Knowledge Discovery based Research Papers Recommender System using Improved K-means Techniques

[1]Sandip S. Rabade,[2]Shweta A. Joshi

[1] Student ME(Computer),[2]Prof. ME(Computer)
[1]Department of Computer Engineering
[1]Flora Institute of Technology, Pune,India.

***Abstract*–The main objective of recommender system is to provide correct and useful recommendations that makes user happy and satisfied. The users are interesting in accessing the document collection which contains the available information. Clustering is the main analytical method used in data mining. For data clustering the generally accepted algorithm is k-means. The similar kind of data presented in the large data sets are tried to be clustered together using k-means. The one of the limitation of the traditional K-means algorithm is that it require a large computational time. Searching and retrieving also reading research documents is more time consuming. To overcome this problem we develop a search engine for recommending research papers which is based on improved K-means algorithm that provide best results with reduced time complexity. To store the papers the MongoDB database is used which can support large number of simple read/write operations per second. MongoDB is NoSQL database. Search engine used is based on clustering and text mining.**

***IndexTerms*–Clustering, Text mining, K-means algorithm, MongoDB, NoSQL**

I. Introduction

A "recommender system" is a fully functional software system that applies at least one implementation to make recommendation. We are going to develop search engine that are going to retrieve the research papers based on user entered query in the search box. To faster theses searching and retrieving operation we use the mongoDB database [13] as this is well known NoSQL database. MongoDB contains big amount of data than MySQL database.To get the faster results and clustering we used the K-means clustering algorithm.

Data Mining and knowledge Discovery in data are attracting a significant amount of research, industry and media. Document clustering is one of the most major techniques to group 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 , news articles , patents and research papers and so on.

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction. 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 (KDT) for the first time mentioned in Feldman et al. deals with the machine supported analysis of text. It uses methods from information extraction, information retrieval also NLP (natural language processing) and connects them with the algorithms and technique of knowledge discovery database (KDD), machine learning, statistics and data mining.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. So, innovative questions for the new data mining methods take place. One problem is that we now have to deal with problems of from the data modeling perspective unstructured data sets.

The existing system uses MySQL database which is structured database and contain limited amount of data. The previous system applying K-means algorithm on MySQL database and calculate the time complexity of this algorithm. As the load on MySQL database increases the resulting algorithm performance degraded. We are going to use the MongoDB database to perform the faster operation than MySQL. MongoDB contain huge amount of data and store the data on multiple servers in distributed manner. As the load on MongoDB increases, this situation does not degrade the performance of system and algorithm. MongoDB is purerly NoSQL database. Our motivation is to improve the time complexity of K-means algorithm to get the faster clustering results on large database.

II. RELATED WORK

There are already several improvements have been done on traditional k-means algorithm by researchers. Every researcher tries to improve extent to which time is well used for k-means algorithm. Several techniques have been applied to find the better initial centroids.

In traditional k-mean algorithm k random objects are selected as primary cluster center [2]. For every iterations distance between each data object and center is calculated. After that using this distance every data object is assigned to nearest cluster. This process is continued until a convergence criterion is met so that this algorithm requires high number of computation and hence takes an extensive time.

M. Fahim et al. [2] introduced the better method of designating the data objects to appropriate clusters with less computational time. In this approach the initial centroids are randomly selected. As the centroids are selected randomly there is no guarantee to produce unique clustering results.

In the K.A. Abdul Nazeer et al. [4] approach two methods are used in enhanced k-means algorithm. The first method is used to find the better initial centroids. The second method is used to assign the data points to the suitable cluster with less time. This approach is useful for improving the time complexity of k-means algorithm. Zhang Chen et al. [5] approach tries to avoid the random selection of initial center. In this approach initial centroid algorithm based on k-means is proposed.

Koheri Arai et al. [6] approach combine the k-means algorithm and hierarchical algorithm to find the better initial centriod. Bhattacharya et al. [7] proposed Divisive Correlation Clustering Algorithm (DCCA) which take k number of cluster as a input. Cluster is created without taking the initial centroids. Sharfuddin Mahmood et al. [8] introduced an algorithm that assign the objects to the appropriate clusters by checking the check point value. In this approach performance of k-mean algorithm is improved. Fang Yuan et al. [9] approach find the initial centroid systematically.

