Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.163 – 167

Knowledge Discovery using Improved K-means Technique for Research Documents

Sandip S Rabade
PG Student, Department of Computer Network, Flora Institute of Technology, Khopi, Pune, India
sandiprabade@gmail.com

Prof. Bharat A. Tidke
Guide, Department of Computer Network, Flora Institute of Technology, Khopi, Pune, India
batidke@gmail.com

Abstract - Clustering focuses to organize a collection of data items into clusters, such that items within a cluster are more "similar" to each other than they are to items in the other clusters. The k-means method is one of the most widely used clustering techniques for various applications. Applications like Searching, Retrieving as well as Reading research Documents are more Time consuming because we need more time for searching or reading single papers or document, so it is required that use enhanced search engine which is based on fastest reading algorithm which provides best output or results. So we are proposed Enhanced architecture with improved k-means algorithm, which proposes a method for making the algorithm more effective and efficient, so as to get better clustering with reduced complexity. It will search the base keyword or string of the content from the knowledge database. Proposed work uses the search engine based on clustering and text mining

Keywords: Text mining, Clustering, k-means algorithm, Enhanced k-means algorithm.

1. INTRODUCTION

Data Mining and knowledge Discovery in data are attracting a significant amount of research, industry and also media. Document clustering is one of the most major techniques to all type documents automatically. This technique is to divide a given set of documents into a certain number of clusters automatically. Each cluster obtained by this technique represents a topic, which is different from the other topics. Thus, it enables a user to have an overall view of the topics contained in the documents so that this technique is often applied to the analysis of web data [13], news articles [12], patents and research papers [1] and so on.

Knowledge Discovery in Databases is an automatic, exploratory analysis and modeling of large data repositories in web. KDD is the organized process of identifying valid, useful, and understandable patterns from large and complex data sets. Data Mining is the core of the KDD process, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis . In the document clustering, the first step of preprocessing is term extraction from a set of documents. After the term extraction process, various clustering methods can be applied By utilizing these extracted characteristics of terms. Text mining or knowledge discovery from text for the first time mentioned in Feldman et al. deals with the machine supported analysis of text [5]. It uses methods from information extraction, information retrieval also natural language processing and also connects them with the algorithms and technique of knowledge discovery database , machine learning, statistics and data mining. a similar procedure is selected as with the knowledge discovery database process, where by not data in general, but the analysis of text documents are in focal point. One problem is that we now have to deal with problems of from the data modeling perspective unstructured data sets.

We can refer to interrelated research areas if we define text mining. For each of them, we can give a different definition of text mining [7], which is motivated. The use of K-Means algorithm allow us to implement semi-supervised learning clusters using an algorithm so as to help to recognize approximate the text to search using predefined patterns and the implementation of a cluster algorithm for consultations within the database manager My SQL i.e. database manager that allows free use of multithreading, multi-search and multi-user in order to obtain scientific research papers in web. But for large data sets the computational complexity of the original k-means algorithm is very high. So finally the algorithm results in different types of clusters depending on the random choice of initial centroids. Researchers made several attempts to improve K-means Algorithm.

This paper deals with a method for improving the accuracy and efficiency of the k-means algorithm. Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly categorized into two types: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm separates the dataset into desired number of sets in a single step [9]. Numerous methods have been proposed to solve clustering problem.

2. IMPLEMENTATION

2.1 Search Architecture Model

first part of the architecture needed in our difficulty where the user begins a search interface for entering text within the information used in the process of searching patterns within a knowledge base in order to obtain parameters for the selection of cluster where the search was implemented [1], once achieved the search is conducted within the database in the process of localization Documents. The implementation of data mining to solve a queries involves the need to implement a methodology focus into the analysis of pattern into the texts, where there are several methodologies are used built-oriented identify of attributes that will be reviewed, so this kind of methodologies are not used for our implementation as the proposed work need a methodology to be acceptable evolutionary behavior.

In figure 1 shows the search architecture model [10], in search architecture model the text mining is mainly used and as a first part of the architecture as a problem outline where the user starts a searching through various interfaces for entering text within the information used in the process of searching patterns within a knowledge base for accessing parameters for the selection of cluster where the actual search was implemented, once the searching is completed within the database in the process of localization papers.

Below figure1 shows proposed system architecture.

