ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Design and Development of Novel Sentence Clustering Technique for Text Mining

Rakhi S.Waghmare, Prof. Ram Mangrulkar, Prof.Vaishali Bhujade

Student, Dept. of CSE, BDCOE, RTMNU University, Wardha, India

HOD, Dept. of CSE, BDCOE, RTMNU University, Wardha, India

Assistant Professor, Dept. of CSE, BDCOE, RTMNU University, Wardha, India

ABSTRACT: Clustering is the process of grouping or aggregating of data items. The sentence clustering is used in variety of applications i.e. classify and categorization of documents, automatic summary generation, etc. In text mining, the sentence clustering plays a vital role this is used in text activities. Size of clusters can change from one cluster to another. The traditional clustering algorithms have some problems in clustering the input dataset. The problems such as, instability of clusters, complexity and sensitivity. To overcome the drawbacks of these clustering algorithms. this paper proposes a algorithm called Hierarchical Bisecting K Means Clustering Algorithm (HBKMCA) is extension of FRCA which is used for the clustering of sentences. Contents present in text documents contain hierarchical structure and there are many terms present in the documents which are related to more than one theme hence HBKMCA will be useful algorithm for natural language documents for finding the action word with the help of RiWordNet dictionary And the using the page rank algorithm and EM method to finding the EM value. In this algorithm single object may belong to more than one cluster.

KEYWORDS: Text mining, natural language processing, HBKMCA Algorithm, Sentence Clustering.

I. INTRODUCTION

Clustering techniques can help in this data discovery and data analysis. Clustering the sentences is mainly useful in Information Retrieval (IR) Process. Clustering text at the sentence level and document level has many differences. Document clustering partitions the documents into several parts and cluster those parts based on the overall theme. It doesn't give much importance to the semantics of each sentence in the document. So there may be content overlap or bad coverage of theme will happen in the case of multi document summarization. Each data element in hard clustering method belongs to exactly one cluster.

Sentence clustering plays an important role in many text processing activities. For example, various authors have argued that incorporating sentence clustering into extractive multi document summarization helps avoid problems of content overlap, leading to better coverage. However, sentence clustering can also be used with in more general text mining tasks. For example, consider web some novel information from a set of documents initially retrieved in response to some query. Clustering algorithms are used in many Natural Language Processing (NLP) tasks to grouping words and documents to entire languages. They are popular tools to use to group similar items together. Irrespective of the specific task (e.g., summarization, text mining, etc.), most documents will contain interrelated topics or themes, and many sentences will be related to some degree to a number of these .We now highlight some important differences between clustering at these two levels, and examine some existing approaches to fuzzy clustering. Clustering is an unsupervised method to divide data.

Since the goal of this methods is to analysis text data with different methods. Our main aim is to generate new algorithm to get the more accuracy with compare to previous algorithm and to use the new preprocessing i.e. NLP preprocessing to find the action word in given text and to reducing the time for analyzing the document.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

II. LITERATURE REVIEW

Andrew Skabar, and Khaled Abdalgader was proposed a novel fuzzy clustering algorithm that operates on relational input data; i.e., data in the form of a square matrix of pairwise similarities between data objects. This algorithm uses a graph representation of the data, and operates in an Expectation Maximization framework in which the graph centrality of an object in the graph is interpreted as alikelihood. [1]. In this paper, J.Dafni Rose, Divya D. Dev, C.R.Rene Robin was proposed the MLCL algorithm.The MKCL Algorithm works in three phases. Firstly, the linked keywords of the genetic based extraction method are identified with a Must Link and Cannot Link algorithm (MLCL). Secondly, the MLCL algorithm forms the initial clusters. parameters. Finally, the clusters are optimized using Gaussian The proposed method is tested with datasets like Reuters-21578 and Brown Corpus [3].

Mario G.C.A. Cimino, Beatrice Lazzerini, Francesco Marcelloni was proposed the TS system to build a dissimilarity matrix which is fed as input to an unsupervised fuzzy relational clustering algorithm, denoted any relation clustering algorithm (ARCA), which partitions the data set based on the proximity of the vectors containing the dissimilarity values between each pattern and all the other patterns in the data set. [2]. Christy Maria Joy1, S. Leela was proposed the Hierarchical Fuzzy Relational Clustering Algorithm and it is a hybrid method which is a combination of general hierarchical clustering concepts with fuzzy relation models that is existing fuzzy clustering algorithms.This method used Cosine similarity.[4]. This paper Vasileios Hatzivassiloglou, Judith L. Klavans was present a statistical similarity measuring and clustering tool, SIMFINDER , that organizes small pieces of [5].

Erkan was introduce a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. LexRank, for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences.[6]. Ronan Collobert and Jason Weston was describe a single convolutional neural network architecture that, given a sentence, out- puts a host of language processing predictions: part-of-speech tags, chunks, named entity tags, seman- tic roles, semantically similar words and the likelihood that the sentence makes sense (grammatically and semantically) using a language model[7].Rekha Jain and Dr G. N. Purohit was discuss the commonly used algo- PageRank, Weighted PageRank and HITS. This paper PageRank and Weighted PageRank algorithms calculates the score at indexing time and sort them according to importance of page where as HITS calculates the hub and authority score of n highly relevant pages[8]. Benjamin C.M. Fung, Ke Wang and Martin Ester were proposing the notion of frequent itemsets, which comes from association rule mining, for document clustering. Frequent itemsets are also used to produce a hierarchical topic tree for clusters. By focusing on frequent items, the dimensionality of the document set is drastically reduced [9]. This paper Saranya.J, Arunpriya.C was survey the different clustering algorithm. The main goal of this survey is to present an overview of the sentence level clustering techniques.. We can obtain the more efficient method or we may propose the new technique to overcome the problems in these existing approaches.

III. PROPOSED ALGORITHM

In this work, we analyze how one can take advantage of the efficiency and stability of clusters, when the data to be clustered are available in the form of similarity relationships between pairs of objects. More precisely, we propose a new Hierarchical bisecting k means clustering algorithm, based on the existing fuzzy C-means (FCM) algorithm, which does not require any restriction on the relation matrix. This HBKMCA algorithm is applied for the clustering of the text data which is present in the form of text files. HBKMCA will give the output as clusters which are grouped from text data which is present in a given documents.

In this HBKMCA algorithm, NLP per processing technique to use to finding the action word with help of Riwordnet dictionary and then sorting the data. Next we used the PageRank algorithm to calculating the similarity measure. EM method to calculate the EM value for each sentence of text .Last we apply the Hierarchical bisecting k means clustering algorithm will give the output as clusters which are grouped from text data which is present in a given documents.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

A. Natural Language Processing

Natural language processing (NLP) is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI).The NLP is mainly used is to finding the action word in text document. Common NLP tasks in software programs today include:

1. Sentence segmentation, part-of-speech tagging and parsing..
2. Named entity extraction.

The availability of standard benchmarks has stimulated research in Natural Language Processing (NLP). Part-Of-Speech tagging (POS) and chunking (CHUNK) is two NLP standard task processor.

1 Part-Of-Speech Tagging

POS aims at labeling each word with a unique tag that indicates its syntactic role, for example, plural noun, adverb,. A standard benchmark setup is described in detail by Toutanova et al. (2003. The best POS classifiers are based on classifiers trained on windows of text, which are then fed to a bidirectional decoding algorithm during inference.

2 Chunking

Chunking also called shallow parsing, chunking aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrases (NP or VP). Each word is assigned only one unique tag, often encoded as a begin-chunk (e.g., B-NP) or inside-chunk tag (e.g., I-NP).

B. Page Rank Algorithm

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element E is referred to as the PageRank of E and denoted by. PageRank can be calculated for collections of documents of any size. PageRank values to more closely reflect the theoretical true value. The original PageRank algorithm was described by Lawrence Page and Sergey Brin in several publications. It is given by

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

Where,

PR(A) is the PageRank of page A,

PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1.But usually used the o.85 damping factor value.

We describe the application of the algorithm to data sets, and show that our algorithm performs better than other fuzzy clustering algorithms. Proposed algorithm, we describe the use of Page Rank. In the Page Rank is used as a graph centrality measure.

C. Expectation Maximization

An expectation-maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. The probabilities calculated from E-step are restimated with the parameters in M-step.

D. Hierarchical Bisecting K Means Clustering Algorithm

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. We used in proposed methodology hierarchical bisecting k means clustering algorithm, it the divisive clustering algorithm to start the one cluster and get the output in different cluster.

The general idea of the Hierarchical Bisecting K Means clustering is the partitioning of the data items into a collection of clusters. The data points are Many existing clustering techniques have difficulties in handling extreme outliers .This algorithm is an extension of fuzzy relational clustering algorithm. An expectation maximization (EM) algorithm is an iterative process, in which the model mainly depends on some unobserved latent/hidden variables. This algorithm is particularly used in finding maximum likelihood estimates of parameters.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

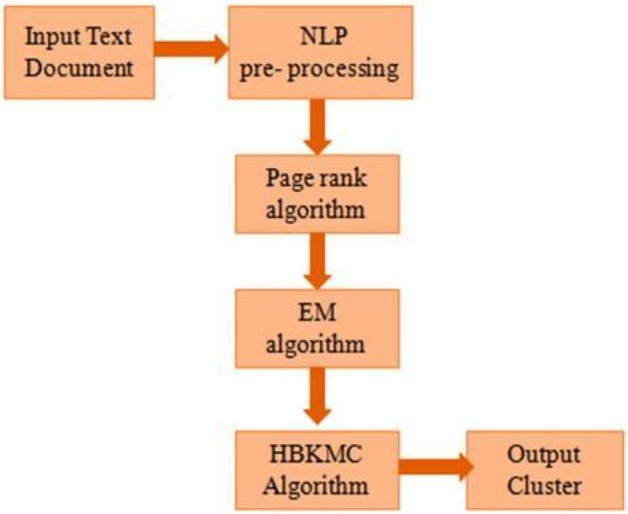

Fig. Flow of HBKMCA Algorithm

In the above figure Text Files are given as the input to the propose algorithm. The NLP are finding the action word by using the RiWordnet Directory .The similarity measure will be calculated with the help of PageRank algorithm. It will calculate the sentence related to number of themes. The EM method is calculating the EM value for each sentence. Then HBKMCA algorithm is applied on it. It will generate different set of clusters which contain sentences in it. Finally, comparison is made between the two algorithms on the basis of effectiveness of result, time complexity, and space complexity. From result, efficient algorithm can be identified and result obtained in this module will be useful.

IV. SIMULATION RESULTS

The Cluster evaluation may be either supervised, in which case external information (usually known class labels associated with the instances) is used to measure the goodness of the clustering; or unsupervised, in which case no external information is used. In the following, L w1; w2 ; . . . is the set of clusters, C c1 ; c2 ; . . . is the set of classes (for supervised evaluation), and N is the number of objects.

4.1 Purity:

The fraction of the cluster size that the largest class of object assigned to that cluster. Two widely used external clustering evaluation criteria are purity and entropy [18]. the purity of cluster j is.

$$Purity=1N\sum i=1kmaxj|ci\cap tj|$$

Where, N is number of objects(data points),k is number of clusters & ci is a cluster in C and tj is the classication which has the max count for cluster ci.

4.2 Entropy

The entropy score denotes the *lower bound* on the number of bits required to represent each value in the data stream. Combining these operations into one formula produces the formula for computing the entropy (H) of a discrete random variable:

$$H(P) = -\Sigma_{i:n}\, p(s_i) * \log(p(s_i))$$

Where, $P(s_i)$ is the probability of class i in P

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

4.3 Rand index:

The Rand index computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. One can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. But before using this formula we calculate the contingency table and the table is,

	Same cluster	Different clusters
Same class	TP	FN
Different classes	FP	TN

Using above table it can be computed the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

Where,

TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

4.4 F Measure:

The F-measure can be used to balance the contribution of false negatives by weighting recall . Let precision and recall be defined as follows

$$P = \frac{TP}{TP+FN}$$
$$R = \frac{TP}{TP+FN}$$
$$F = \frac{2*PR}{P+R}$$

where ,P is the precision rate ,R is the recall rate and F is the F measure.

	TIME (msec)			
Datasets	NLP	PRA	EMA	HBKMCA
D1	1147	23058	22909	22748
D2	1362	1840	1800	1900
D3	1374	610	500	560
D4	1312	79714	80312	79580
D5	1392	32997	32982	32947
D6	1562	87813	87939	88442
D7	1292	16803	41191	41201
D8	1252	15520	15920	15550

Fig 4.1 Time Analysis with Different Datasets.

`In the above table show the different text data set.of text dacuments.We analysis the different text document time requried to each techniques.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

Technique	Purity	Entropy	Rand index	Fmeasure
HBKMCA	0.816	0.256	0.870	0.822
FRECCA	0.800	0.324	0.862	0.601
ARCA	0.622	0.451	0.815	0.462
Spec.cluster	0.690	0.475	0.800	0.444
k-medoids	0.720	0.457	0.779	0.459

Fig 4.2 Clustering Evaluations in Different parameter

The various technique different ratio of the evaluation text data. The above table show the comparison for HBKMCA technique evaluation and previous technique evaluation result.

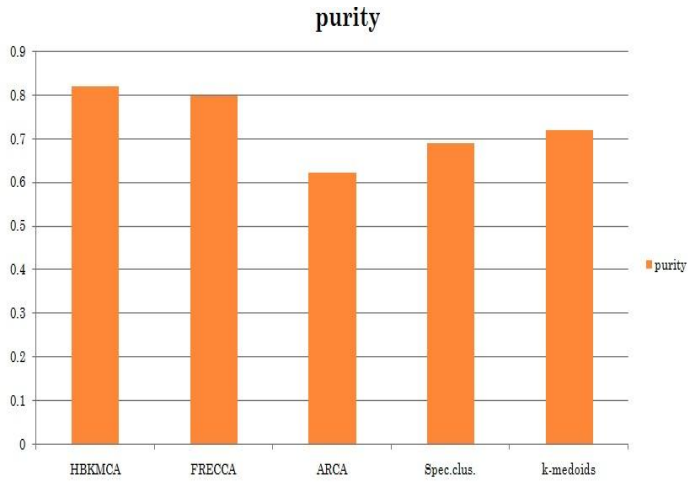

Fig 4.3 Purity Comparisons Graph of Different Clustering Algorithm

In above graph show the analysis of purity factor for HBKMCA technique and pervious technique. And show the purity of HBKMCA technique is increase .

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

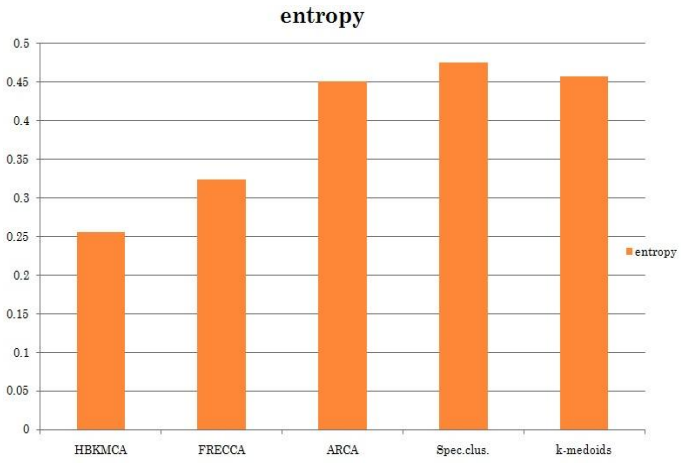

Fig 4.4 Entropy Comparison Graph of Different clustering Algorithm

In above graph show the analysis of entropy factor for HBKMCA technique and pervious technique. And show the entropy of HBKMCA technique is decrease .