Nirmala Pudota, Antonina Dattolo, Andrea Baruzzo, Felice Ferrara, Carlo Tasso propose Automatic keyphrase extraction and ontology mining for content based tag recommendation [12]. Collaborative tagging represents for the Web a potential way for organizing and sharing information and for heightening the capabilities of existing search engines. However, due to the lack of automatic methodologies for generating the tags and supporting the tagging activity, many resources on the Web are deficient of tag information and recommending opportune tags is both a current open issue and an exciting challenge. This paper approaches the problem by applying a combined set of techniques and tools (that uses tags, domain ontologies, keyphrase extraction methods) thereby generating tags automatically. The proposed approach is implemented in the PIRATES (Personalized Intelligent tag Recommender and Annotator TEStbed) framework, a prototype system for personalized content retrieval, annotation, and classification. A case study application is developed using a domain ontology for software engineering.

III. OVERVIEW OF SYSTEM

The objective of the proposed recommender system is to provide correct and useful recommendation that makes user happy and satisfied. The proposed recommender system uses several techniques to achieve this objective. Data mining make best use of large data warehouses to extract the useful information. User can retrieve the knowledge from the stored data using data mining techniques. Data mining is frequently used in knowledge discovery system. The main objective is to propose a enhanced architecture with improved k-means algorithm, which propose a method for making the algorithm so effective and efficient, so as to get better clustering results.

The proposed system defines the search architecture model where the actual activities of the system are clearly defined. This search architecture model provides the user interface where user entered the search query and retrieves the documents. The proposed system is implemented on two kind of databases one is MySQL and another one is MonogoDB. The differences between these two databases are MySQL is relational database and MongoDB is NoSQL database. The storage capacities of both the databases are different. Relational databases have little capabilities to horizontal scale over many servers. A key feature of NoSQL system is “share nothing” horizontal scaling-replicating and partioning data over many servers. In the proposed system, improved K-means algorithm is applied on both kind of database and their performance is compared on different processors.

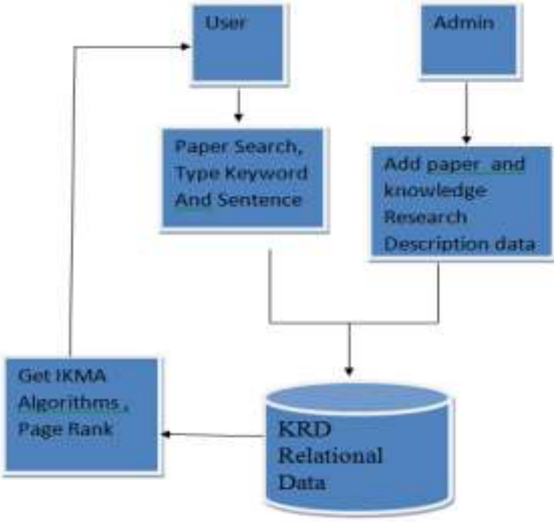

User
Admin
Paper Search, Type Keyword And Sentence
Add paper and knowledge Research Description data
Get IKMA Algorithms, Page Rank
KRD Relational Data

Figure 1: Proposed Research paper Recommender System Search Architecture Model.

The above Fig.1 shows the search architecture model of proposed system. This search architecture model provides the interface for both user and admin. Admin can store the papers in the database and also enter the title of paper, year of publication, abstract part of paper and domain of paper in the database. Admin can access both mongoDB and MySQL database. When the users enter the query or type keyword and sentence for paper search in search box, the IKMA(Improved K-means Algorithm) applied on database. The k-means algorithm generates the resulting clusters that contain the similar kind of paper documents and retrieve those documents to the user. The papers are retrieved to the user in year wise sorted manner i.e. recent year paper is retrieved first and then others. The user can also read the title, abstract and domain of the each retrieved papers.KRD is knowledge research data where information about each paper is stored like paper title, abstract, paper domain and year of publication. When user type keyword or sentence for research paper, that keywords or sentence are matched to these knowledge research data. By applying IKMA algorithm the clusters of similar documents as per user typed keyword are created and that resulting documents are retrieved to the user.

The below Fig. 2 shows the proposed system architecture, dataset component represent the monogoDB or MySQL database. When user enters the query, the pre-processing is done on database by improved k-means algorithm. The resulting cluster of documents is created and that documents are output to the user.

