Data Mining based on Hashing Technique

Krishan Rohilla[1], Shabnam Kumari[2], Reema[3]

[2,3]A.P., Department of CSE, Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India

[1]M.Tech scholar., Deptt of CSE, Sat Kabir Institute of Technology & Management, Bahadurgarh, Haryana, India

Abstract: Data Mining is an important aspect for any business. Most of the management level decisions are based on the process of Data Mining. One of such aspect is the association between different sale products i.e. what is the actual support of a product respected to the other product. This concept is called Association Mining. According to this concept we define the process of estimating the sale of one product respective to the other product. We are proposing an association rule based on the concept of Hardware support. In this concept we first maintain the database and compare it with systolic array after this a pruning process is being performed to filter the database and to remove the rarely used items. Finally the data is indexed according to hashing technique and the decision is performed in terms of support count.

Keywords: Apriory, Clustering, Hashing, Data mining Techniques, Decision Trees.

1. INTRODUCTION

Data mining refers to extracting or mining the knowledge from large amount of data. Data collection and storage technology has made it possible for organizations to accumulate huge amounts of data at lower cost. Exploiting this stored data, in order to extract useful and actionable information, is the overall goal of the generic activity termed as data mining.

*1.1.*How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

*1.2.*Elements of data mining:

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.

- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

*1.3.*Parameters of Data Mining:

Data mining parameters include:

1.3.1. Regression - In statistics, **regression analysis** includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables

1.3.2. Sequence or path analysis - looking for patterns where one event leads to another later event.

1.3.3. Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok).

1.3.4. Clustering - finding and visually documenting groups of facts not previously known.

1.3.5. Decision Trees – Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.

*1.4.*Levels of analysis:

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships

2. Architecture of Data Mining

To best apply advanced techniques, it must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large data warehouse.

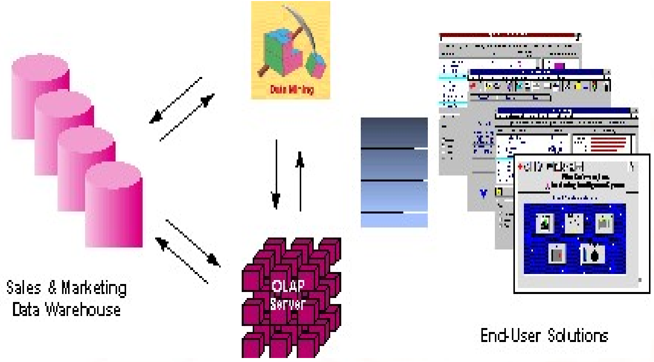

Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

3. Problem Definition

Current researches on data mining are based on simple transaction data models. Given an item set {item$_i$} and a transaction set {trans$_i$}, an association rule is defined as an implication of the form, X→Y, where X and Y are non-overlap subsets of {item$_i$}. In classification data set, an item can be viewed as {attribute, value} pair. Two important related quantities are confidence c, which is the percentage of transactions including X and Y to transactions including X, and support s, which is the percentage of transactions including X and Y to all transactions. Classification association rule (CAR) is then $X \rightarrow c_i$ where c_i is a class label. A training data set is such a set of data items that for each item, there exists a class label associated with it. A classifier is a function that maps attributes to class labels. In general, given a training data set, classification is to build a class model from the training data set such that it can be used to predict the class labels of unknown items with high accuracy.

*3.1.*Association rule Mining:

Association rule mining finds interesting association or correlation relationships among a large set of data items. It first discovers frequent item sets satisfying user-defined minimum support, and then from which generates strong association rules satisfying user-defined minimum confidence. The most famous algorithm for association rule mining is Apriori algorithm.[2] Most of the previous studies on association rule mining adopt the Apriori-like candidate set generation-and-test approach.Apriori algorithm uses frequent (k – 1)-itemsets to generate candidate frequent k-itemsets and use database scan and pattern matching to collect counts for the candidate itemsets. Recently, J. Han et al critiqued that the bottleneck of Apriori algorithm is the cost of the candidate generation and multiple scans of database. Han’s group developed another influential method for discovering frequent pattern without candidate generation, which is called frequent pattern growth (FP-growth). It adopts divide-and-conquer strategy and constructs a highly compact data structure (FP-tree) to compress the original transaction database. It focuses on the frequent pattern (fragment) growth and eliminate repeated database scan. The performance study by Han’s group shows that FP-growth is more efficient than Aproiori algorithm.[3]

*3.2.*Classification rule mining:

Classification rule mining is to build a class model or classifier by analyzing predetermining training data and apply the model to predict the future cases. Besides other techniques for data classification such as decision tree induction, Bayesian classification, neural network, classification based on data warehousing technology, and etc.The associative classification or classification based on association rules is an integrated technique that applies the methods of association rule mining to the classification. It typically consists of two steps:

3.2.1. The first step finds the subset of association rules that are both frequent and accurate using association rule techniques.

