A comparative study of IE Techniques and their algorithms

Mahmood Ashraf, Qazi Qayyum, Shazad Hussain
Department of Information Technology Bahauddin Zakariya University Multan Pakistan

Abstract— Information extraction from databases is a key research area and this area is familiar among the researcher. Data mining having great importance for the organization and it provides the reason to increase the profit of the organization. So data mining is an interesting field for the researcher. Many application are using to provide the facilities for data mining .hence data mining techniques are needed to determine the customer habits, according to business point of view, these techniques are used for the improvement in applications that provide the facilities for data mining, these applications may be internet online services and data where housing etc. These applications provide the new openings for the business and also increase its profit. Due to such demand this article is written on one information extraction technique clustering with its algorithm comparison and information extraction mechanism is presented.

Index Terms— Information extraction, clustering, clustering algorithm, K-Means Clustering, Hierarchical Clustering, DB Scan Clustering, Density Based Clustering, OPTICS, EM Algorithm.

1 INTRODUCTION

In recent past, our abilities of data[1] gathering and collection have been expending quickly. The most frequently utilization of scanner tags for most business items and the numerous business and the government exchanges, are being computerized, and the advancement in the tools that are used for data gathering, these are all the responsible to make the data huge. A huge number of databases have been utilized as a part of government administration, engineering data management, scientific data management, business management and many other different applications. It is noticed that the quantity of these types of database continuously increasing very fast due to the accessibility of capable and reasonable database systems. This unstable development in the database and data has created the needs and requirements of new tools and techniques that can be efficiently handle the information into knowledge and significant information. As a result data mining gets importance in the area of research.

Data miming some time called knowledge discovery from database. Data mining is a process of extracting the important information from the databases that are not known and potentially valuable data before, high level and valuable information, interesting knowledge, regularities can be mined from the given or relevant datasets, from the data base and it is examined from various points of views. So in this way huge data is consider a reliable source for data verification and generation of information, such extracted information can be used for query processing, information management, for decision making and also used to control the process, it can also be used in various other applications. Moreover developing applications giving the services for example www and online services and these services needed different techniques for extracting the information, these services provide an easy way to measure the behavior of the user. It helps to make the services better and opening the new door for the business.

Due to such importance and demand this article is written on data mining and its technique. There are various types of methods and techniques using to extract the information from the

Databases, such as classification, clustering, association rules, characterization etc. There are two ways to implement the information extraction learning one is supervised and the 2nd **is** unsupervised. [10]

Where as a **supervised learning** is a process in IE where learning starts from the known data labels, this label data is used to make a model then we it is used to predict the target class for the input data. All

the regression and classification algorithms are studied under the supervise learning ("e.g. logistic regression, decision trees, K-nearest neighbors, support vector machines, Naïve Bayes, Random forest, Linear regression, Polynomial regression, SVM for regression etc".). In **unsupervised learning** the training data is not giving at the start, learning starts from the unlabeled data to differentiate the input data, all the algorithms of clustering are come under the umbrella of unsupervised learning(e.g. K-means clustering, Hidden Markov Models and Hierarchal clustering etc.). So in this respect clustering is an unsupervised learning technique of data mining. In this article Clustering and clustering algorithms are discussed in detail.

2 Data mining Mechanism.

Information extraction mechanism is a method in which [2]information are structurally extracted from a database. There are different phases to store the data in the database and also used different ways to mine this information. Here simple architecture is discussed that explains the mechanism of information extraction from the database. At the first stage corpus processing done and then store this type of data to the database, it may be parse tree database (PTDB) and the information extraction process can be seen from the parse tree query language (PTQL). The parse tree query language (PTQL) judge mentor convert this query into keyword based queries and SQL queries, which are judged by IR engine and RDBMS. An inverted index is created by the index creator for corpus. This index helps the IR engine to evaluate the query and take part as a part of the query evaluator.

Interface is provided for the user to put the query and get the result, this interface having to input modes.

1) Extraction mode for specified query
2) Extraction mode for pseudo relevance feed back

2.1Extraction mode for specified query

In the 1st extraction mode user can easily and directly place the query for information extraction process while the interface provide the facility to provide key word based queries against such types of queries information retrieval search engine retrieved the most appropriate information or sentences from database. Query generator helps to uncover the patterns that are related to grammar by looking the hierarchical top ranked words that are automatically increase the queries that are based on primary keywords and then create a query or parse tree query language query, after the evaluation of query, results are presented to the user. Mover explainer module is available that explain the source of the result against query by displaying the acceptable formats of the sentences provided in the extracted information. This provides the guidance to the user how to improve his query to extract the demanded information.

2.2 Extraction mode for pseudo relevance feed back

At the 2nd extraction mode automatic analysis a method is provided. Manual part of the relevance feedback automates here. By doing the performance of retrieval process is improved without any extension in interaction. Extraction pattern that is based on the key word base queries can be express by adjusting the value of" m" user can be able to achieve the optimize recall or precision extracted results. Information extraction mechanism is illustrated in fig.1

3 Data mining Techniques:

For data mining from databases there are different kinds of techniques are available for different kind of knowledge. Recently there are different mining techniques and system are using. Data mining methods are classified and categorized on the bases of types of databases and types of information to be discovered. However some data mining techniques like association rules, classification, characterization, and clustering. In our research we will discuss clustering in detail with clustering algorithms by using wake tool. However different tools [11] can be used for information extraction process. Some of them are discussed under.

1) **NIME:** NIME is broadly and helpful data mining tool and having more than 1000 of various type of operators. This tool is very supportive for clustering incorporates Hierarchical, K-mediods, K-means clustering, Fuzzy, C-means and self organizing tree algorithm (SOTA).
2) **Orange:** orange data mining plate form is very simple. It uses the several operators. This tool is

helpful for SOM (self organizing maps), K-means, Hierarchical clustering and multi-dimensional scaling (MDS).

3) **RapidMiner Community Edition:** It is also an important tool used for clustering purpose. It is mostly used visual data mining tool. This tool supports the different algorithms of clustering like support vector clustering, Hierarchical clustering, Top down clustering, K-mediods and K-means etc.
4) **Tanagra:** It is also a tool for clustering used for variety of clustering methods like K-means, SOM, Hierarchical clustering and learning vector Quantizers (LVQ).
5) **Weka:** It is a most familiar clustering tool, whose algorithms are using for different toolkits e.g. RapidMiner. Weka supports the different algorithms like DBSN, COBWeb, EM and K-means clustering etc.

Fig: 1 IE Mechanism

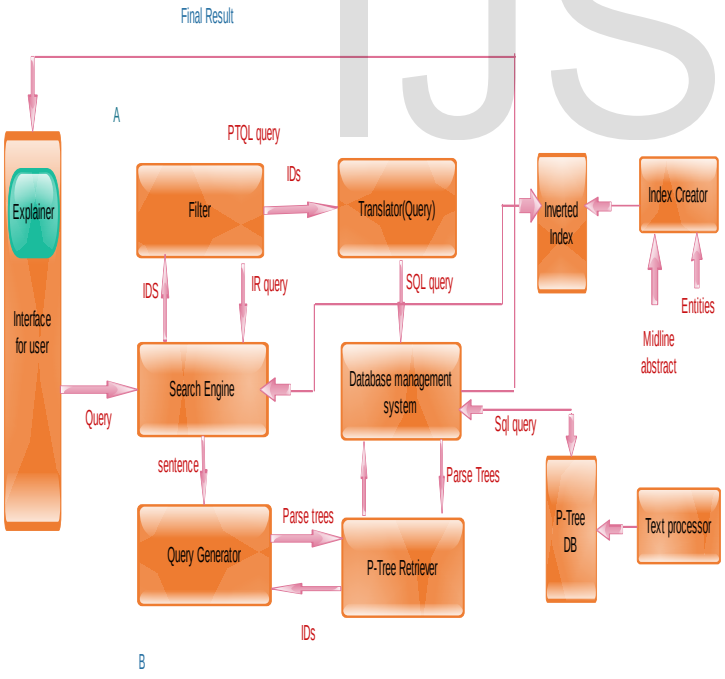

Fig:1 Information Extraction Mechanism

3.1 Cluster versions:

Cluster version provides [12] the function levels that are presents on the cluster. Versioning actually a technical term used for a special technique that permits the cluster to consist the nodes at various release levels and these nodes are completely interoperates by deciding the protocol level to be utilized for the communication.

Basically there are two versions of cluster

1) Potential cluster version
2) Current cluster version

1) Potential cluster version:

It is the most advanced level cluster version where the functions of given node is available. This version provide the communication mechanism and nodes are enabled to communicate with the other nodes of the cluster.

2) Current cluster version:

This version provides the facility that what is the current version being used for the operation of all the clusters. This version provides the communication among the cluster's nodes.

Fig: 2

3.2 Clustering

Clustering is the most important and familiar technique to extract the[3] information from the data base. In clustering groups are formed same type of data and different types of data in the form of cluster. Clustering is unsupervised data mining technique. Clustering deals by finagling the structure in a group and the data is unlabeled. Clustering organize the objects of same methods. So in this way we can say clusters are objects of similar clusters and dissimilar objects having other cluster.

Actually in clustering we put the similar data in one group of cluster and dissimilar data are grouped in other cluster. It is not an easy task to find and categorize these objects. Another important use of clustering is that to find the relevance data knowledge. For this purpose different clustering techniques are used and also presented in this research. Techniques are used to represent each cluster and cluster centers. Cluster center is helpful to find the heart of each cluster is present. So clustering techniques are used

and the cluster algorithms divide the data into clusters of given numbers Fuzzy and k-means etc.

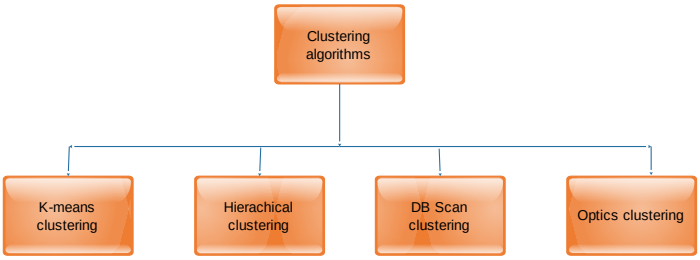

Fig:-3 clustering algorithms

3.2.1 K-means

It is a method for the analysis of cluster. The basic purpose of k means clustering to make n observations, in the form of k-clusters where every observation make the relation with the nearest mean cluster, so in this respect k-means called k-means algorithm where k is the cluster's number. The membership of the clusters can be determined by making[4] calculations of center of every group and the every object is assigned to the nearest centroid of the group. This technique is very helpful to minimize the dispersion within the cluster by reallocating the members of the cluster.

3.2.2 Hierarchical clustering:

This algorithm makes hierarchy or tree of the[3] cluster. It is also called dendrogram. It contains nodes and each node contains child clusters, these nodes are called parent nodes and the partitioning is done by covering those parents those are common.

Steps: from bottom to top.

1. Begin with one point
2. Relative clusters are added recursively, These clusters should be more than one
3. When k clusters achieved then stop adding clusters.

Top to down scheme:

When we are using top down approach then following steps are used.

1. take cluster
2. Divide the cluster into small clusters
3. When k-number cluster is achieved then stop.

3.2.3 DB Scan algorithm:

This algorithm works on the density[5] based clustering. In which data spaces are partitioned by lower and higher density of the objects in which regions they are placed. Maximum connected points in data spaces are known as clusters. For such type of clustering we use DBSCAN algorithm in which random shape clusters are discovered. So proper clusters, location[3] size and shapes of the clusters can be determined by using DBSCAN clustering techniques

3.2.4 EM clustering algorithm:

EM stand for expectation Maximization algorithm. It is a famous statistical based algorithm, which is based on distance. EM algorithm modeled the dataset like linear combination of different multiple normal distribution, the parameters of the distribution are determined by the algorithm, which increase the quality of measure of the model. It is also known as log-probability. This algorithm is used for clustering, Due to the following reasons.

a. It has solid statistical base.
b. This algorithm is suitable for noisy data.
c. Number of clusters can be putted according to the need.
d. It handles the large clusters
e. It provides good coverage and also provides good start.

3.2.5 Optics clustering algorithm:[2]

"Ordering points to identify the clustering structure (Optics) "it is also familiar clustering algorithm. When data mining is applied on spatial data, then a special algorithm is needed to mine the data. This algorithm was introduced by "Michael Ankerst, Markus M. Breuning, Hans-peter kriegel and Jorg sander. It is based on the idea of DBSCAN. This algorithm remove the one main drawback of the DBSCAN, in DBSCAN there is a hurdle of varying density, this problem is created when we are finding the clusters. This problem is minimized by using the Optics algorithm by ordering the points of the database in such a way that the spatially confined points are ordered to the closest neighbors, however special distance is maintained between the points which shows the density, and also verify the both points that they are the member of the same cluster and this is the requirement for clusters to be accepted. So in this respect optics reduce the need of density for the clus-

tering and it provides the spatial mechanism to make the clusters for data mining. So in this way there is no more interest in DBSCAN algorithm

4 COMPARISON BETWEEN DIFFERENT ALGORITHMS:

By making the comparison among the above studied algorithm, we can find the best one, a comparison table is constructed for this purpose. This comparison table will provide the short summary about the clustering algorithms and provide the information about the algorithm. However this comparison is taken on the following four factors.

a. Dataset size
b. Dataset type
c. Cluster numbers &
d. Software type

Table.1 comparison table among different clustering algorithms

ALGORITHMS	DATASET SIZE	DATASET TYPE	CLUSTERS	SOFTWARE TYPE
HIERARCHICAL CLUSTERING ALGORITHM	SMALL AND MASSIVE DATASET	RANDOM AND IDEAL DATASET	SMALL AND LARGE NUMBERS OF CLUSTERS CAN BE USED.	LNK NET PACKAGE AND TREE VIEW AND CLUSTER PACKAGE
K-MEANS CLUSTERING ALGORITHM	SMALL AND MASSIVE DATASET	RANDOM AND IDEAL DATASET	SMALL AND LARGE NUMBERS OF CLUSTERS CAN BE USED	LNK NET PACKAGE AND TREE VIEW AND CLUSTER PACKAGE
DBSCAN ALGORITHM	SMALL AND MASSIVE DATASET	RANDOM DATA SET.	FOR LARGE AND SPATIAL DATA[6]	R'S FPC PACKAGE[7]
EM-ALGORITHM	SMALL AND MASSIVE DATASET	RANDOM AND IDEAL DATASET	SMALL AND LARGE NUMBERS OF CLUSTERS CAN BE USED	LNK NET PACKAGE AND TREE VIEW AND CLUSTER PACKAGE
OPTICS ALGORITHM	SMALL AND MASSIVE DATASET	___	___	__

5 Data set for experiment

Data set is taken from the website https://www.data.gov/education/[8] to test the comparison among them and clustering algorithm. This dataset is about "Integrated postsecondary education". This data set contains the information about every university, college, technical and vocational colleges that provides the financial aids to the students, this data set consists of

- Year wise enrollment.
- Graduation rates.
- Programs.
- Staff and faculty
- Institutional prices
- Financial aids to the students.

Information about each educational institution in the 2013 IPEDS and universities contained by the file this contains, name, address, city, state and zip code, currently active or not.

Clustering algorithms are applied on the dataset and calculate the results.

Table: 2 clustering algorithm comparison table.

A quantitative comparison is taken among the different algorithms of clustering for this purpose above said data is taken and evaluation is made with the help of weka tool (3.6.9 v). Following results are taken that are presented in the given table Table-2 on the bases of these results we can judge the most efficient clustering algorithm.

Table: 2

NAME	NUMBER OF CLUS-TERS	NUMBER OF CLUSTER INSTANCES	ITERATION NUMBERS	ER-RORS	MODEL BUILT TIME	LOG LIKE-LIHOOD	UN CLUS-TERED IN-STANCES
EM ALGO-RITHM	2	68(68%) 32(32%)			0.52 SEC-ONDS	-7.28232	0
K-MEANS	2	3877(50%) 3892(50%)	12	118745.68767477981	1.14 SEC-ONDS		0
DBSCAN	0				37.15 SECONDS		7769
HIERAR-CHICAL CLUSTERING	—	—	—	—	—	—	—
OPTICS	0				34.59 SECONDS		7769
DENSITY BASED CLUSTERS		3954(51%) 3815(49%)		118745.68767477981	1.16 SEC-ONDS	-289.27245	

6 CONCLUSION

Information extraction or data mining is very fast[1] growing field. Recently new systems and t results of the research are presented in the field of information extraction. Researchers from all over the world have taken part and add their contribution in the research area of information extraction systems and methods.

In short it is a difficult task to explain the data mining methods comprehensively. In this article one reasonable method of data mining along with algorithms are discussed comprehensively. Moreover[9] we have discussed the information extraction frame work, how information are extracted from the database and also a comparison have taken among the different clustering algorithm so that we can pick the best suited algorithm, when we are using the clustering technique to extract the information from the database. A number of research hurdle have been acknowledged which are needed to windup clearly actual research inclines in the coming years. This research area is sensitive because on the base of mined information discussions have been taken so this way data mining plays vital role in the business, so highly skilled professionals are needed for making the analysis and having the complete knowledge about the data mining methods/ techniques and having the ability to understand the different factors on the bases of comparison have been taken.

References

[1] Chen, M.-S., J. Han, and P.S. Yu, *Data mining: an overview from a database perspective*. IEEE Transactions on Knowledge and data Engineering, 1996. **8**(6): p. 866-883.W.-K. Chen, *Linear Networks and Systems*. Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)

[2] Tari, L., et al. *GenerIE: Information extraction using database queries*. in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. 2010. IEEE K. Elissa, "An Overview of Decision Theory," unpublished. (Unplublished manuscript)

[3] Verma, M., et al., *A comparative study of various clustering algorithms in data mining*. International Journal of Engineering Research and Applications (IJERA), 2012. **2**(3): p. 1379-1384.C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992. (Personal communication)

[4] Abbas, O.A., *Comparisons Between Data Clustering Algorithms*. Int. Arab J. Inf. Technol., 2008. **5**(3): p. 320-325S.P. Bingulac, "On the Compatibility of Adaptive Controllers," *Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory*, pp. 8-16, 1994. (Conference proceedings)

[5] Gao, J. and S. Buffalo, *Clustering Lecture 4: Density-based Methods.*

[6] Wang, W., J. Yang, and R. Muntz. *STING: A statistical information grid approach to spatial data mining*. in *VLDB*. 1997.L. Hubert and P. Arabie, "Comparing Partitions," *J. Classification,* vol. 2, no. 4, pp. 193-218, Apr. 1985. (Journal or magazine citation)

[7] *Data Mining Algorithms In R/Clustering/Density-Based Clustering*.

[8] *Higher Education Datasets*. Available from: https://www.data.gov/education/.

[9] Williams, K., et al. *Scholarly big data information extraction and integration in the CiteSeer χ digital library*. in *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*. 2014. IEEE.

[10] http://dataaspirant.com/2014/09/19/supervised-and-unsupervised-learning/

[11] http://www.butleranalytics.com/10-free-data-mining-clustering-tools/

[12] https://www.ibm.com/support/knowledgecenter/en/ssw_ibm_i_7/rzaue/rzaigplanclusterversions.htm.