FRIENDLY APPROACH ON BUSINESS INTELLIGENCE USING WEB DATA MINING

Jyosyula Satyavenu Gopala Krishna, G V S CH S L V Prasad, Shaik Shamshad Begam, Konda Sreenu, Donavalli Venkata Vidya Deepthi

1Associate Professor in Computer Science and Engineering Department, SirCRRCoE, Eluru, 2Associate Professor in CSE Department, Anurag Engineering College, Kodad(Md), Nalgonda
3Pursuing M. Tech from Andhra University, SirCRRCoE, Eluru, 4Assistant Professor in Computer Science and Engineering Department, SirCRRCoE, Eluru, 5Assistant Professor in Computer Science and Engineering Department, SirCRRCoE, Eluru, Email: 1jsvgk4321@gmail.com, 2zeeveeyes@yahoo.com, 3sreenukcupid@gmail.com, 4shamshad1284@gmail.com, 5deepthidonavalli@yahoo.com

Abstract—Data warehouse is a place where we store data which is collected from different sources like supermarket, hospital, airport, government office etc., This valuable information should be retrieved when necessary from the data tombs. Data mining (the advanced analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. For retrieving data one should use data mining techniques. There are lots of mining techniques available globally. The algorithms used here are used to display accurate results. This paper deals with business. Where business contains lot of data from years. Based on query requirement it should be mined and display accurate result. This paper deals with Query, Data mining, Database and algorithms for mining data.

Index Terms—Data Warehouse, Data mining, Database, Query.

I. INTRODUCTION

Data Warehouse

In computing, a data warehouse or enterprise data warehouse (DW, DWH, or EDW) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from multiple disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting.

The typical ETL-based data warehouse uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data. The Data Warehouse Architecture is shown as in Fig. 1.
A data warehouse constructed from an integrated data source systems does not require ETL, staging databases, or operational data store databases. The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. Unlike the ETL-based data warehouse, the integrated source data systems and the data warehouse are all integrated since there is no transformation of dimensional or reference data. This integrated data warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated source data systems.

Data warehouses can be subdivided into data marts. Data marts store subsets of data from a warehouse.

This definition of the data warehouse focuses on data storage. The main source of the data is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support (Marakas & O'Brien 2009). However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

Data Mining

Data mining (the advanced analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is discovery, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data
collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

A data set (or dataset) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. Nontabular data sets can take the form of marked up strings of characters, such as an XML file.

Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Six Classes and Tasks

Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors and require further investigation.

- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- Regression – Attempts to find a function which models the data with the least error.

- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Database Management Systems

A Database Management System (DBMS) is a set of programs that enables you to store, modify, and extract information from a database, it also provides users with tools to add, delete, access, modify, and analyze data stored in one location. A group can access the data by using query and reporting tools that are part of the DBMS or by using application programs specifically written to access the data. DBMS’s also provide the method for maintaining the integrity of stored data, running security and users access, and recovering information if the system fails. The information from a database can be presented in a variety of formats. Most DBMSs include a report writer program that enables you to output data in the form of a report. Many DBMSs also include a graphics component that enables you to output information in the form of graphs and charts. Database and database management system are essential to all areas of business, they must be carefully managed. There are many different types of DBMSs, ranging from small systems that run on personal computers to huge systems that run on mainframes. The following are examples of database applications: computerized library systems, flight reservation systems, and computerized parts inventory systems. It typically supports query languages, which are in fact high-level programming languages, dedicated database languages that considerably simplify writing database application programs. Database languages also simplify the database organization as well as retrieving and presenting information from it. A DBMS provides facilities for controlling data access, enforcing data integrity, managing concurrency control, and recovering the database after failures and restoring it from backup files, as well as maintaining database security. Example Table relation with other Table as shown in Fig. 2.
Data mining can also be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Information obtained – such as universities attended by highly successful employees – can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

Another example of data mining, often called the market basket analysis, relates to its use in retail sales. If a clothing store records the purchases of customers, a data mining system could identify those customers who favor silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical, or inexact rules may also be present within a database.

Market basket analysis has also been used to identify the purchase patterns of the Alpha Consumer. Alpha Consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society. Analyzing the data collected on this type of user has allowed companies to predict future buying trends and forecast supply demands.

II. EXISTING SYSTEM

An existing to preparing a data set for analysis is generally the most time consuming task in a data mining project, requiring many complex SQL queries, joining tables and aggregating columns. Existing SQL aggregations have limitations to prepare data sets because they return one column per aggregated group.

