A New Approach to Automated Summarization based on Fuzzy Clustering and Particle Swarm Optimization

Anshita
PDM College of Engineering,
Bahadurgarh, Haryana

Rahul Kumar Yadav
PDM College of Engineering,
Bahadurgarh,Haryana

Sugandha Singh, PhD
PDM College of Engineering,
Bahadurgarh,Haryana

ABSTRACT

Automated Summarization of the text is now become an important aspect as it makes the meaning of documents easy to understand and easy to read. Automated summarization is the process of decreasing a text document with a computer system to be able to develop a synopsis that retains the main points associated with document this is certainly initial. Once the irritating dilemma of information overload is continuing to grow, and as the total amount of data has increased, so has fascination with automated summarization. A typical example of the application of summarization technology such as for example Bing and Document summarization is another. There are number of clustering algorithms which have been used in the past as clustering plays significant role in summarizing of the documents. In this paper, we discussed about the existing clustering algorithms. We also proposed a hybridized algorithm based on the combination of fuzzy C-Means and Particle Swarm Optimization. In the last, we compared our proposed algorithm results with the existing clustering algorithms.

Keywords

Hybridized, Clustering, Particle Swarm Optimization , Fuzzy C-Means

1. INTRODUCTION

Developing data mining algorithms for streaming and text data have emerged as an important problem. For streaming data, the assumption is that the data records can be examined only once. One of the main applications is to cluster large amounts of distributed network monitoring data. Simply integrating the distributed data in real time is expensive and for this reason low cost clustering algorithms for streaming data are of more interest. Clustering is a data mining technique used to group a set of objects into clusters, with the purpose of low intra-cluster distances and high inter-cluster distances. Application of clustering algorithms include fraud detection in the telecommunications industry, IDS, Web document clustering, Wireless sensor networks, Web Mining, Text Mining, Information Retrieval etc.

Document Clustering is the grouping of relevant objects is a difficult task in mining owing to the high-dimensionality and sparse nature of text documents. It requires efficient algorithms which can address this high dimensional clustering problem. This plays a crucial role in web based applications and text data mining.

There are different clustering algorithms which are used for clustering purpose which has been discussed in this paper like K-Means, Fuzzy C-Means, particle Swarm Optimization. These algorithms play significant role in summarizing the document. Summarization of the document is important from the users point of view as it enhances the understanding of the user , saves time , provides more accuracy etc.

Further this paper is divided in different section as , in section II the work which has been done till now is discussed , in section III different existing clustering algorithms are discussed and in section IV proposed algorithm is discussed.

In section V evaluation is done and on the basis of evaluation comparison is provided in section VI. Section VII concludes the conclusion and in section VIII future work is discussed.

2. LITERATURE SURVEY

In this paper, the impact of using phrases in the vector space model for clustering documents in Swedish has investigated in different ways. The investigation is carried out on two text sets from different domains: one set of newspaper articles and one set of medical papers. The use of phrases do not improve results relative the ordinary use of words. The results differ significantly between the text types. This indicates that one could benefit from different text representations for different domains although a fundamentally different approach probably would be needed [1].

In this paper, the results of an experimental study of some common document clustering techniques: agglomerative hierarchical clustering and K-means is presented. (In this both a “standard” K-means algorithm and a “bisecting” K-means algorithm is used.) Results indicate that the bisecting K-means technique is better than the standard K-means approach and (somewhat surprisingly) as good or better than the hierarchical approaches that we tested [2].

In this paper, the problem of organizing and browsing the top ranked portion of the documents returned by an information retrieval system is considered. Effectiveness of a document organization in helping a user to locate the relevant material among the retrieved documents as quickly as possible is studied. In this context it is examined that a set of clustering algorithms and experimentally show a clustering of the retrieved documents can be significantly more effective than traditional ranked list approach. It is also shown that the clustering approach can be as effective as the interactive relevance feedback based on query expansion while retaining an important advantage -- it provides the user with a valuable sense of control over the feedback process [3].

In this paper, a statistics-based approach for clustering documents and for extracting cluster topics is described relevant (meaningful) expressions (REs) automatically extracted from corpora are used as clustering base features. These features are transformed and its number is strongly reduced in order to obtain a small set of document classification features. This is achieved on the basis of principal components analysis. Model-based clustering analysis finds the best number of clusters. Then, the most important REs are extracted from each cluster and taken as document cluster topics [4].

In this paper, most state-of-the art document clustering methods are modifications of traditional clustering algorithms that were originally designed for data tuples in relational or transactional database. However, they become impractical in real-world document clustering which requires special handling for high dimensionality, high volume, and ease of browsing. Furthermore, incorrect estimation of the number of clusters often yields poor clustering accuracy. In this thesis, we propose to use the notion of frequent item sets, which comes from association rule mining, for document clustering. The intuition of our clustering criterion is that there exist some common words, called frequent item sets, for each cluster. We use such words to cluster documents and a hierarchical topic tree is then constructed from the clusters. Since we are using frequent item sets as a preliminary step, the dimension of each document is therefore, drastically reduced, which in turn increases efficiency and scalability [5].

In this paper, data mining (DM) brings knowledge and theories from several fields including databases, machine learning, optimization, statistics, and data visualization and has been applied to various real-life applications. A large amount of data mining articles have been published. The goal of this study is to establish an overview of the past and current data mining research activities from the title and abstract more than 1400 textual documents collected from premier data mining journals and conference proceedings. Specifically, this study applied document clustering approaches to determine which subjects had been studied over the last several years, which subjects are currently popular, and describe the longitudinal changes of data mining publications [6].

In this paper, the World Wide Web has become an important medium for disseminating scientific publications. Many publications are now made available over the Web. However, existing search engines are ineffective in searching these publications, as they do not index Web publications that normally appear in PDF (Portable Document Format) or PostScript formats. One way to index Web publications is through citation indices, which contain the references that the publications cite. Web Citation Database is a data warehouse to store the citation indices. In this paper, we propose a mining process to extract document cluster knowledge from the Web Citation Database to support the retrieval of Web publications. The mining techniques used for document cluster generation are based on Kohonen's Self-Organizing Map (KSOM) and Fuzzy Adaptive Resonance Theory (Fuzzy ART). The proposed techniques have been incorporated into a citation-based retrieval system known as Pub Search for Web scientific publications [7].

In this paper, a novel document clustering method based on the non-negative factorization of the term-document matrix of the given document corpus is proposed. In the latent semantic space derived by the non-negative matrix factorization (NMF), each axis captures the base topic of a particular document cluster, and each document is represented as an additive combination of the base topics. The cluster membership of each document can be easily determined by finding the base topic (the axis) with which the document has the largest projection value. The experimental evaluations show that the proposed document clustering method surpasses the latent semantic indexing and the spectral clustering methods not only in the easy and reliable derivation of document clustering results, but also in document clustering accuracies [8].

In this paper, a novel algorithm for document clustering is presented. This approach is based on distributional clustering where subject related words, which have a narrow context, are identified to form meta-tags for that subject. These contextual words form the basis for creating thematic clusters of documents. In a similar fashion to other research papers on document clustering, the quality of this approach with respect to document categorization problem is analysed and show it to outperform the information theoretic method of sequential information bottleneck [9].

In this paper, a probabilistic model for online document clustering is proposed and non-parametric Dirichlet process prior to model the growing number of clusters, and use a prior of general English language model as the base distribution to handle the generation of novel clusters. Furthermore, cluster uncertainty is modeled with a Bayesian Dirichlet multinomial distribution. Empirical Bayes method to estimate hyper parameters based on a historical dataset is used. A probabilistic model is applied to the novelty detection task in Topic Detection and Tracking (TDT) and compared with existing approaches in the literature [10].

3. EXISTING CLUSTERING ALGORITHMS

3.1 K-Means

K-Means is a partitioning method which is used to analyse data and checks the observation of the data. It is also known as hard C-Means clustering algorithm. Partitioning of the objects is done in such a way that makes objects within the cluster so close than that the objects in other clusters.