3.2.2. The second step employs the rules for classification.

4. Previous Work

Recent researches on the integration of association rule mining and classification rule mining. Recently, Bing Liu et al proposed Classification Based on Association rules (CBA) algorithm as an integration of classification rule mining and association rule mining.[4]

The integration was done by finding a special subset of association rules called class association rules (CARs) and building a classifier from the CARs. The main strength of CBA algorithm is its ability to use the most accurate rules for classification, which explains its better performance compared with some original classification algorithms such as C4.5. Liu's research group also proposed some methods to deal with the problems of the original CBA algorithm such as single minimum support and not being able to generate long rules for many datasets. The performance of the algorithm was improved by using multiple minimum support (S_{min}) instead of a single S_{min}, and combining CBA algorithm with other techniques such as decision tree method.[5,6] More recently, Wenmin Li et al critiqued some weakness of Liu's approach as follows: (1) simply selection a rule with a maximal user-defined measure may affect the classification accuracy, (2) the efficiency problem of storing, retrieve, pruning, and sorting a large number of rules for classification when there exist a huge number of rules, large training data sets, and long pattern rules. They proposed a new associative classification algorithm: Classification based on Multiple Association Rules (CMAR). The experimental result shows that CMAR provides better efficiency and accuracy compared with CBA algorithm. The accuracy of CMAR is achieved by using multiple association rules for classification. The efficiency of CMAR is achieved by extension of efficient frequent pattern method, FP-growth, construction of a class distribution-associated FP-tree, and applying a CR-tree structure to store and retrieve mined association rules.[7] (Both CBA algorithm and CMAR algorithm will be discussed in detail later in the section of related work.)

5. Proposed Work

6.

In this research work we are proposing a new architecture for the association rule mining. The complete concept the proposed work is based on two main concepts

- Hash Based System
- Pipelined system

The system architecture is inspired from the hardware enhancement. As the architecture is followed by any hardware system same approach is being proposed in this work to find the association between the selling produces

The complete work is divided in 3 states:

- In first modules the data will be collected and stored into the hardware system. In this system the dataset is being compared with the systolic array.
- In the second module the pruning process will be performed. It is actual the filtration process to clear all such items that are not part of frequently used item list. We can setup the association rules based on the frequently selling items. If some item is being sold rarely any need to establish any association rule onto it. This process will be done by Pruning
- In third stage, on the dataset collected from the customer transaction a hash table will be maintained. On the basis of this dataset the actual decision support will be calculated and the results will be derived

7. Conclusion:

DataIn this research we conclude that with the help of hash based pipelining technique products in market can be sold faster because in HAPPI technique it removes bottleneck problem thereby providing faster throughput and our sales process becomes faster because due to indexing hasing process becomes faster. Firstly items are kept in systolic array then items which are not in close proximity with each other are trimmed or removed from the filter then put into hash table filter so that duplication of items get removed so in this way. It solves our bottleneck problem

Acknowledgement

I would like to thank my guide Ms. Shabnam Kumari for her indispensible ideas and continuous support, encouragement, advice and understanding me through my difficult times and keeping up my enthusiasm, encouraging me andfor showing great interest in my dissertation work, this work could not finished without her valuable comments and inspiring guidance.

References:

[1] Xingquan Zhu, Ian Davidson, “Knowledge Discovery and Data Mining: Challenges and Realities”, ISBN 978- 1-59904-252, Hershey, New York, 2007.

[2] Joseph, Zernik, “Data Mining as a Civic Duty – Online Public Prisoners Registration Systems”, International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[3] Dr. Lokanatha C. Reddy, A Review on Data mining from Past to the Future, *International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011* [2]. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,From Data Mining to Knowledge Discovery in Databases, AI Magazine Volume 17 Number 3 (1996)

[4]http://www.slideshare.net/Annie05/sequential-pattern-discovery-presentation

[5]http://dataminingtools.net/wiki/introduction_to_data_mining.php

[6] http://www.dataminingtechniques.net

[7] http://www.slideshare.net/huongcokho/data-mining-concepts

[8] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf Retrieved 2008-12-17..

[9] “Data mining and ware housing”. Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5

[10] Weiyang Lin, Sergio A. Alvarez and Carolina Ruiz “Collaborative Recommendation via Adaptive Association Rule Mining” (2000)

[11] A Data Mining Framework for Building A Web-Page Recommender System

[12]Jorge, A., Alves, M. A. and Azevedo, P. “Recommendation with Association Rules: A Web Mining Application” in Proceedings of Data Mining and Werehouses, a sub-conference of information society 2002, EDS. Mladenic, D., Grobelnik, M., Josef Stefan Institute. (October 2002)

[13] Eui-Hong (Sam) Han and George Karypis “Feature-Based Recommendation System” Conference on Information and Knowledge Management **(**2005**)**

[14] Barry Smyth, Kevin McCarthy, James Reilly, Derry O'Sullivan, Lorraine McGinty and David C. Wilson “Case-Studies in Association Rule Mining for Recommender Systems” (2005)

Books:

[1]. Arun K. Pujari, *Data Mining Techniques*

[2]. Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques