**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: 6.017**

*IJCSMC, Vol. 6, Issue. 1, January 2017, pg.234 – 237*

# Comparative Study of Algorithms used for Data Mining: A Review

**Jyoti Mahajan**

Assistant Professor, Department of Computer Engineering, Govt. College of Engg. & Technology, Jammu, India
jmahajan1972@gmail.com

***Abstract— Data mining actually refers to extraction of data in some regular patterns as desirable by the users. This deals with the discovery of patterns of knowledge from the databases. There are various algorithms which deal with the mining processes of the data i.e. each of the algorithms performs certain steps to preprocess the data from a data repository, convert it into useful information and finally transform it into an understandable format for the further use. The data mining process deals with full or partial automatic analysis of data by cluster analysis, anomaly detection and following association rules. The heuristic technique which make calculations and based on them form a model from given pattern of data is a data mining algorithm. This paper discusses about various algorithms along with the pitfalls that the algorithms suffer.***

***Keywords— K-mean Clustering; Ranking; Mining; Association Rules; Brute Force Technique.***

## I. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD)[1] an interdisciplinary field of computer science,[2][3][4] is the analytical process based on computational methods of discovering patterns from the data warehouses involving techniques from artificial intelligence, machine learning, operational research, business intelligence and research methodology.[2] It performs the analysis of raw data, pre-processing and mine the required data, form the model based on the data set, generate inferences from the given pattern of data, calculate the precision and accuracy of the generated patterns post processing and finally the online updating of the information.[2].The popular book "Data mining: Practical machine learning tools and techniques with Java"[5] was initially named "Practical machine learning", and the term "data mining" actual promoted the marketing.[6]. The general terms are “data analysis" or "analytics" but in actual methods it’s a technique of artificial intelligence and machine learning. The data mining process deals with full or partial automatic analysis of data by cluster analysis, anomaly detection and following association rules. For example, the data mining extracts multiple groups of the data from the repository which generates accurate prediction results by modelling it into a form of a decision support system. Neither the data collection and preparation nor interpretation of the data model nor the result generation and reporting are data mining steps but do add on to the knowledge discovery in databases. The heuristic technique which make calculations and based on them form a model from given pattern of data is a data mining algorithm. To create a model, firstly the data is analysed then the desirable patterns are extracted, the results obtained specify optimal for the generation of a model and finally when the parameters are applied to the model it generates certain patterns and statistics. The results of a mining model can be in various forms as in form of clusters or in form of decision trees or as a mathematical model or a set of certain protocols for the results. All of the data mining algorithms can be programmed by APIs, or by using the data mining components in SQL Server Integration Services.

## II. ALGORITHMS USED

Various algorithms have been designed for mining the data and assigning ranks to the data sets from the datarepository. These algorithms can generate results in various forms as clusters, graphs, association rules, matrices and coordinate values.

### 1. Hits Algorithm

Hypertext Induced Topic Search was the algorithm given by Klienberg where the search is query dependent. [7]. There are two kinds of web pages – hubs and authorities. The HITS algorithm consider World Wide Web as directed graph G(V,E), where V are vertices and E are edges. The vertices are the pages and edges are the links between the pages. It has two steps: Sampling Step which forms a set of desired pages for the given user defined query and Iterative Step which finds the Hubs and Authorities using the output of sampling step. The edges forms the in degree which gives the authority pages in the graph and the out degree give the hub pages in the graph. In this algorithm the query is given according to the requirement of the user and the algorithm traverse through the base set S and then from the base set it forms the root set R by adding all the pages along with the OUT(R) and IN(R) forming an adjacency matrix L where the entry is 1 if the link exist between any two pages and 0 if the link does not exist. Then from the set R two sets are obtained set A and set H where $A = L^T.H$ and $H = L.A$. Then report the pages with largest value of x coordinate as authorities and the pages with largest value of y coordinate as hubs. HITS algorithm is query dependent so helps in easy retrieval of data but is sometimes slow in action, may lead to spams and also leads to topic drift. It has to restart the computation every time a new page is added.

### 2. Page Rank Algorithm

Page Rank algorithm is based on assigning priority or rank to each page. The pages are categorized are Very Relevant Pages (VR), Relevant Pages (R), Weakly Relevant Pages (WR) and Irrelevant Pages (IR).Very Relevant Pages (VR) are pages which contain the most relevant information in context with the query. Relevant Pages (R) are relevant butnot with important information about a query. Weakly Relevant Pages (WR) has the keywords from the query but no relevant information. Irrelevant Pages (IR) have neither the relevant information norkey words [7]. Page Rank algorithm is based on a Crawler- retriever the contents of the required query forming a set of relevant pages, Indexer- stores by giving indices to relevant pages, Ranker- assigns a priority or rank to each page and the Retrieval Engine- performs a look up in the rank indices based on highest rank value. Google is based on Page Rank algorithm. It is query independent algorithm so no such spam is generated and it is faster. It can rank the pages even in the offline mode. Due to lack of query dependence it favors pages accessed former and the latter pages have lesser number of links.

### 3. K – Means Clustering Algorithm

K-means clustering algorithm is based on the grouping of organized clusters [8]. It works on the principle of increasing the intra class similarities and decreasing the interclass dissimilarities. This technique is used in color pallets on old graphical displays, web mining, medical imaging and vector quantization. K-Means is an unsupervised form of learning it consists of attributes n- total number of data elements in a data set to be clustered, k- centroid for each cluster, i- number of elements in each cluster and d- distance vector to associate the data items in each cluster. Initially the k centroids are defined for each of the cluster. Retrieve each point from the data set and associate it to the nearest centroid. Recalculate the K new centroids so that the clusters formed contains approximately same number of elements. The data points in the data set are again rebound to the K new centroids formed. This way the process is repeated until there is no further change in the position of K centroids. The main objective is to minimize the objective function which actually calculates the distance vector. The distance vector is calculated based on Minskowski, Manhattan (Block) or Euclidian distance. This algorithm hence is of utmost importance where simplification and pattern detection are required. The pitfalls are that there’s no way to initialize it until k is known and is based on mean values so the outliers may lead to discrepancies.

### 4. Online Page Importance Computation

With the pitfalls of the HITS algorithm the OPIC gained importance. It’s based on easy computations whenever a new page is added. It uses less number of resources and whenever a page is visited at the same time it’s importance is calculated. The basic idea is that cash is initially distributed to all the pages in web equally. Any page we access we know about the pages its pointing to at no cost. The cash of the page is recorded in the history. Now, when a page i is being crawled i.e. it’s being referred to or accessed then its cash is equally distributed to all the pages to which the former page was pointing to and hence it’s added to the credit history of the page. Reset the value of cash of the former page i to 0. This is done each time we read a page. Thus cash and credit history are two parameters on which OPIC is based on. Cash is the sum of cash obtained by the page since the last time it was crawled and Credit History is the sum of the cash obtained by the page since the start of the algorithm until the last time it was crawled. Main memory stores the cash while the disk stores the credit. It provides enough information to compute the importance of the pages and dynamic graphs can be handled by this analytical model. It requires less storage resources, less CPU, memory and disk access and is easy to implement. [9]

### 5. Apriori Algorithm

Apriori algorithm was given by R Aggarwal and R Srikant who applied it on transactional databases for the extraction and mining of the relevant data by applying a brute force method. This algorithm is based on the association rules and support and confidence parameters. The association rule so formed can be Boolean or quantitative. In order to find the frequent item set check the minimum support threshold (minsup) as in case if the greater the support value of a data item from the minsup more frequent is the data item. For a given transaction T goal of the association rules is find all data items with support value greater than or equal to minsup and confidence value greater than or equal to minconf (minimum confidence value). Apriori uses a "bottom up" approach where each data item is taken and extended to form subsets of related items and then these subsets or groups of items are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count item sets efficiently. It generates candidate item sets of length K from item sets of length K - 1. Then it prunes the candidates which have an infrequent sub pattern. It actually follows a two-step approach joining and pruning. Joining includes the generation of the frequent item sets and pruning includes the formation of association rules to check for all the data items in the sets. [10]. Thus it actually follows a brute force technique. It's very cumbersome to retrieve the frequent data sets in case of large transactional databases which add to the pitfalls of Apriori algorithm. Also it's quite a time consuming process if there are large number of data items set with same support value as the algorithm will repeatedly scan the database.

## III.COMPARATIVE ANALYSIS

The algorithms are compared in the following Table 1

| Algorithm | Hits | Page Rank | K-Means | Opic | Apriori |
|---|---|---|---|---|---|
| **Mining technique** | WSM and WCM | WSM | Clustering | WSM and WCM | Brute Force Technique |
| **Working** | Compute hub and authority scores of n highly relevant pages on the fly. | Compute scores at indexing time. Results are sorted acc. to high rank of pages. | Compute position of data items based on the centroids of the clusters formed | Compute cash and credit history of all the recently crawled pages | Compute support threshold based on which association rules are generated |
| **Input parameters** | Backlinks, Forward Links & content | Backlinks | Value of k and data items | Contents and dynamically obtains all links | Minimum support threshold and contents |
| **Advantage** | Query dependent | Faster and does not support spam | Simple and easy implementation | less CPU, memory and disk access | Easy retrieval for frequent sets |
| **Disadvantage** | Topic drift and Efficiency problem | Query independent | Dependent on K | Lot many calculations need to be done | Time consuming and requires more space |
| **Complexity** | < O(log N) | O(log N) | O(log k) | O(1) | O(N) |

**Table 1: Comparison of various algorithms**

### IV. CONCLUSION

Data mining actually deals with a lot of algorithms and these algorithms can be implemented using various techniques of artificial intelligence which includes techniques like neural networks, fuzzy logic, Support Vector Machine, Bayesian networks, back propagation techniques and can be implemented using an API, a MATLAB tool, SPSS or a JAVA platform. Depending on the requirement of the user and the data to be mined any of the above algorithms can be used. The outcome of the above compared algorithms can be in form of decision trees, association rules or clusters.

# REFERENCES

[1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[2] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.

[3] Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

[4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

[5] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.

[6] Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project". Journal of Machine Learning Research 11: 2533–2541.

[7] Page Ranking Algorithms for Web Mining by Rekha Jain and Dr. G. N. Purohit International Journal of Computer Applications (0975 – 8887)Volume 13– No.5, January 2011.

[8] Washio T, Nakanishi K, Motoda H (2005) Association rules based on level wise subspace clustering. In: Proceedings. of 9th European conference on principles and practice of knowledge discovery in databases. LNAI, vol 3721, pp. 692–700 Springer, Heidelberg.

[9] S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A first experience in archiving the french web. ECDL, 2002.

[10] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.