For document clustering setback, this way allocates every single of the document to one of the K number of clusters. An effectual cluster here will be a globe whereas centroid is believed to be its center of gravity. Centroid is also known as center point which is used to characterize every cluster and it is the point whose coordinates can be achieved by finding the average of the coordinates of each point which are given to clusters[11].

3.2 K-Means Particle Swarm Optimization (KPSO)

This hybrid begins alongside the K-Means module to produce an early clustering consequence and next PSO is requested to produce globe optimized cluster.

In KPSO, PSO is a stochastic tool for optimization that can be used easily to resolve various optimization issues. A ‘swarm’ implies to a grouping of a number of optimal solutions where each optimal result is known as a ‘particle’.

3.3 Fuzzy C-Means (FCM)

FCM is a popular soft clustering approach that combines features of K-Means and Fuzzy technique and it is an unsupervised clustering algorithm and it is widely applied in many fields like agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition [12].

FCM is also called a data clustering technique [13] as it grouped n clusters from a data set and direct every data point to every cluster and provides high degree of connection to specified cluster.

Algorithm consists of following steps [14]:

We have to fix c where c is (2<=c<n) and then select a value for parameter m and there after initialize the partition matrix

$U^{(0)}$. Each step in this algorithm will be labeled as r where r= 0,1,2…

1) We are to calculate the c center vector {Vij} for each step.

$$V_{ij} = \frac{\sum_{k=1}^{n} u_{ik}^{m'} \times x_{kj}}{\sum_{k=1}^{n} u_{ik}^{m'}}$$

2) Calculate the distance matrix D[c,n].

$$D_{ij} = \left[\sum_{j=1}^{m} \left[x_{kj} - v_{ij}\right]^2\right]^{[1/2]}$$

3) Update the partition matrix for the r^th step,U® as follows:

$$u_{ik}^{r-1} = \frac{1}{\sum_{j=1}^{c} \left[\frac{d_{ik}^{r}}{d_{jk}^{r}}\right]^{2/[m'-1]}}$$

If $\|U^{(k+1)}-U^{(k)}\| < \delta$, condition comes then we have to stop else go to step 2 by updating both the membership grades and the cluster centers iteratively [14].

4. PROPOSED WORK

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that keeps the most important things of this document that is original. Due to the fact nagging issue of information overburden has exploded, and also as the total amount of data has increased, so has desire for automated summarization. Technologies that may make a synopsis that is coherent into consideration factors such size, composing design and syntax.

A typical example of the utilization of summarization technology is the search engines. Typically, there are two methods to summarization which are automated removal and abstraction. Extractive practices work by picking a subset of present terms, phrases, or sentences in the text that is initial form the summary. In contrast, abstractive methods develop an internal representation that is semantic then make use of normal language generation techniques to develop a summary that is nearer to what a person might generate. Such a summary might explicitly consist of words not present in the first.

Research into abstractive techniques is an ever more crucial and study that is active, nonetheless as a result of complexity constraints, study up to now features concentrated primarily on extractive techniques. Clustering method is a widely found in unaided data arrangement, self-pattern exploring, and retrieving of data. High quality clustering approach plays a job this is certainly vital properly moving, summarization & company of data. The papers to be clustered are internet news articles, abstracts of analysis documents.

This work provides a strategy to document problem this is certainly clustering categorizing a set of archives into categories of appropriate archives on foundation of feature delicate clustering where a 'feature' could be the concept contained in the document. The documents useful for clustering are selected to be a group internet development articles as they are by the bucket load on the internet and want to precisely be categorized. A approach this is certainly hybridized intelligence based approach Particle Swarm Optimization (PSO) with standard partitioning clustering algorithms K-means and Fuzzy-C-Means will likely be applied to handle such high-dimensional clustering over. The recommended clustering approaches (in other words. K+PSO and FCM+PSO) picked for comparison and execution tend to be hybrids of conventional partitioning K-Means and FCM with Particle Swarm Optimization (PSO) method.

5. EVALUATION

The performance of documents clustering algorithms is evaluated according to the following internal and external validity measures:

Entropy: This is an information theoretic external validity measure. It analysis that on what basis documents of all categories are divided within each cluster. Entropy value 1 depicts worst entropy and value 0 depicts best entropy. The value of entropy represents 0 where each pile constitutes archive from a specific category only. The value of Entropy for each cluster *j* is computed as:

$$E_j = - \sum_i p_{ij} \log (p_{ab})$$

Here p_{ab} represents chances where a member of cluster b relates to class *a*. Computation of summation of entropy for m, set of the collection is done in the way as summation of entropies of individual cluster calculated from n_j Size of individual cluster at the place where summation will be obtained for all classes:

$$ECS = \sum_{j=1}^{m} \frac{n_j * E_j}{n}$$

Where *n* denotes total number of data points (clusters)

F-Measure: This is a non-information theoretic external validity measure. It predicts the degree with which all individual clusters have archives from the original division. F-Measure ranks clustering essence from 0 to 1 that is the value of F-measure will be 1 when all categories have their associated cluster containing the same document sets. For cluster b *&* class a.

$$F(a,b) = \frac{(2 * \text{Recall } (a,b) * \text{Precision } (a,b))}{((\text{Precision } (a,b) + \text{Recall } (a,b))}$$

Here,

$$\text{Recall}(a,b) = \frac{n_{ab}}{n_a}$$

$$\text{Precision}(a,b) = \frac{n_{ab}}{n_b}$$

Where n_{ab} denotes frequency of members of class a in cluster b, n_b denotes members of cluster b and n_a denotes frequency of members of class a.

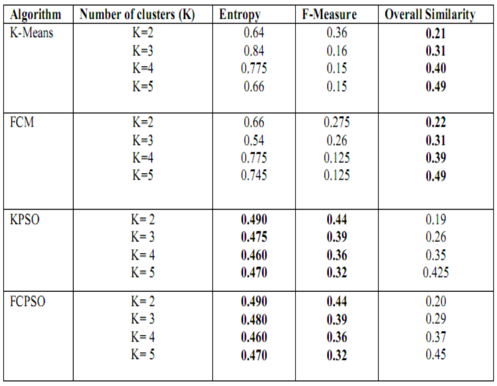
Overall Similarity: This is an internal validity measure for clustering quality with no background information. It measures the degree of cohesiveness between the documents. Cohesiveness of clusters may be adopted like an internal validity metric. It can be computed using weighted similarity between two documents as below:

$$\text{Overall similarity} = \frac{wx.wz}{\|w.x\|^2 + \|w.z\|^2 - wx.wz}$$

Taking the Reuters data set and finding the evaluation pattern for each discussed algorithm. The table below shows values of performance measures for varying values of K for all algorithms

Number of Reuters data set documents selected = 1504
Number of clusters=2 to 5

Table 1 : Depicting values of performance measures for each algorithm



Algorithm	Number of clusters (K)	Entropy	F-Measure	Overall Similarity
K-Means	K=2	0.64	0.36	**0.21**
	K=3	0.84	0.16	**0.31**
	K=4	0.775	0.15	**0.40**
	K=5	0.66	0.15	**0.49**
FCM	K=2	0.66	0.275	**0.22**
	K=3	0.54	0.26	**0.31**
	K=4	0.775	0.125	**0.39**
	K=5	0.745	0.125	**0.49**
KPSO	K= 2	**0.490**	**0.44**	0.19
	K= 3	**0.475**	**0.39**	0.26
	K= 4	**0.460**	**0.36**	0.35
	K= 5	**0.470**	**0.32**	0.425
FCPSO	K= 2	**0.490**	**0.44**	0.20
	K= 3	**0.480**	**0.39**	0.29
	K= 4	**0.460**	**0.36**	0.37
	K= 5	**0.470**	**0.32**	0.45

The numbers in bold depict better values of an evaluation measure for a clustering.

6. COMPARISONS

Comparison of all the clustering algorithm is provided on the basis of the internal and external validity measures.

