Data warehousing and data mining

Abstract

This paper aims to discuss about data warehousing and data mining, the tools and techniques of data mining and data warehousing as well as the benefits of practicing the concept to the organisations. It also includes the trends and application in data warehouse and data mining in current business communities.

Keywords

Database, data warehouse, data mining, database management.

Introduction

Organisation uses information systems to record and retrieve data from daily transactions. The information systems via the database that link to it provides valuable data for making important and strategic decisions in regards to the well-being of a company. An organisation can predict the expectation that is yet to come from the data that they possessed. The data can also be used to provide possible solutions to overcome the problems that they faced, and even, they can use the data to obtain competitive advantage in their business environment. Database has reduces, if not in some place, vanish the old method of storing and keeping the information, that is, through the usage of the traditional filing system. The change towards digitization of data and the establishment of data repository has created a new term in the field of information systems, new position in the organisation, and a new way of doing business and daily transactions in human life.

This paper will discuss further about the two terminologies which is data warehouse and data mining from the perspective of database management in the organisation. At the same time, this paper will also include some cases and issues about data warehouse in the organisation according to real situation based on the literatures.

According to William H. Inmon, data warehouse is a set of integrated, subject oriented databases designed to support Decision Support Systems (DSS) functions, where each series of data is precise to some period of time. It is said that data warehouse contains atomic data and lightly conclude the data.

On the other hand, data mining is the search for valuable information in large volumes of data (Weiss & Indurkhya, 1998). It is the process of nontrivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques (Technology Forecast, 1997; Piatetsky-Shapiro and Frawley, 1991). As mentioned earlier, many organisations nowadays use computers especially through the usage of information system to collect particulars of business transactions such as records of banking operations, sales of retails, productions of factory, telecommunications and other transactions. Consequently the data mining tools are used to expose positive potentials and association from the data collected.

Background of data warehousing and data mining

The following part point up the historical evolution of the database and directly discuss about data warehouse and data mining. A brief history of data warehousing and data mining are included. Furthermore is the issues faced in the early years of implementing the concept of data warehousing and data mining and where both concepts are useful.

Data warehousing started in the late 1980s from the IBM lab and the responsible researchers are Barry Devlin and Paul Murphy. They started by the development of business data warehouse for decision support surroundings. In the early 1990s, it became a trend for organisations to meet the growing demand for organising information.

However Haisten (1999), a columnist for Information Management Website, mentioned that the concept of data warehouse take shape in early 1970s through a study that started out at MIT with the aim to provide optimal technical architecture.

And now, the next generation of data warehousing called Trend in Data Warehouse (TDWI) is mushrooming and become popular in many organisations that use information as their vital capitals.

The emergence of data mining began in the late of 1980s and it flourished by 1990s. There are three roots that can be traced back along three family lines on the origin of data mining, which are the classical statistics, artificial intelligence, and machine learning. In order to automate the process of extracting the data which are increased every single time, human has increased the power of computer and data storage. For that reason, the amount of data becomes huge and more complex. Primarily, Bayes' theorem (1997) and Regression analysis has identify patterns in data. The data mining is actually the process or method by using greater discovering in computer science engineering such as neural networks, clustering process, genetic algorithm and decision trees. Data mining can be said as a method to help with the collection of observation of behaviour.

Ayre (2006) stated in his paper that today’s data mining techniques is due to the work of mathematician, logicians, and computer scientist join together to create Artificial Intelligence (AI) and Machine Learning dated back from the 1950s. That was a very basic spark for data mining ideology. As mention earlier, in the 1960s, AI and statistic practitioners created new algorithm such as regression analysis, maximum likelihood estimates, neural networks, bias reduction, and linear model.

Also in 1960s, the field of information retrieval (IR) made its contribution in the form of clustering techniques and similarity measures. At these time techniques were applied to text document, but they would later be utilized when mining data in databases and other large, distributed data sets (Dunham, 2003).

In 1997, Connecticut-based Gartner Group report has mentioned about data mining and artificial intelligence are at the top five ranking of major technology areas that will clearly have a main crash transversely the whole scope of business unit within the incoming three to five years. Presently, data mining techniques and tools are being prolonged to the variety of areas. For instance, the data mining tools like intelligent text-mining system will extract the text waste pertinent to user queries.

The above is the process of how the data is transport to database and data warehouse and selection process by using data mining techniques and technology. And then it show us how the information form by the translating the data to be deploy in business.

Approaches of data warehousing and data mining in various industries

The industry of finance, sales and marketing, administration and others should see information as corporate source but the many local narrow systems that held that information simply did not give way the incorporated commercial viewpoint that was required. (Inmon, 2007)

Even though operational data is a greater asset to the organisation, it seemed data is usually not making use to its full capable. Therefore, data warehouse basically is to enable users' appropriate access to breaking apart and complete view of the organisation, supporting forecasting and decision-making process at the managerial stage. Additionally, data warehouse can achieve information consistency by carry data from dissimilar data foundations into centre of database. Users from different department for instances, can view the data from consistent single one place repository. The layer of data in data warehouse makes the information consistent by enable data around the data warehouse to be describe in business terms as against to using database terminology. The establishment of data that enforce how business terms are declared or calculated are also defined in the metadata layer and then served to the users. Because of the data in the data warehouse is non-volatile but it must be design to adapt the changes periodically. It is because terminologies use in business cannot run from changes.

Mannino and Walter (2004) in their study about the refreshment of data warehouse stated that data warehouse refreshment is a complex process comprising many tasks, such as extraction, transformation, integration, cleaning, key management, history management, and loading. This study is base on interviewed of 13 organisations and the author conclude that daily refresh during nonbusiness hours were the most common policy.

Sometimes data warehouse is not fully utilized by organisation or it being used by company but not all departments. In a case studied by Payton (2005) conclude that there are three factors why data warehouse is disappointed them. It is because; marketing’s lack of trust in the data in CDW (Corporate data warehouse); marketing’s low perceived quality of the data; and marketing’s perceived lack of incorporation of their needs in the design of the data warehouse and data warehouse interface.

Data mining in the industries like information provider as library involved in digital libraries gain benefits from it as they found the method to classify information automatically and apply new way to clustering the subject called MetaCombined the project. Besides database, data mining can be useful in a variety data types like text, spatial data, temporal data, images, and other complex data.

Data warehousing and data mining in telecommunication

The telecommunication industry is fast fitting the main user of high quantity information system. The problem faced by telecommunication industry is the generation of information which is too fast and in tremendous condition. The difficulties occur when a user, either a manager or high executive, needs access to stored information. If the time is not the issue to search what they want in that kind of stored data where they put in different places, it will not be an issue at all but time limitation is consuming. For instance, in order to produce a report regarding subscriber, an executive need to extract the data, do some analysis, and some other step to make it presentable to their officer. What else can enhance all this besides technology? The exact question to ask is; what is the technology that can be very helpful in this situation? The answer is through the application of data warehousing and data mining.

In real case studied by Papaiacovous, Bramblet, and Burgess (n. d) in a paper titled ‘ Data Warehouse: A telecommunication Business Solution’; they described about the difficulties to produce report. They then design personalized systems which exceed the traditional borders of data warehousing systems by assembling and keeping only important data, analyzing and transforming the data, and then summarizing and rearranging it in according to the demands of the user.

Another interesting article by Gomez (1998), expressed the hope that cellular companies and other communications firms to strongly consider data warehousing as a way to achieve competitive advantage. The author also reviews new way to data warehousing that have established successful in compliant concrete business benefits. Service providers realize due to the competition in the marketplace, they need to provide the best for their customer or risk to lose them. It is because customer can simply change their telecommunication service provider if they are not satisfied with their current provider. So the provider must get the knowledge in customer’s hand about what they want actually. After all the data about the customer are collected via online and phone survey, a data warehouse can enhance the executive to analyze and segments customer into groups by their product usage patterns, demographic characteristics, etc.

Telecommunications companies produce tremendous quantity of data. These data consist of call detail data, which describes the calls that cross the telecommunication networks; network data, which explain the position of the hardware and software components in the network, and customer data. Data mining can be used to uncover useful information buried within these data sets.

Telecommunication companies might counter fraud from customer that intends to use the service without paying for it. It happens when the users register and manipulate the registration information. The most regular way for identifying fraud is to construct a profile of customers calling behaviour and compare recent activity against this behaviour. Thus, this data mining application relies on deviation detection. The calling behaviour is captured by summarizing the call detail records for a customer.

Here is the issue on data mining. In the customer case study by the company ECtel n order to sell their data mining product for fraud detection called FraudView noted that selling data mining product to a telecommunication provider has been traditionally difficult because they don’t have data mining experts on staff who can work conventional data mining tools. Additionally, there are many ways to run away from paying for telecommunication services, from stealing phone card to bypassing phone circuitry. ECtel created FraudView, the solution that uses SPSS Inc.'s advanced data mining workbench, which enable the detection of telecommunications fraud in real time.

Data mining in telecommunication industries is not limited to detect fraud only but it also can be used as network fault isolation, marketing or customer profiling, etc. This is owing to the three main sources of telecommunication data which are call detail, network, and customer data.

Data warehouse and data mining in financial services

How a retail bank can truly understand and predict its customers' needs to the point where it can design product and services that suit those needs? One way of looking at customers can be from the standpoint of channel usage. In the UK's Llyods Bank/TSB merger, data were sourced from both their data warehouse, and then used to segment the customer base by service channel usage. Customers were allocated to segment on their usage of the following channels: ATMs, automated (direct debits/standing orders), cards (credit card and debit) and telephone (Peppard, 2000).

Financial institutions struggle with the large amount of data on every transaction deal. Data warehouse helps financial service organisations to analyse large, complex, and rapidly growing data volumes in a quicker way for better decision making and faster speed back to the market.

Fundamentals of data mining in finance are coming from the need to forecast multidimensional time series with high level of noise, accommodate specific efficiency criteria, make coordinated multiresolution forecast, and also incorporate a stream of text signals as input data for forecasting models (Kovalerchuck & Vityaev, 2002).

As noted by Kovalerchuck & Vitayaev, four main reason why data mining need to be implemented in finance is because the emergence of high volume databases such as commercial data warehouse and computer automated data recording; advances in computer technology such as faster and bigger computer engines and parallel architectures; fast access to vast amounts of data, and the ability to apply computationally intensive statistically methodology to these data.

Data mining is used to forecast the target variable, performing the contribution varies in percent within today's closing price and the price five days later, along with next day's prediction.

Data warehouse and data mining in health service

In healthcare there is not much transaction as business environment. The data is about outpatient, visit's to doctor office, procedure and so forth. Instead of numerical data, healthcare has textual description if the different medical counters. And there is a little bit problems here, where the technology that own a old method of data warehouse is created to manage process of transacting data that is very conquered by arithmetical information. When textual, non-transactional information is come across, the old method data warehouse technology nowadays is simply at a defeat to handle healthcare information. (Inmon, 2007).

Then, if the data is not a number but a textual; it must be kept with different understanding of phrase. It just likes a different language. In order to be standardized, there has to be creation of same vocabulary for instance, with the purpose to gain understanding for all. Then it can be kept in the data warehouse.

In a case study written by Kumar and Raval (n. d), they traced a large global pharmaceutical, which has a huge data of clinical trials for a number of drugs projects. Due to data collection and analyses operations that are broadening across the world, it is harder to implement data standards. Even harder to enforce was the programming and validation standards that are required of pharmaceutical companies. Primarily, a data warehouse is an operational middle ground and disparate and incompatible to a big quantity of systems put together to diverse collection from end user platform.

In another case, Whiting (2001) reported a healthcare name Intermountain Health that used data warehouse to make an analysis handling provided to its cardiovascular patients for five years. From the result, it improves service provided after the patients return home.

These are the data mining in healthcare and insurance where it can give beneficial such as providing claims analysis, it means determine which medical procedure are claimed together. It helps in predicting which customer will buy new policies and can identify behaviour pattern or risky customer and also prevent fraud.

Data warehouse and data mining in retail industry

The challenge in retailer business actually is inundate of data, the battle of data and expired data. To cope with these challenges, many retailers are building unified repositories of data known as data warehouse.

In the early implementation of data warehousing technology in 1990s, the retail business has gained benefits of practical data warehouse. From the daily historical sales reporting database created over past few years ago, retailer can expanded the use of analytical systems to support and produce vital decision.

The retail industry is going through a transformation. Data warehouse enable retailers to carry out on their major products, including activities such as inventory replacement, purchasing, and vendor management across multiple other multiple. Financial planning, adjusting for stock outs to seed a top-down financial plan provides all of the data necessary to support well-organized process for the confirmation of invoice accuracy to strategy-based pricing solution.

Simple application that can implement the concept of data mining for retail industries are SQL server 2008 and Microsoft Office Excel 2007. To stay competitive, retailer must understand not only current consumer behaviour but must also be able to predict future consumer behaviour. Accurate prediction and an understanding of customer behaviour can help retailers keep customers, improve sales, and extend the relationship with their customers. SQL server 2008 provide predictive analysis through data mining and Microsoft Excel 2007 offer data mining capabilities that can help retailers make better decision.

The application that is common for business retail in data mining such as market basket analysis, fraud detection, database marketing, sales forecasting, and also merchandise planning and allocation. Data mining is so beneficial in retailer industries!

Recommendations

In the business world a transaction is repeated again and again and many of them deal with data in numerical. The same activity repeats with different customers and different figures. To release from this mess, data warehouse and data mining provide solution. Even though data warehouse and data mining is a strategic investment to the business world but it can be risky without a proper understanding of the concept. Governance or control is important to support the implementation of data warehouse and data mining. There must be a proper standard to ensure compatibility in processing the data especially for textual data used in the health industry. There should also be a policy and to manage the data warehouse. It is highly recommended that to be successful in the implementation of data warehouse or/and data mining, an organisations are required to have extensive or comprehensive knowledge about the data in their company. This is to guarantee that a well structured data warehouse can be constructed. A well structured data warehouse consequently will help organisation to exploit via data mining the data that they have. Organisation should also know what exactly they want to implement in their organisation so that the right tools for data mining can be used. And finally, a strong support from top management is important to deploy data warehouse and data mining because the investment on these is not cheap.

Conclusion

Insufficient of data is no longer a trouble but lack of ability to breed valuable information from data is the issue today. The answer for those issues is through the implementation of data warehouse and the power to use data mining techniques and tools. Nevertheless, the realisation and the awareness of data warehouse and data mining in the organisation should take into consideration many aspects regardless of what industries. The aspects include support of the top management, understanding of the data needed by the organisation, governance and policy, the right design of the data warehouse, and the right tools or techniques for data mining.

Bibliography

- Dunham, M. H. (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.
- Kovalerchuk, B., & Vityaec, E. (2002). Data mining in finance advances in relational hybrid methods. USA: Kluwer Academic Publisher.
- Wang, J. (2003). Data mining opportunities and challenges. USA : Idea Group Publishing.
- Keng Siau. (2003). Advanced Topics in database research. USA : Idea Group Publishing.
- M. Kumar Sagar., & Raval, H. (n. d). Data warehousing in pharmaceutical and healthcare: an industry perspective. Retrieved January 10, 2010 from: http://www2. sas. com/proceedings/sugi24/Dataware/p115-24. pdf
- Mannino, V. M., & Walter, Z. (2006). A framework for a data warehouse refresh policies. Decision Support System, 42, 121-143. Retrieved January 10, 2010 from: www. sciencedirect. com

- Syncort Inc. (2010). Business drivers and enabling technologies for clickstream data warehouse initiatives [White Paper]. Retrieved from www. syncsort. com/clickstream
- Balog, K. (2004). An intelligent support system for developing text classifies. Retrieved January 10, 2010 from: http://balog. hu/itm/thesis. pdf
- Sang Jun Lee , & Keng Siau. (2001). A review of data mining techniques. Industrial Management and Data System. 101/1, 41-46. Retrieved January 10, 2010 from: http://www. emerald-library. com/ft
- Karthik Jayashankar. (2007). Data mining tools for analytics application in retail. Information Management Online. Retrieved January 10, 2010 from: http://www. information-management. com/white_papers/10000547-1. html
- Hackney, D. (1999). A data warehouse is subject-oriented. Are they any rules to go about defining the subjects? Information Management Online. Retrieved January 25, 2010 from: http://www. information-management. com/news/1331-1. html
- Adelman, S., & Moss, L, (1999). Data warehouse goals and objectives. Part 3: Long term objectives. Information Mangement Online. Retrieved January 25, 2010 from: http://www. information-management. com/issues/19991101/1564-1. html
- Bertman, J. (2005). Dispelling myth and creating legends for your e-biz intelligence warehouse. [Power Point Slides]. Retrieved from www. dgigusa. com
- Luja´n-Mora, S., Trujillo, J., & Il-Yeol Song. (2006). A UML profile for multidimensional modeling in data warehouse. Data & Knowledge

Engineering, 59, 725-769. Retrieved January 25, 2010 from: http://www. sciencedirect. com. ezaccess. library. uitm. edu. my/science? _ob= MImg&_imagekey= B6TYX-4HWXJXG-1-2R&_cdi= 5630&_user= 6533825&_pii= S0169023X0500176X&_orig= search&_coverDate= 12%2F31%2F2006&_sk= 999409996&view= c&wchp= dGLbVtz-zSkWA&md5= 35d7b25297f3ee013bded90b43ecf5bb&ie=/sdarticle. pdf

- Shin-Yuan Hung, Yen, D., C., & Hsiu-Yu Wang. (2006). Applying data mining to telecom churn management. Expert System with Application, 31, 515-524. Retrieved February 12, 2010 from: www. elsevier. com/locate/eswa
- Weiss, G., M. (n. d). Data mining in telecommunications. Retrieved February 12, 2010 from: http://citeseerx. ist. psu. edu/viewdoc/download? doi= 10. 1. 1. 60. 955&rep= rep1&type= pdf
- Lamont, J. (2000). Datawarehousing in the telecommunications industry. KMworld Magazine. Retrieved February 12, 2010 from: http://www. kmworld. com/Articles/Editorial/Feature/Data-warehousing-in-the-telecommunications-industry-9153. aspx
- Gomez, J. (1998). Data warehousing for the telecom industry. Information Management Online. Retrieved February 12, 2010 from: http://www. information-management. com/issues/19981201/260-1. html
- Papaiacovou, D., Bramblett, L., D., & Burgess, J. (n. d). Data warehouse: A telecommunicaitons Business Solution. Retrieved February 12, 2010 from: http://www2. sas. com/proceedings/sugi22/DATAWARE/PAPER135. PDF

- Thompson, B. (2005). Information and communications technology and industrial property. Journal of Property and Investment Finance, 23 (6), 506-5015.
- Peppard, J. (2000). Customer Relationship Management (CRM) in financial service. European Management Journal, 18 (3), 312-327.
- Rogers, G., & Joyner, E. (n. d). Mining your data for health care quality improvement. Retrieved February 12, 2010 from: http://www2. sas. com/proceedings/sugi22/DATAWARE/PAPER135. PDF
- Silver, M., Hua-Ching Su., Dolins, S. B. (n. d). Case study: how to apply data mining techniques in a healthcare data warehouse. Retrieved February 12, 2010 from: http://www. himss. org/content/files/jhim/15-2/him15208. pdf
- Bach, M., P., & Cosic, D. (2008). Data mining usage in health care management: literature survey and decision tree application. Med Glas, 5 (1), 57-64. Retrieved February 12, 2010 from: http://www. ljkzedo. com. ba/M8_10. pdf
- Inmon, B. (2007). Data warehousing in a healthcare environment. Administration Newsletter. Retrieved February 12, 2010 from: http://www. tdan. com/view-articles/4584
- McEachern, C., Stern, L, & Bell, L. (1998). Data warehousing in the health care industry – Three perspective. Information Management Online. Retrieved February 12, 2010 from: http://www. information-management. com/issues/19980301/696-1. html
- Whiting, R. (2001). Data analysis to health care's rescue. IT helps health-care group identify best clinical practices. Infrormation Week.

Retrieved February 12, 2010 from: http://www. information-management. com/issues/19980301/696-1. html

- Haisten, M. (1999). The next stage in data warehouse evolution, part 1. Information Management Online. Retrieved February 12, 2010 from: http://www. information-management. com/news/946-1. html
- Ayre, L., B. (2006). Data mining for information professionals. Retrieved February 12, 2010 from: http://techessence. info/files/Ayre_DataMiningForInformationProfessionals_June2006. pdf
- Ross, D. (2005). Retail data warehousing – the-state-of-the-art. BeyeNetwork. Retrived February 12, 2010 from: http://www. b-eye-network. com/view/769
- Adams, M. (2008). Microsoft SQL server predictive analytics for the retail industry. Retrieved February 12, 2010 from: http://74. 125. 153. 132/search? q= cache: kCA9HUfe0VcJ: download. microsoft. com/download/6/9/d/69d1fea7-5b42-437a-b3ba-a4ad13e34ef6/PredAnalyticsRetail. docx+Predictive+Analytics+for+the+Retail+Industry+SQL+Server+Technical+Article&cd= 1&hl= en&ct= clnk&gl= my
- Russom, P. (2009). Next generation data warehouse platforms. Retrieved February 12, 2010 from: http://download. 101com. com/pub/tdwi/Files/TDWI_BPR_NextGenDWPlatforms_Q409_r. pdf
- Payton, F., C., & Zahay, D. (2005). Why doesn’t marketing uset he corporate data warehouse? The role of trust and quality in adoption of data ware-housing technology for CRM applications. Journal of Business & Industry Marketing. 20 (4), 237-244. Retrieved February 12, 2010 from: www. emeraldinsight. com/0885-8624. htm