Data mining

Business

Introduction Data is information often in the form of facts or figures obtained from experiments or surveys, used as a basis for making calculation or drawing conclusions. Mining is the process of removing minerals or in some cases acquiring information from its source.

Therefore data mining is or sometimes called knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is technically the process of finding correlations or patterns among dozens of fields in large relational database. Usually this kind of data is stored in massive storages called warehouses. This is where sorting of the data and its organization is done (Alhaji, 2007). Problem in context and its Motivation I am researching about data mining because with the many case of fraud and unaccountability of many people, I saw a need to change the current situation and give solutions to these problems that seem to be here to stay.

This is one of the many problems that data mining can be useful. I found out in my studies that the internet is one of the areas that anyone can gather information. Without such equipment and their update, it can be very difficult to do research work. For instance, if the Wikipedia was not updated, we could be getting wrong information which may not be up to date. Therefore, this topic is not chosen at random but it came as a motivation because it has so many benefits that cannot be disputed. The fact that data mining can be used to in every sector and every area that you can think of is what motivated me to do a research on it.

Data mining is not something that you can dispute and it can be used by anyone anywhere to benefit himself and to grow his own business. It can be done in small scale or in large scale and being very economical; it’s a very useful tool for the prediction of the future with a high level of certainty. Objectives These are some of the objectives of this research a) To find the unseen pattern in large volumes of historical data which will help in the efficient management of an Organization. b) Prediction of unknown or future values of selected variables c) Description in terms of patterns that can be interpreted by human beings. Ways of Data Mining There are so many ways that data can be collected.

We can use observation, counting, use of questioners, interviews etc. whichever way you use randomness is required for the data to yield accurate/better results for any research work. Data mining can be translated to useful knowledge about historical patterns and future trends. For instance it can be used to calculate the demand and supply relationship or to see if the company is making any progress i. e.

if it is making profit or losses. Usually, enormous data is stored in data warehouses. From these warehouses, information can be retrieved or updated at any particular time whenever it’s required as long as you have authorization from the relevant authority. Information in the data warehouses is mostly organized into departments and in a chronological way for ease of access to a particular section that one needs to visit. Usually there are four categories of sorting /mining data: a) Classes This is data that is stored in predetermined groups.

For example in a restaurant, it can be used to determine when and what a customer usually orders. This is very important especially when selecting a special meal for the day. b) Associations Data can be mined to determine association to people or things or even departments. This makes it easy for accessibility and correlation between different items. This is also useful in modeling where parallel information can be used to come up with new models. c) Clusters This is when the data is grouped based on relationship or consumer preference.

Particularly in the supermarkets, items are arranged in such a way that those things that are closely related like the food stuffs is put in one section with drinks. This is very helpful when it comes to preservation matters. Foods stuff which is put near strong scented items like perfumes; they smell less fresh due to the strong scent from the perfumes. d) Sequential patterns This is simply done after a long time of observation and a trend or a behavior that can be predicted with a high level of certainty. This is what is called modeling.

This is one way of establishing trend automatically to solve different problems. Elements of Data Mining In data mining there are processes involved in order to ensure that the whole process is done flawlessly and in a very organized manner to ensure that the information stored is accurate and is maintained for further use. The first process is the extraction of data. This can be in form of research work in the field or in the laboratory or in the field like counting people in the case of census. This data is usually disorganized and not well arranged because in most cases it’s not done by one person and therefore the information has to be put together (Keong, 2006). It is usually referred to as crude data in terms of mathematics.

This kind of information is to be collected in a random manner to ensure accuracy and to make the data more precise. This data should not be rounded up because by doing so this can increase the error in the calculation and can give wrong results. Therefore it should be recorded in with keenness and with high precision to ensure accuracy of the projections or the analysis. After the data has been put together, it is now entered into the computer. This is the process called data entry. All this information is fed in the data warehouse.

The second step is to store and manage the data in multidimensional database system. This is simply the organization of data for easier manipulation and calculation. Basically, at this stage is where manual labor is required and it’s also one of the most important stages. This is so because; if wrong information is entered in the computer it would automatically yield wrong results. Hence high level of concentration is required just as much as in the first step. This is not only tedious but also expensive to a company since it has to employ many people to do the job.

