structure of sentences in documents, a well text clustering results are achieved. The new concept-based model consists of two components. The 1st component is concept-based statistical analyzer that analyses semantic structure of every sentence to extractthe sentence concepts using *ctf* measure. The second module is the concept based similarity measure which is capable of performs an accurate calculation of pair-wise documents. The quality of the text clustering results achieved by the concept based model surpasses that of traditional weighting approaches significantly.

We have tried with k-means algorithm for clustering the documents after concept based similarity measure. But it hasn't given the best results when compared with remaining algorithms. We have experimented with four data sets namely, Reuter's dataset, Brown dataset, ACM dataset and 20 news groups. We have done concept based model in three steps, they are text pre-processing, concept based analysis, and concept based similarity measure and clustering using k-means algorithm

There are 2 drawbacks for concept-based model. One limitation is that an English sentence has to be grammatically correct. If an English sentence is not grammatically correct, then the semantic role labeler may not have an output. Another limitation is that if an English sentence is extremely short. For example, if a sentence contains only three words. i.e." Mike plays football". Then the subject, verb, and the object will have the same weighting. Though, this is never happen in regular documents as sentences tend to be much longer.

8. FUTURE WORK

There are a number of suggestions to extend this work. One direction is to link the presented work to web document clustering, categorization, and retrieval. The intention is to apply the same approach on web documents. In this case, a text extraction process is required before applying the proposed approach to web documents. This direction can open new horizons in many business applications. For example, clustering stock markets of the user's we web blogs based on the sentence meaning can lead to know which stock to buy. Another future direction is to investigate the usage of WordNet to extract the synonyms, hypernyms, and hyponyms and their effect on document clustering, categorization, and retrieval results, compared to that of traditional methods. In this case, the model analyzes terms and their corresponding synonyms and/or hypernyms on the sentence and document levels. Thus, if two documents contain different words and these words are semantically related, the model can measure the semantic-based similarity between the two documents.

9. REFERENCES

[1] [Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43

[2] Navathe, Shamkant B., and ElmasriRamez, (2000), "Data Warehousing and Data Mining", in "Fundamentals of Database Systems", Pearson Education pvtInc, singapore, 841-872.

[3] HaralamposKaranikas and BabisTheodoulidis Manchester, (2001), "Knowledge Discovery in Text and Text Mining Software", Centre for Research in Information Management, UK

[4] https://en.wikipedia.org/wiki/Concept_mining

[5] P. Kingsbury and M. Palmer, "Propbank: The Next Level of Treebank," Proc. Workshop Treebanks and Lexical Theories, 2003.

[6] G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol5, 513-52

[7] G. Salton and C. Buckley. Term Weighting Approaches in AutomaticText Retrieval, 1960, Information Processing and Management, 24, Vol 5, 513-523

[8] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proceedings of the 20th VLDB Conference, 1994

[9] Agrawal R, Imielinski T, Swami A, "Mining association rules between sets of items in large databases". Proc of the 1993ACM SIGMODInternational Conference on Management of data

[10] Bing Liu, Yiming Ma, "Discovering unexpected information from your competitors 'Web Sites in Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, USA.

[11] An Efficient Concept-Based Mining Model for Enhancing Text Clustering, Shady Shehata, Member, IEEE, FakhriKarray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE 2010

[12] Concept mining from natural language texts, Rockai V. Dept. of Cyber. & Artificial Intelligent, Tech. Univ. of Kosice, Kosice, Slovakia Mach. M IEEE 2012

[13] Concept Mining using Association Rules and Combinatorial Topology Sutojo, A, San Jose State University, San Jose IEEE 2007

[14] Webpage Clustering and Concept Mining, an Approach to Intelligent Information Retrieval. Fang Li, Martin Mehlitz, Li Feng, Huanye Sheng, DEPT of CSE, Shanghai Jiaotong University, Shanghai ,China IEEE 2006

[15] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs: Prentice Hall, 1988.

[16] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, ser. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons Inc., 1990.

[17] K. J. Cios, W. Pedrycz, and R. W. Swiniarski, Data mining methods for knowledge discovery," IEEE Transactions on Neural Networks, vol. 9, no. 6, pp. 1533{1534, 1998.

[18] B. V. Dasarathy, Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, 1991.

[19] D. R. Hill, A vector clustering technique, "in FID-IFIP, Samuelson(ed), N-H 1968, 1967.

[20] A survey paper on Concept Mining in Text documents K.n.s.s.v.prasad, ,Dr.S.K.Saritha, ,Dixa saxena. International Journal of Computer Applications (0975 – 8887)Volume 166 – No.11, May 2017