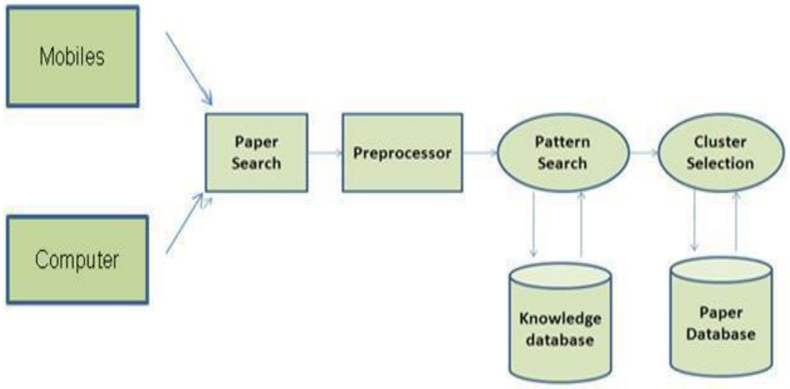


Figure 1. System Architecture

In the search architecture we will inputs the location of the research to get text patterns and work reading line by line and it is supported by a knowledge portion that is fed into a rank semi-automatically with the information collected from items previously stored. Before searching the pattern of knowledge we will apply preprocessing for reducing time and size complexity of process, so we use stop word elimination approach as well as steaming. Here we uses the search pattern into the research paper for accessing the database searching patterns of knowledge, with this the proposed work generate there engine to make pattern matching needed in the problem outline start with a pattern matching that read the article searching a similar pattern compared with the knowledge base once we are locate in where it relates is selected cluster on which will be uploaded from the article in the database.

2.2 Improved K-means Algorithm

In the improved k-means algorithm we are enhancing the performance of k-means clustering algorithm. In the previous method the initial centroids are selected randomly, so this method is very sensitive to the initial starting points and it does not guarantees to produce the unique clustering results. In the paper [2] authors uses two methods for finding initial clustering i.e finding initial centroids and assigning data points to appropriate clusters. so we are proposed an enhanced algorithm to improve the accuracy and efficiency of the k-means clustering algorithm. In this paper we are proposed a new approach for finding the better initial centroids with minimized time complexity. In the proposed algorithm first we will checking, the given data set contain the negative value attributes or not. If the data set contains the negative value attributes then we are transforming the all data points in the data set to the positive attribute value in the given data set. Here positive space is subtracting the each data point attribute with the minimum attribute value in given data set. Transformation is required, because in the proposed algorithm we will calculate the distance from origin to each data point in the data set. So, when we are selecting the different data points then we will get the same Euclidean distance from the origin. Because of that we will get result is in incorrect selection of the initial centroids. So to overcome this problem all the data points are transformed to the positive space. then for all the data points as we will get the unique distances from origin. If there is all positive values in data set then the transformation is not required.

In the next step of algorithm, for each data point we will calculate the distance from origin. Then, the original data points are sorted accordance with the sorted distances. After sorting partition the sorted data points into k equal sets or numbers. In each set take the middle points or mean value as the initial centroids. These initial centroids lead to the better unique clustering results. Next, for each data point the distance calculated from all the initial centroids. The next stage is an iterative process which makes use of a heuristic approach to reduce the required computational time. The data points are assigned to the clusters having the closest centroids in the next step. ClusterId of a data point denotes the cluster to which it belongs. NearestDist of a data point denotes the present nearest distance from closest centroid.

```
Require: D = {d1, d2, d3,..., di,..., dn } // Set of n data points.
di = { x1, x2, x3,..., xi,..., xm } // Set of attributes of one data point.
k // Number of desired clusters.
Ensure: A set of k clusters.
Steps:
1: In the given data set D, if the data points contains the both Positive and negative attribute values then go to step Otherwise go to step 4.
2: Find the minimum attribute value in the given data set D.
3: For each data point attribute, subtract with the minimum Attribute value.
4: For each data point calculate the distance from origin.
5: Sort the distances obtained in step 4. Sort the data points Accordance with the distances.
6: Partition the sorted data points into k equal sets.
7: In each set, take the middle point as the initial centroid.
8: Compute the distance between each data point di (1 <= I <= n) to all the initial centroids cj (1 <= j <= k).
9: Repeat 10: For each data point di, find the closest centroid cj and assign di to cluster j.
11: Set ClusterId[i]=j. // j:Id of the closest cluster.
12: Set NearestDist[i]= d(di, cj).
13: For each cluster j (1 <= j <= k), recalculate the centroids.
14: For each data point di,
14.1 Compute its distance from the centroid of the present nearest cluster.
14.2 If this distance is less than or equal to the present nearest distance, the data point stays in the same cluster.
Else
14.2.1 For every centroid cj (1<=j<=k) compute the distance d(di, cj).
End for;
    Until the convergence criteria met
```

Algorithm 1. The ***Enhanced Method*** *algorithm*

In the next step, for each cluster the new centroids are calculated by taking the mean of its data points. Then for each data point the distance calculated from the new centroid of its present nearest cluster. If this distance is less than or equal to the previous nearest distance, then the data point stays in the same cluster, otherwise for each data point we are need to calculate the distance from all centroids. After calculated the distances, the data points are assigned to the appropriate clusters and then Cluster Id's are given and new Nearest Dist values are updated. This reassigning process is repeated until the convergence criterion is met.

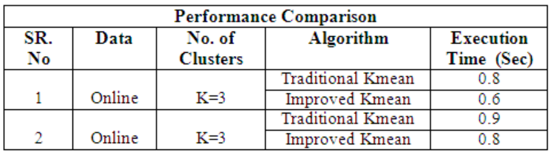
3. RESULT

The implementation of the search engine using data mining scheme works by using a document search to research this for easy portability to multiple platforms helping a simple web search interface where the user enter the information that want and finally showing a list of papers or documents available. The Figure 3 shows the interface to be used for searching and browsing of research papers with a simple structure that helps the user to identify a single text area where the user enter text on the search and finally at the bottom were generated the results of this search. The time in the search takes one minute to search in all the database knowledge, making a best time than the usual when the researcher are searching in the web. The figure 4 shows results of the search in the defined clusters in the form of all extension files, with their own value of comparison using the full text in the search engine and compared with their own category in this case data base.