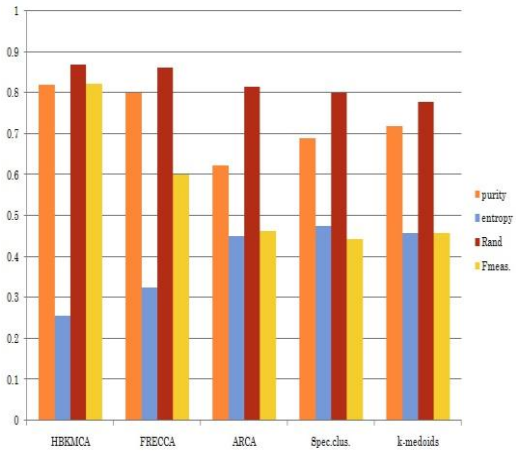

Fig 4.5 Performance Evaluation Graph of Different Clustering Algorithm

In above graph show the performance evaluation for HBKMCA technique and pervious clustering algorithm. And show the comparison graphs of HBKMCA algorithm and previous clustering algorithm.

V. CONCLUSION AND FUTURE WORK

The HBKMCA algorithm was motivated by oue interest in Hierarchical clustering of sentence level text, and the need for an algorithm which can accomplish this task based on text input data. the result we have presented show that the algorithm is able to achieve better performance of clustering than FRECCA, ARCA,ect. The Hierarchical clustering algorithm give the solution for Fuzzy clustering algorithm such as time accuracy and overlapping the cluster data. We analyzed how algorithm combine different results in order to obtain the stable clusters. HBKMCA also can work with any relational clustering algorithm rather than BKMC algorithm not. The algorithm also be used with general text mining applications. In the case of sentence clustering, first we collect different dataset in text document and apply the NLP pre-processing for finding action word in text document and then applying the page rank and EM algorithm to

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 9, September 2015

calculate the similarity value of each sentence. The HBKMCA cluster to reduce the overlapping of sentence and missing data in cluster output.

REFERENCES

1. Andrew Skabar,Member,IEEE, and Khaled Abdalgader ,"Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm",IEEE Trans. Knowledge and Data Eng., vol. 25, no. 8,pp. 1138-1150,No1,January 2013.
2. Richard Khoury ,"Sentence Clustering Using Parts-of-Speech" I.J. Information Engineering and Electronic Business, 2012, 1, 1-9.
3. KamalSarkar.",SentenceClustering-based Summarization of Multiple Text Documents", International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, July 2009.
4. N. Duhan, A.K. Sharma and K.K. Bhatia, Page Ranking Algorithms: A Survey,Proceedings of the IEEE International Conference on Advance Computing, 2009.
5. Ronan Collobert,Jason Weston," A Unied Architecture for Natural Language Processing:Deep Neural Networks with Multitask Learning", NEC Labs America, 4 Independence Way, Princeton, NJ 08540 USA
6. J.Durga, D.Sunitha, S.P.Narasimha, B.Tejeswini Sunand "A Survey on Concept Based Mining Model using Various Clustering Techniques" International Journal of Advanced Research in Computer Science and Software Engineering 2012.
7. Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceeding of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
8. Erkan and D.R. Radev, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization,"J. Articial Intelligence Research, vol.22, pp. 457-479, 2004.
9. Rekha Jain, Dr. G. N. Purohit "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications (0975 - 8887), Volume 13 No.5, January 2011.
10. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceeding of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.
11. Ruchika R. Patil and Amreen Khan,"Bisecting K-Means for Clustering Web Log data", International Journal of Computer Applications (0975 - 8887) Volume 116 - No. 19, April 2015.

BIOGRAPHY

Rakhi Waghmare is a Mtech student in the Computer Science &Engg. Department, College of BDCOE Sewagram, RTMNU,University. India. Her research interests are Data Mining , HCI, Algorithms,etc. Her completed the bachelor degree in RTM Ngpupur university.