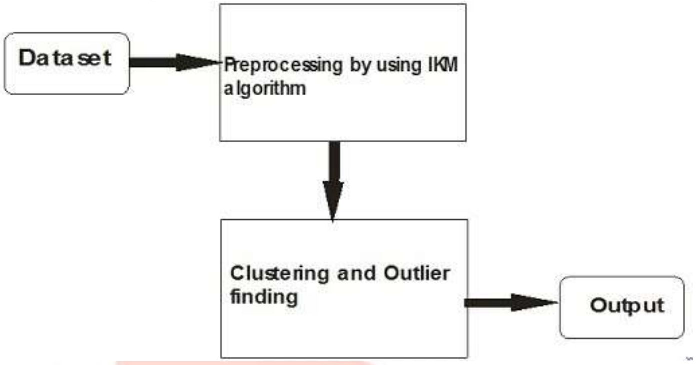

Dataset
Preprocessing by using IKM algorithm
Clustering and Outlier finding
Output

Figure 2: Proposed System Architecture

IV. Methodology and Improved K-means Algorithm

The mathematical model implemented in the proposed system is as follow:

- **Design using Set Theory**

- S={Ta,A,Op,Su,Fa}
- Ta={T1,T2,T3,…….Tn} No. of input as keyword/sentence.
- A= {a1} each user having only one account and contains user ID.
- Op={Op1,Op2,Op3,….Op6} Operations performed by User authority
- Su-success {connection establish, sending or retrieving data successfully}
- Fa-failure {not a valid user, incorrect password}

- **Mathematical representation:**

1) To add new account :

$$\sum_{i=1}^{n} \boldsymbol{Ai} = \mathbf{1} \qquad \textbf{(1)}$$

$Note: 1 = true, 0 = False$
Then user is already exists.
Else
Add new account in it.
Ai contain user id.

2) To login in Account:

$$\sum_{i=1}^{n} \boldsymbol{Ai} = \mathbf{1}\ \textbf{then}$$
$$\sum_{i=1}^{n} \boldsymbol{Ai} = \sum_{i=1}^{n} \boldsymbol{Ud} = \mathbf{1} \qquad \textbf{(2)}$$

(Verifies password in User Details according to the User authority ID present in Account)
If and only if account information and user details are match i.e. account information which contain User id if match with user Id and Password of user authority details then Login Success
Else
Login failed

3) User Authority Operations:

$$\sum_{i=1}^{n} Opi = 1 \ then \ Operation \ success \qquad (3)$$

// perform algorithm for check sentence similarity
Else
Operation Failed
// show result “similarity not found”.

- **Improved K-means Algorithm**

In the traditional k-means algorithm the initial cluster center is chosen by randomly selecting any data member. After that the distance between

data objects and cluster centers is calculated. The data object nearest to the cluster center is included in that cluster. This process is repeated until all the objects are not properly placed in the appropriate cluster. This algorithm requires high number of computation.

The traditional k-mean algorithm work by following equation [11]:

$$\underset{S}{arg\,min} \sum_{i=1}^{k} \sum_{X_j \epsilon S_i} \left\| X_j - \mu_i \right\|^2 \qquad (4)$$

This formula represents a given set of observations (X1,X2,…,Xn), where each observation represent an element of cluster with d-dimensional real vector [11]. S= {S1,S2,….,Sk) that is the no. of clusters where μi is the mean of point in Si.

We are going to propose an effective method that enhances the time of existing k-means algorithm. We are going to proposed a new way of finding the better initial centroid with less time complexity. In the proposed algorithm we are finding in the given dataset whether data points attribute value are positive or negative. If the dataset contain negative attribute value then transformation is required to convert the data point attribute value to positive value. The positive value is calculated by performing the subtraction operation. In subtraction operation minimum attribute value in the given data set is subtracted from each data point attribute value. When all the data points in the datasets are positive we get unique distance from origin.

In the next step of algorithm, the distance between each data point and origin is calculated. The distances calculated from this step are sorted. By using this sorted distances all the data points are sorted to place in appropriate clusters. After performing the sorting operations all the sorted data points are placed into e equal sets. In each set the central point is taken as initial centroid that gives better clustering results. The distance between each data point and all initial centroids is calculated. In the next step the data points are assigned to the appropriate cluster having initial centroid. The heuristic approach is used at every iteration of assigning data points to cluster to minimize time complexity. The Improved K-means Algorithm Steps are as follows:

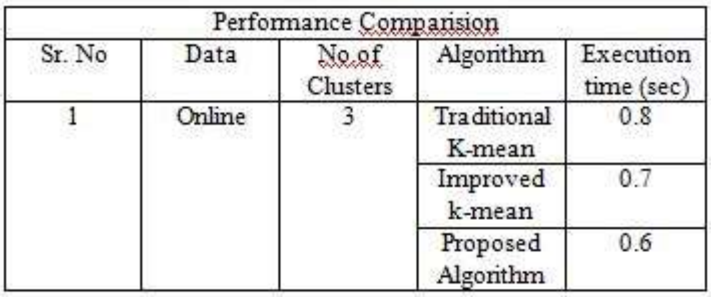
```
Require: P = { p1, p2,p3 ,.., pi,..., pz} // set of n data point
pi = { a1, a2, a3,…, ai,…., am} // set of attributes
e // desired number of cluster
Check: e number of clusters
1. Find the better initial centroid.
2. Check the given data set P contain the negative value attribute or not then go to step 3 and 4 otherwise go to step 5
3. If the dataset contain negative attribute value find the minimum attribute value.
4. Transform all the data point in the given data set to positive value by subtracting minimum attribute value from each data point attribute value.
5. Calculate the distance between origin and each data point in the data set. Then the original data points are sorted accordance with sorted distance.
6. Locate the data points that are sorted into e equal sets.
7. In each set calculate the average value and take it as initial centroid.
8. For each data point pi (1<=i<=z) calculate the distance between each data point and all initial centroids cl (1<=l<=e)
9. Repeat step 8
10. For each data point pi, find the nearest initial centroid cl and assign pi to cluster cl.
11. Set CID[i] = l;
12. Set ClosestDist[i] = dist( pi, cl)
13. For every cluster l (1<=l<=e), recalculate the  centroid.
14. For every data point pi,
    14.1 Calculate the distance between the centroid of the present cluster and data point.
    14.2 Check the distance if it is equal to or less than present nearest distance, the data point remains in the same cluster and find the threshold value
    // This used for improve performance
    Else
    14.2.1 For every centroid cl (1<=l<=e) compute the distance dist(pi,cl)
    End For;
```

The above procedure mentioned in the algorithm is repeated until convergence criterion is reached where there is no need to move centroid any more.

V. Experimental Result and Comparative Analysis

We are going to compare performance of the resulting outputs of original k-means algorithm, improved k-means algorithm and our proposed advanced algorithm in terms of time. The below table 1 shows the experimental results of execution time required for each algorithm.

Table 1: Performance Comparison in terms of time



Performance Comparision				
Sr. No	Data	No of Clusters	Algorithm	Execution time (sec)
1	Online	3	Traditional K-mean	0.8
			Improved k-mean	0.7
			Proposed Algorithm	0.6

The graphical representation of above table is shown below:

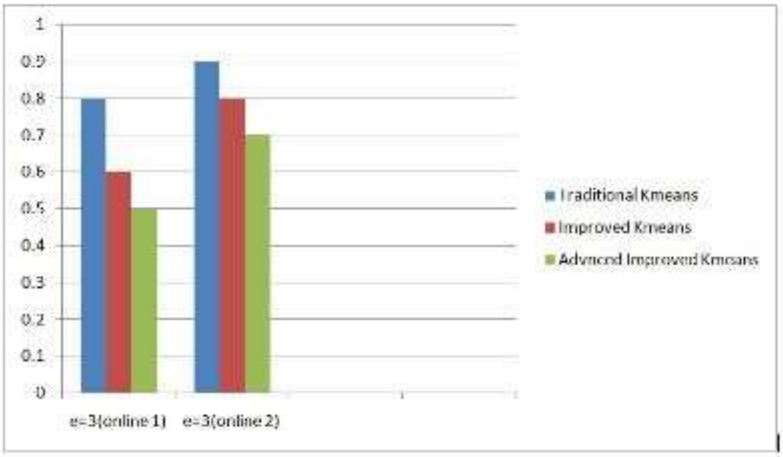


Figure 3: Graphical Representation of K-means, Improved K-means and Advanced Improved K-means Algorithm

We consider the different number of records and apply the traditional k-means, improved k-means and our proposed advanced algorithm on those records. The output generated by each of these algorithms we check in terms of time i.e. time required to retrieving the result. The resultant time required for each algorithm is shown in above table. The method proposed in our algorithm find the initial centroid systematically. For every experiments performed on traditional algorithm, improved algorithm and our proposed advanced algorithm the required time is computed and average time of all the experiment is taken. The performance comparisons of these three algorithms are shown in Table 1. The resulting values in terms of time presented in the table shows that proposed advanced algorithm is producing better results in fewer amount of computational time.

Initially, the data cube and assign it to NULL i.e. Values of all positions (indexes) of data cube are initialize to NULL. Number of positions of data cube are same as multiplication of value of each dimension of data cube or multiplication of domain of condition and decision attribute.

Table 2: Performance comparison between MongoDB and MySQL with respect to execution time on three different processors.

Sr. No.	Data	No. of Cluster	Configuration	Execution Time(sec)
1.	MongoD0B	K=8	Core2Dual, 500 GB HDD, 2.2 GHz, 2GB RAM	0.9
	MySQL			1.8
2.	MongoDB	K=8	I3 Processor, 500 GB HDD, 2.5 GHz,2GB RAM	0.8
	MySQL			1.2
3.	MongoDB	K=8	I5 Processor, 500 GB HDD, 2.6 GHz, 4GB RAM	0.6
	MySQL			1.0

We analyze the degree and exact conformity of proposed system advanced improved k-means algorithm by performing the operation on MongoDB [13] and MySQL database. Table 2 shows performance comparison in terms of document accessing time, retrieving time between MongoDB and MySQL on three different processors. The proposed system is implemented on both MongoDB and MySQL database. We check the proposed system result output time on two databases on three different processors. Core2Dual, I3 and I5 processor we used to check the resulting time of proposed system. Whatever the resulting time proposed system get is shown in above table 2. The time we get is dynamic. Time depends on the processor speed, RAM, database access time and retrieval time and also the algorithm execution time in terms of resulting clusters creation. Each time we run the system on three different processor we get the different times which is not static. Time also depend on the server response time. Proposed system used the Apache Tomcat 7.0 server. Each time when user enter the query for research paper, that query is transfer to the server then server retrieve the resulting documents to the user side.

VI. ACKNOWLEDGEMENTS

I hereby take this opportunity to record my sincere thanks and heartily gratitude to Guide Prof. Shweta A. Joshi useful guidance and making available to me his intimate knowledge and experience in making **" Knowledge Discovery based Research Papers Recommender system using Improved K-means Techniques"** as a project and i express my sincere gratitude towards people who have worked in this area.

VII. CONCLUSION

The k-means algorithm is the most popular algorithm for assigning object to specific cluster. But this traditional k-means algorithm requires high number of computation and takes a large amount of time. In this paper the proposed advanced algorithm gives better performance as compared to traditional k-means algorithm. This proposed algorithm find the better initial centroid and assign the data objects to appropriate clusters efficiently with less time complexity. This algorithm is beneficial to accessing research papers. The proposed system used the MongoDB database which is advanced NoSQL database system which gives the better and faster result and also stores the large amount of documents than MySQL.

REFERENCES

[1] S. AL Manseer, A. Malibari, "Improved Teaching method of Data Mining course", I.J.Modern Education and computer science, Second Volume, Page-15-22,2012

[2] 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.

[3] M. Yedla, S. R. Pathakota, T.M. Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, Vol. 1(2):121-125, 2010.

[4] K. A. Abdul Nazeer, M.P. Sebastian, "Improving the Accuracy and efficiency of the K- Means Clustering Algorithm", International Conference on Data Mining and Knowledge Engineering (ICDMKE). Proceeding of the World Congress on Engineering(WCE-2009), Volume : 1 , 2009.

[5] Chen Zhang, Shixiong Xia, "K-Means Clustering Algorithm With Improved Initial Center", ISBN: 978-0-7695-3543-2, pp: 790-792.

[6] 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.

[7] 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.

[8] Sharfuddin Mahmood, "A Proposed Modification of K-Mean Algorithm", I.J. Modern Education and Computer Science, 2015,6,37-42.

[9] 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.

[10] E.alan Cavillo, Ale jandro Padilla, Jaime Munoz, " Searching Research Papers using clustering and text mining", IEEE 2013.

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

[12] Nirmala Pudota, Antonina Dattolo, Andrea Baruzzo, Felice Ferrara, Carlo Tasso, "Automatic keyphrase extraction and ontology mining for content based tag recommendation", Artificial Intelligence Laboratory Department of Mathematics and Computer Science University of Udine, Italy.

[13] Rupali Arora, Rinkle Rani Aggarwal " Modelling and Querying Data in MongoDB", International Jouranal of Scientific and Engineering Research, Volume 4, Issue7, July-2013.

[14] Sandip Rabade, Bharat Tidke " Knowledge Discovery using Improved K-means Techniques for Research Documents", IJCSMC, Vol.5,Issue.3, March 2016, pg.163-167.

[15] Sandip Rabade, Bharat Tidke, " Research Papers Recommender System using Knowledge Discovery", Fifth Post Graduate Conference of Computer Engineering, CPGCON2016.

[16] J. Han, M. Kamber, J.Pei, "Data Mining-concepts and techniques", Third Edition, Chapter:7, Page:401.

[17] D. Sharmila Rani, V.T. Shenbagamuthu," Modified K-Means Algorithm for Initial Centroid Detection", IJIRCCE, Vol.2, Special Issue 1, March 2014.