4. RESULT COMPARISON

We apply both the algorithms original and Improved for the different number of records. Both the algorithms original and Improved need number of clusters as an input. In the basic k-means clustering algorithm set of initial centroids are required. The proposed method finds initial centroids systematically. The proposed method requires only the data and number of clusters as inputs. The basic K-means clustering algorithm is executed for the different data values of initial centroids. In each experiment the time was computed and taken the average time of all experiments. Table 1 represents performance comparison of both the algorithm and Figure 5 shows the performance graph of comparison of the Traditional and Improved k-mean clustering algorithms. The experiments results show that the proposed algorithm is producing better results in fewer amounts of computational time and accuracy compared to the basic k-means algorithm.

Table 1: performance Comparison in Terms of Time



Performance Comparison				
SR. No	Data	No. of Clusters	Algorithm	Execution Time (Sec)
1	Online	K=3	Traditional Kmean	0.8
			Improved Kmean	0.6
2	Online	K=3	Traditional Kmean	0.9
			Improved Kmean	0.8

5. ACKNOWLEDGEMENTS

I hereby take this opportunity to record my sincere thanks and heartily gratitude to PG Coordinator Prof. Bharat A. Tidke useful guidance and making available to me his intimate knowledge and experience in making "**Knowledge Discovery using Improved K-means Technique for research Documents**" as a project and i express my sincere gratitude towards people who have worked in this area.

6. CONLUSION

The k-means algorithm is widely used for clustering large sets of data. But the standard algorithm do not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. Moreover, the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times, during every iteration of the loop. This topic presents an enhanced k-means algorithm which combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. The previous improvements of the k-means algorithm compromise on either accuracy or efficiency. The proposed algorithm is to be more accurate and efficient compared to the original k- means algorithm. This proposed method finding the better initial centroids and provides an efficient way of assigning the data points to the suitable clusters. So it is very effective to accessing research papers using Enhanced clustering algorithm with improving in time as shown in graph. So in future it can be possible to implement this work for various or different datasets as well as paper repositories for accessing various documents efficiently.

REFERENCES:

[1] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A.Ramadan, "An Efficient enhanced k-means clustering algorithm", journal of Zhejiang University, 10(7): 16261633, 2006.

[2] K.A.Abdul Nazeer and M. P. Sebastian,"Improving the accuracy and e_ciency of the k-means clustering algorithm", in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1,London, UK,July 2009.

[3] Chen Zhang and Shixiong Xia,"K-means Clustering Algorithm with Improved Initial center", in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, "A New Algorithm to Get the Initial Centroids", proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.

[5] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means:
an algorithm for Centroids initialization for k-means",department of information science and Electrical Engineering Poli technique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[6] A. Bhattacharya and R.K.De,"Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression pro_les", bioinformatics, Vol. 24, pp. 1359-1366, 2008.

[7] Jieming Zhou, J.G. and X. Chen,"An Enhancement of K means Clustering Algorithm", in Business Intelligence & Financial Engineering, BIFE ^a09. International Conference on, Beijing,2009.

[8] Chakrabarti, S., "Mining The Web: Discovering knowledge from hypertext data", Part 2. 2003.

[9] Jieming Zhou, J.G. and X. Chen, "An Enhancement of K means Clustering Algorithm", in Business Intelligence and Financial Engineering, BIFE 09. International Conference on, Beijing,2009.

[10] Sachin Shinde, Bharat Tidke, "improved k-mean algorithm for searching research papers", IJCSCN ,ISSN 2249-5789,vol 4(6), pp.197-202 dec 2014.

[11] Aukasz Machnik, "Documents Clustering techniques", in Annales UMCS Informatica Lublin-Polonia Sectio AI,p 401 411,2004. International Conference on, Beijing,2009.

[12] Marijus Bernotas, Kazys Karklius, Remigijus Laurutis, Asta Slotkien A, "The Peculiarities Of The Text Document Representation, Using Ontology And Tagging-Based Clustering Technique", 124x Information Technology And Control, Vol.36, No.2,2007.

[13] Anand M. Baswade, Prakash S. Nalwade, "Selection of Initial Centroids for k-Means Algorithm", in International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 7,pg.161 ^a 164, July 2013,

[14] E.AlanCalvillo,Alejandro Padilla,Jaime Munoz"Searching Research papers using clustering and text minig" IEEE 2013.

[15] Vishwanath Bijalwan ,Vinay Kumar,Pinki Kumari and Jordan Pascual, "KNN based Machine Learning Approach for Text and Document Mining", in International Journal of Database Theory and Application,Vol.7, No.1 ,pp.61-70,(2014).