Disadvantages of Existing System:

1) Existing SQL aggregations have limitations to prepare data sets.
2) To return one column per aggregated group.

Previous Process Flow:
As shown in Fig. 3.

III. PROPOSED SYSTEM

Our proposed horizontal aggregations provide several unique features and advantages. First, they represent a template to generate SQL code from a data mining tool. Such SQL code automates writing SQL queries, optimizing them and testing them for correctness.

Advantages of Proposed system:
1) The SQL code reduces manual work in the data preparation phase in a data mining project.
2) The SQL code is automatically generated it is likely to be more efficient than SQL code written by an end user.
3) The data sets can be created in less time.

The data set can be created entirely inside the DBMS.

Proposed Process Flow:
As shown in Fig. 4.
Module Description:

Module 1 : Admin Module

Admin will upload new connection form based on regulations in various states. Admin will be able upload various details regarding user bills like a new connection to a new user, amount paid or payable by user. In case of payment various details regarding payment will be entered and separate username and password will be provided to users in large.

Module 2 : User Module

User will be able to view his bill details on any date may be after a month or after months or years and also he can to view the our bill details in a various ways for instance, The year wise bills, Month wise bills, totally paid to bill in EB. This will reduce the cost of transaction. If user thinks that his password is insecure, he has option to change it. He also can view the registration details and allowed to change or edit and save it.

Module 3 : View Module

Admin has three ways to view the user bill details, the 3 ways are 

i) SPJ
ii) PIVOT
iii)CASE

i) SPJ : While using SPJ the viewing and processing time of user bills is reduced.

ii) PIVOT : This is used to draw the user details in a customized table. This table will elaborate us on the various bill details regarding the user on monthly basis.

iii)CASE : Using CASE query we can customize the present table and column based on the conditions. This will help us to reduce enormous amount of space used by various user bill details. It can be viewed in two difference ways namely Horizontal and Vertical.

In case of vertical the number of rows will be reduced to such an extent it is needed and column will remain the same on other hand the Horizontal will reduce rows as same as vertical and will also increase the columnar format.

Module 4: Download Module:

User will be able to download the various details regarding bills. If he/she is a new user, he/she can download the new connection form, subscription details etc. then he/she can download his /her previous bill details in hands so as to ensure it.

IV. PROBLEM SOLUTION

The following example gives the working details of the problem. Carrefour S.A. is a French multinational retailer headquartered in Boulogne-Billancourt, France, in Greater Paris. It is one of the largest hypermarket chains in the world (with 1,395 hypermarkets at the end of 2009), the second largest retail group in the world in terms of revenue, and the third largest in profit (after Wal-Mart and Tesco). Carrefour operates mainly in Europe, Argentina, Brazil, China, Colombia, Dominican Republic, United Arab Emirates and Saudi Arabia, but also has shops in North Africa and other parts of Asia, with most stores being of smaller size than hypermarket or even supermarket. Carrefour means "crossroads" in French. Previously the company head office was in Levallois-Perret, also in Greater Paris. Fig. 5. shown log of Carrefour.

![Carrefour Log](image)

Fig. 5. Carrefour Log.

The first Carrefour store opened on 1 January 1958 in suburban Annecy near a crossroads (carrefour in French). The group was created by Marcel Fournier, Denis Defforey and Jacques Defforey and grew into a chain from this first sales outlet. In 1995 it merged with Promodès, known as Continent, one of its major competitors in the French market. Marcel Fournier, Denis Defforey and Jacques Defforey had attended several seminars in the United States led by "the Pope of retail" Bernardo Trujillo, who influenced them to move forward with Carrefour idea. The Carrefour group was the first in Europe to open a hypermarket, a large supermarket and a department store under the same roof. They opened their first hypermarket on 15 June 1963 in Sainte-Geneviève-des-Bois, near Paris in France. In April 1976, Carrefour launched a private label Produits libres (free products – libre meaning free in the sense of liberty as opposed to gratis) line of fifty foodstuffs, including oil, biscuits (crackers and cookies), milk, and pasta, sold in unbranded white packages at substantially lower prices.

In September 2009, Carrefour updated its logo. In May 2011, Carrefour reviewed its situation under conditions of stagnant growth and increasing competition in France from rivals including Casino Guichard-Perrachon SA, and planned to invest €1.5 billion ($2.1 billion) to change the supermarket with
the new concept of Carrefour Planet in Western Europe.

There are nearly 8000 Carrefour centers around the world. Carrefour announced membership cards to each family member who purchase goods in Carrefour. When the reach difference milestones of amount they are benefited with offers. Customer can use membership card in any center world wide. There are nearly 10 lakhs customers who visit and purchase goods in all centers. Among them 5 to 6 lakhs are members to Carrefour. Customer can use membership card any where in world. Customer who is having membership card can check the details of payment by login to website and card expires after 3 years of card date. Website logins are created by Administrator of Carrefour on receiving membership card. Customer can check his bill payment details every day or every week or every year or pervious year on purchase. Even customers are acknowledged before reaching milestones and after reach miles on payment. Each center is having its own database servers and all servers are interconnected with each other in distributed manner. After payment of bills customer can check the details of his payment amount total. Official website of Carrefour with membership option is shown in Fig. 6. Same as banking system, each card is unique. Purchase of goods at any center will be maintained at one server database.

Fig. 6. Official website of Carrefour.

User module deals with Fig. 7.

Fig. 7. User module with current amount paid details.

Module 3:

SPJ - While using SPJ the viewing and processing time of user bills is reduced. According Supplier-Parts-Products table we can left join tables that are having bill details of customer as shown in Fig. 8 and 9.

![Fig. 8. Consider 4 Tables.](image)

![Fig. 9. Joining Tables based on SPJ.](image)

Similarly, the joining of tables is done based SPJ. As example given in below Fig. 10.

![Fig. 10 For calculating Aggregate of Sum of Bill Amounts.](image)

Case is to get \( \text{SUM}(\text{Bill Amount}) \) purchased by the Customer in varies centers.

Variable \( X = \text{SUM}(\text{T1(Bill No1, ..., Bill Amount1), ..., Tn(Bill No, ..., Bill Amo\text{\text{ntn}}))} \)

Therefore Variable \( X = \text{SUM}(\text{Bill Amount1} + \text{Bill Amount2} + \ldots + \text{Bill Amo\text{\text{ntn}}}) \)

Finally, Variable \( X = \Sigma \text{SUM}(\text{Bill Amount}) \), so like that every day, every week, every month, every year or previous year bill should added.

PIVOT : This is used to draw the user details in a customized table. This table will
elaborate us on the various bill details regarding the user on monthly basis. Same as SPJ in above figures, a pivot should be maintained or constructed based on query by user or administrator. Pivot can be used for data mining the tables horizontally and vertically. Vertically direction searches that table which matches query based on calendar and horizontal search for Bill Amount in each table as shown in Fig. 11. Below Fig. 11 will make a pivot table with the matching of columns.

![Pivot Table](image)

Fig. 11. Pivot Table to get Table with match of Columns.

Last module here is download module. User will be able to download the various details regarding bills. If he/she is a new user, he/she can download the new connection form, subscription details etc. then he/she can download his/her previous bill details in hands so as to ensure it. Obviously, we are maintaining pivot table for each customer, on every payment of bill. Now customer can easily track his bill details whenever necessary to cross check. Not only cross check but also on no entry, customer can take a copy of bill to the office to update his bill payment.

V. CONCLUSION

Data mining is a concept where we require mining data for long years back. Various algorithms are available for mining. Modules give information how to mine data from years. Data mining techniques are expensive and suits large databases which contains years of information. This paper gives some techniques for sorting some problems. This paper deals with aggregate functions in SQL, where it successfully returns accurate results for the query. There is a scope to improve algorithms.

REFERENCES

[1] “Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis” by Carlos Ordonez, Zhibo Chen.

[2] “Approach for Developing Business Statistics Using Data Web Usage Mining” by G V S CH S L V Prasad, Malapati Sri Rama Lakshmi Reddy, Kuntam Babu Rao, Chodagam Suresh Kumar.

[3] “A Web Usage mining framework for Business Intelligence” by Tiwari. S.

[4] “Understanding the scope of Web usage mining & application on of Web Data” by Reddy K.S.

[5] “Data mining: an overview from a database perspective” Ming-Syan Chen, Jiawei Han and Yu P.S

[6] “Privacy-preserving association rule mining” Rakesh Agrawal and Ramakrishnan Srikant

[7] “Privacy preserving association rule mining in vertically partitioned data” Jaideep Vaidya and Chris Clifton

[8] “Privacy preserving association rule mining in vertically partitioned data” Jaideep Vaidya and Chris Clifton

[9] “Efficient data mining for path traversal patterns” Ming-Syan Chan, Jong Soo Park and Yu P. S

[10] “State-of-the-art in privacy preserving data mining” Vanilios S Verykios, Elisa Bertino, Igor Nai Foisuo, Loredana Parasiliti Provenza, Yucel Saygin and Yanniz Theodori dis.