Multilink Constrained k-means Clustering Algorithm for Information Retrieval

M.Parvathavarthini [1] , E.Ramaraj[2]

[1]M.Phil Scholar, [2] Professor

Department of computer science and Engineering
Alagappa University
Karaikudi, India.

***Abstract*— Clustering is traditionally viewed as an unsupervised method for data analysis. However, in some cases information about the problem domain is available in addition to the data instances themselves. In this paper, the popular k-means clustering algorithm can be profitably modified to make use of information with available instances is demonstrated. We can also apply this method to the real-world applications such as University database, hospital database etc. for information retrieval. In this proposed method the University data are collected to perform the k-means clustering algorithm to information retrieval. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Many universities and public libraries use IR systems to provide access to books, journals and other documents. An information retrieval process begins when a user enters a query into the system.**

***Keywords*—** Clustering, Information Retrieval, k-means algorithm, Database.

I. Introduction

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning statistics, and database systems (2). The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is ready for application in the business community because it is supported by three technologies that are as follows:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms
-

In data mining, Cluster is a group of objects that belong to the same class (3). The similar object are grouped in one cluster and dissimilar are grouped in other cluster. A cluster is a subset of objects which are “similar”. A subset of objects such that the distance between any two objects in the cluster is less than the distance between any object in the cluster and any object not located inside it. Clustering is a process of partitioning a set of data or objects into a set of meaningful sub-classes, called clusters. It also helps users understand the natural grouping or structure in a data set.

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text or other content-based indexing. Automated information retrieval systems are used to reduce what has been called "information overload" (10). Many universities and public libraries use IR systems to provide access to books, journals and other documents. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy (1).

There are mainly three techniques for information retrieval as follows: they are limiting the search by field, full text or word search and subject term search (5).

A. Limiting search by field

When searching a database, specify the search by using different search fields. Typical search fields available in databases are author, title, publication year and ISBN number (9). These fields can also be used to narrow search for retrieving the information.

B. Full-text or word search

Full-text or word search is most suited for the beginning stages of research on a topic. Full-text search results in a lot of unnecessary data because it looks for the terms anywhere in the record. It is easy to find relevant results using word searches.

C. Subject term search

A subject term search gives more accurate results than a full-text search, and limits the search to the subject term field. It also includes search results in different languages and when searching international databases that may contain varying terms for the same topic(4).

Retrieval Technique

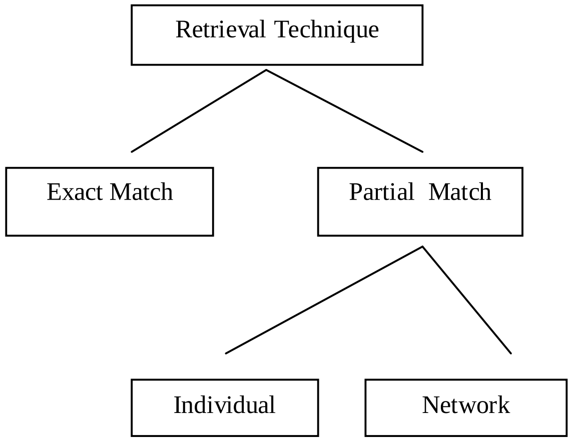


Fig 2.1 Subject term Search

II.LITERATURE REVIEW

T. C. Sprenger et al [6] has been proposed a hierarchical visual clustering method called H-BLOB, which provides an efficient level of-detail strategy and is consequently capable to cluster and visualize very large and complex data volumes. The algorithm is subdivided into two stages: Firstly, a simple and fast clustering strategy – based on edge collapsing that computes a cluster hierarchy. Secondly, improving this hierarchical structure, the next stage visualizes the clusters with nested implicit shapes.

T. Soni Madhulatha [7] explained clustering is a descriptive technique. The solution is not unique and it strongly depends upon the choice of analyst. It described how it is possible to combine different results in order to obtain stable clusters, not depending too much on the criteria selected to analyze data. Clustering always provides groups, even if there is no group structure. When applying a cluster analysis we are hypothesizing that the groups exist. Clustering results should not be generalized**.**

Madhuri V. Joseph et al [8] described a numerous key data mining techniques that have been developing and used in data mining projects. These include statistics, association, classification, clustering, prediction, sequential patterns and decision tree. Basically Data Mining techniques classified into two categories based on their evolution. A comparative study to analyze the similarities and differences between these distinct approaches namely Classical and Next Generation techniques. The basic methods covered here on Classical approaches are Statistics, Clustering and Nearest Neighborhood and that of Next Generation are Trees, Networks and Rules.

III. PROPOSED METHOD

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence. This algorithm typically determines all clusters at once.

A. *Multilink Constrained K-means algorithm*

The multilink constrained K-means algorithm assigns each point to the cluster whose center also called centroid. The center is the average of all the points in the cluster is coordinates the arithmetic mean. It is one of the methods used to partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining.

Step1: Each instance di is assigned to its closest cluster center.

Step2: Each cluster center Cj is updated to be the mean of its constituent instances.

Step3: Distribute all objects to K number of different cluster at random.

Step4: Calculate the mean value of each cluster, and use this mean value to represent the cluster.

Step5: Re-distribute the objects to the closest cluster according to its distance to the cluster center.

Step6: Update the mean value of the cluster.

B. *Structure of Proposed Work*

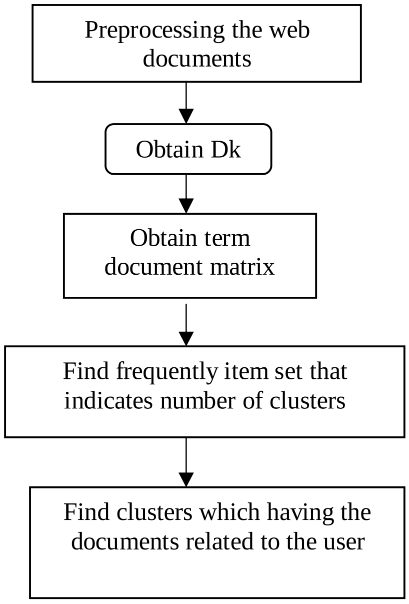

Preprocessing the web documents
Obtain Dk
Obtain term document matrix
Find frequently item set that indicates number of clusters
Find clusters which having the documents related to the user

Fig 2. The entire flow chart of the proposed work

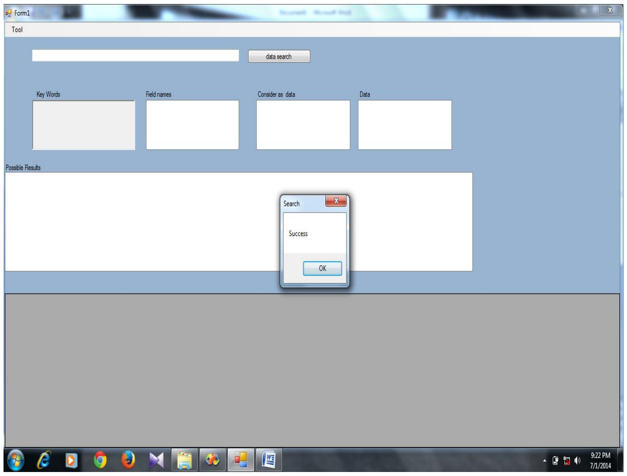
C. *Comparison Analysis*

TABLE I
Parameters comparison

Parameters	**Eclat Algorithm**	**Constrained K-Means Algorithm**
Time	Execution time is more as time wasted in producing record at every time.	Execution time is small than Eclat algorithm
Technique	It uses depth first search approaches and uses intersection of transaction ids list for generating record item sets.	K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.
Memory Utilization	Require large amount of records are produced so require large memory space.	With a large number of variables, K-Means works faster than hierarchical clustering and need small memory space.
Results	The result is suffers from the number of inefficiencies.	Give better results when data set are distinct or well separated from each other.

IV. DEMONSTRATION OF RESULTS

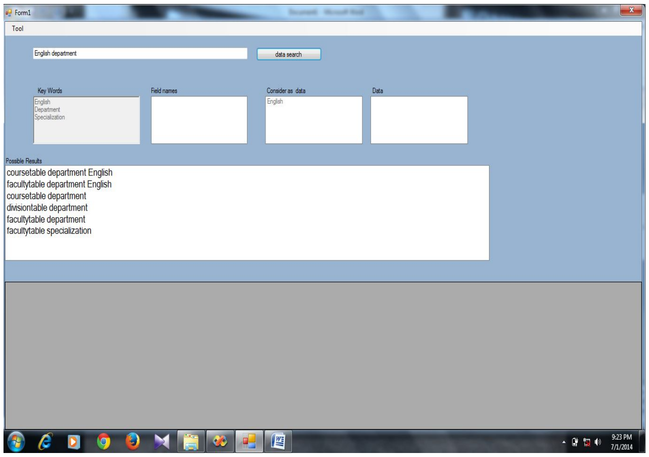
A. *Data Base Selection*



- Enter the data base table using tools selection.
- When the base selected, the success window displayed.
- The total number of records in the dataset is about 1000, which holds the course details, faculty information, subject wise details and relations between the datasets.

D. *Information Retrieval using K-Means Clustering Algorithm Data Base Selection*

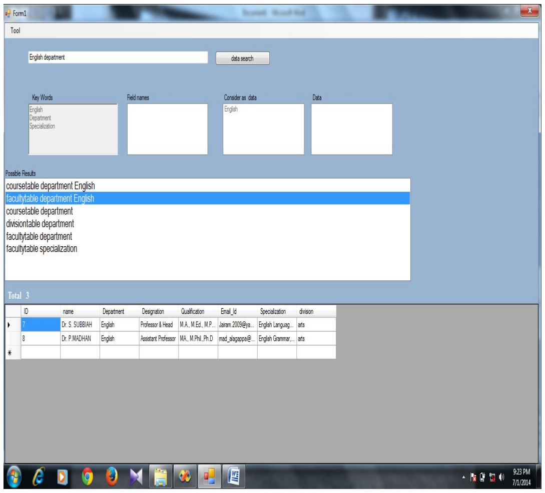
- The obtain the information possible query keywords should be used
- The results were calculated by the clustering k-means algorithm



coursetable department English
facultytable department English
coursetable department
divisiontable department
facultytable department
facultytable specialization

Fig. 1 Example of an image with acceptable resolution

- When user click some extra fields on the possible results table, table 3 clustered the data base and precede some most viewed results.



coursetable department English
facultytable department English
coursetable department
divisiontable department
facultytable department
facultytable specialization

Total 3

	ID	name	Department	Designation	Qualification	Email_Id	Specialization	division
▸	7	Dr. S. SUBBIAH	English	Professor & Head	M.A., M.Ed., M.P...	Jairam.2009@ya...	English Languag...	arts
	8	Dr. P.MADHAN	English	Assistant Professor	MA., M.Phil.,Ph.D	mad_alagappa@...	English Grammar,...	arts
*								

TABLE II

COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING ALGORITHM WITH THE PROPOSED SCHEME FOR COURSE DATASET

Factors	**Eclat Algorithm**	**K-Means**	**Multilink Constrained K-Means**
Database 2	Faculty Table	Faculty Table	Faculty Table
Records	1100	1100	1100
Average Length	28	25	20
CPU Utilization	8.68%	5.36%	2.14%
Memory Utilization	0.06%	0.04%	0.02

TABLE III

COMPARATIVE ANALYSIS OF VARIOUS CLUSTERING ALGORITHM WITH THE PROPOSED SCHEME FACULTY DATASET

Factors	**Eclat Algorithm**	**K-Means**	**Multilink Constrained K-Means**
Database 1	Course Table	Course Table	Course Table
Records	200	200	200
Average Length	33	28	23
CPU Utilization	9.74%	5.01%	3.56%
Memory Utilization	0.04	0.03%	0.02%

V. CONCLUSIONS

Clustering lies at the heart of data analysis and data mining applications. The ability to discover highly correlated region of objects when their number becomes very large is highly desirable, as data sets grow and their properties and data interrelationships change. The biggest advantage of the k-means algorithm in data mining applications is its efficiency in clustering large data sets. However, its use is limited to numeric values. In this proposed method the university data are used to perform the k-means clustering algorithm to retrieve valuable information. Compared between the Eclat algorithm, k-means algorithm produce the exact results on this information retrieval analysis. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.

REFERENCES

[1] Bellot, P., & El-Beze, M. (1999). A clustering method for information retrieval (Technical Report IR-0199). Laboratoire d'Informatique d'Avignon, France.

[2] Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000).

[3] Constrained k-means clustering (Technical Report MSR-TR-2000-65). Microsoft Research, Redmond, WA.

[4] Cardie, C. (1993). A case-based approach to knowledge acquisition for domain-speci_c sentence analysis. Proceedings of the Eleventh National Conference on Arti_cial Intelligence (pp. 798{803). Washington, DC: AAAI Press / MIT Press.

[5] Ferligoj, A., & Batagelj, V. (1983). Some types of clustering with relational constraints. Psychometrika, 48, 541{552.

[6] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall.

[7] Lefkovitch, L. P. (1980). Conditional clustering. Biometrics, 36, 43-58.

[8] MacQueen, J. B. (1967). Some methods for classi_cation and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability (pp. 281{297). Berkeley, CA: University of California Press.

[9] Marroquin, J., & Girosi, F. (1993). Some extensions of the k-means algorithm for image segmentation and pattern recognitionAI Memo 1390). Massachusetts Institute of Technology, Cambridge, MA.

[10] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.