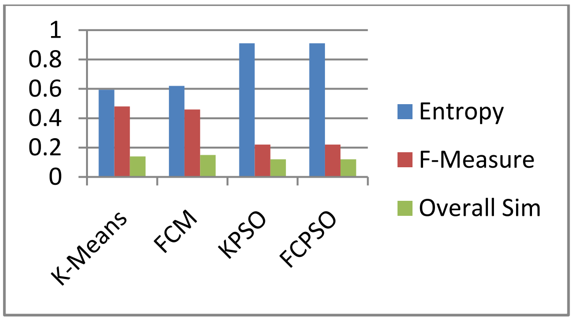


Figure 1: Comparison Graph for Reuters Data Set

7. CONCLUSION

Hybridized algorithms KPSO (K-Means Particle Swarm Optimization) and FCPSO (Fuzzy C-Means Particle Swarm Optimization) perform better than K-Means and Fuzzy C-Means. The cluster quality of FCPSO is better than that KPSO as it deals well with the overlapping nature of documents (which is the real scenario of documents on web) which makes it more suitable to use.

8. FUTURE WORK

Significant work has been done in the field of document clustering using hybrid swarm based approach. We have tried our level best to implement PSO as a hybrid approach to cluster documents. The field of swarm intelligence is still open to many challenges which provide future scope for improvement for document clustering problem.

Since the quality of clustering of documents widely rely on the data set nature; more text datasets varying in the number and type of documents can be explored to judge the effectiveness of the implemented algorithms.

Parameter tuning for inertia weight has not been explored in this thesis. Tuning this parameter can help in better convergence in PSO for more high-dimensional clustering problems.

Labeling of final clusters can also be improved by using an appropriate data structures for representation and storage.

Other external validity measures like purity, accuracy, random index, normal mutual information which have not been explored in this work can also be used for complete validation.

9. REFERENCES

[1] Magnus, Rosell, and Sumithra Velupillai. "The impact of phrases in document clustering for Swedish." In NoDaLiDa 2005, Joensuu, Finland, 2005, pp. 173-179. 2005.

[2] Michael, Steinbach, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." In KDD workshop on text mining, vol. 400, no. 1, pp. 525-526. 2000.

[3] Anton Leuski. "Evaluating document clustering for interactive information retrieval." In Proceedings of the tenth international conference on Information and knowledge management, pp. 33-40. ACM, 2001.

[4] Silva, Joaquim, Joao Mexia, Agra Coelho, and Gabriel Lopes. "Document clustering and cluster topic extraction in multilingual corpora." In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, pp. 513-520. IEEE, 2001.

[5] Benjamin Chin Ming Fung. "Hierarchical document clustering using frequent itemsets." PhD diss., SIMON FRASER UNIVERSITY, 2002.

[6] Yi, Peng, Gang Kou, Zhengxin Chen, and Yong Shi. "Recent trends in data mining (DM): Document clustering of DM publications." In Service Systems and Service Management, 2006 International Conference on, vol. 2, pp. 1653-1659. IEEE, 2006.

[7] Yulan, He, Siu Cheung Hui, and Alvis Cheuk M. Fong. "Mining a web citation database for document clustering." Applied artificial intelligence 16, no. 4 (2002): 283-302..

[8] Wei, Xu, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 267-273. ACM, 2003.

[9] Vladimir, Dobrynin, David Patterson, and Niall Rooney. "Contextual document clustering." In Advances in Information Retrieval, pp. 167-180. Springer Berlin Heidelberg, 2004.

[10] Jian, Zhang, Zoubin Ghahramani, and Yiming Yang. "A probabilistic model for online document clustering with application to novelty detection." In Advances in Neural Information Processing Systems, pp. 1617-1624. 2004.

[11] Soumi Ghosh and Sanjay Kumar Dubey," Comparative Analysis of K-Means and Fuzzy CMeans Algorithms", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[12] Y. Yong, Z. Chongxun and L. Pan, "A Novel Fuzzy C-Means Clustering Algorithm for Image Thresholding", Measurement Science Review, vol. 4, no.1, 2004

[13] S. Chen and D. Zhang, "Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure", IEEE Transactions on Systems, Man and Cybernetics, vol. 34, 1998, pp. 1907-1916.

[14] V. S. Rao and Dr. S. Vidyavathi, "Comparative Investigations and Performance Analysis of FCM and MFPCM Algorithms on Iris data", Indian Journal of Computer Science and Engineering, vol.1, no.2, 2010 pp. 145-151.