This section is labor intensive and the only way to facilitate it with much keenness is by employing more people to do the job. If this is not done, one may end up overworking the staff and due to much stress they may end up entering the wrong figures. After accurately entering all the data in the system, it is then given to the analysts and information technology analysts who intern analyze the data using soft ware available depending on the type of data and the type of result they want to extract from the data. In this stage, we the company is required to have experts who are very good in their jobs to manipulate such enormous data to come up with accuate projections and patterns which can be used to solve other problems. That is, ones a pattern is established, and it can now be used to gather other important information which may be used at a later stage (Allen, 2008).

Finally the analysts present the data in a useful format such as a graph or a table for easier understanding and visualization or predictions. Sometimes the presentation is done using a power point presentation. This is usually done for easier visualization and interpretation of the data collected in the field. This is also done for easier understanding of those people who cannot be able to understand all the calculations done to come up with the outcome. This kind of presentation is done by the employee to his or her bosses as a final document on the findings and conclusions to the prior research. Also the analysts are able to give recommendations to their superiors on how, for instance, to reduce costs and increase returns/profits or give a better way of going about to yield better results for the growth of the company.

Data Predictions Prediction of data is usually done through mathematical calculations and laid out patterns that have been laid out for data prediction. This is actually the essence of this whole process and it is vital that the predictions be correct since from this finding, it gives the way forward for that company. And from such predictions the board can make decisions and draw out strategies to better the company. Also, these predictions allow the company to allocate fund to a particular department for the purposes of growth or making more profit. The major ways of automatic predictions that are commonly used are; a) Automated prediction of trends and behaviors This method is used to predict information in large databases.

Quarries that required extensive hands-on analysis can now be answered directly and faster. Such predictions may include forecasting bankruptcy, defaults and response of a population to a given event. b) Automated discovery of previously unknown patterns: Data mining tools checks and identifies hidden patters in the database for ease of comparison and projections. For instance; identifying products at the supermarket that are purchased together like steak and beer. Using data mining tools is when being implemented on high performance parallel processing, can analyze massive data within a split second making it very reliable.

This means that the user can experiment to come up with more models which can yield very useful information and enable him/ her to understand complex data. This is an added advantage since large databases produce much more accurate results. Other methods that are used in data predictions are; a) Decision trees These are tree shaped structures that represent particular sets of decisions. They generate rules for database classification. Such trees include the Chi Square Automatic Interaction Detection (CHAID) and Classification and Regression Tress (CART) b) Artificial neural networkThese are non-linear predictive models. They are characterized through training and they resemble biological neural network structures (Campilho, 2010).

c) Rule induction This is the extraction of useful patterns based on statistical significance. These statistical significances are calculated and then induced using mathematical calculations with the help of some important formulas. d) Genetic algorithms: These uses genetic combination, mutation and natural selection based o the concept of evolution. e) Nearest neighbor method This technique classifies each record in a dataset on a combination of the k record(s). Some statistical methods are used to establish such classifications.

They use neighborhoods methods to come up with such conclusions (Masseglia, 2008). How Data Mining Works Basically data mining is able to tell you important things that you didn’t know and it can be able to tell you what is going to happen in the future. This is what mathematicians call projection. Through the use of previous data and present information, one can be able to calculate and predict the future with a certain level of certainty. The technique that is used to perform these features is called modeling. Now, modeling is the act of building a set of mathematical relationships based on data from situations where the answer is known and then applying the model to other situations where the answer is yet to be established.

This technique of modeling has brought about better solutions to our day to day problems and it is way too faster than manually calculating variables. Time is money so we are told; hence by applying such techniques we can be able save time and subsequently save money. For instance if you want to establish the age bracket of those that usually purchase the most of your commodity not considering gender but considering where they come from, this can also be done. You can use the laid out patterns to do so by only considering the variables that you think can affect your outcome and those that don’t affect it you leave them. By doing all this you’ll be able to find your desirable results using the similarities in the previous models. Modeling is not only important to find out such information it can also be useful in determining your competitors and give to you a hint on where to establish a branch for your company and expect great income from those branches.

As the marketing director you have access to a lot of information like the credit history of your customers and your clients. This is very useful since you can be able to establish whether or not to give him/her a loan or not i. e. if you work in a bank. Such information is very vital and without it, you may cause the bank to become bankrupt or incur huge losses (Mayr, 2001).

Although the processing of data may be limited by the type of computer being used it is still very helpful in data manipulation. The more the data you have, the larger the computer that is needed so as to processes it. If at all the computer is small with a small processing unit, it can be very difficult to process faster all that data and in turn may cause overheating for the computer. For this purpose, it is recommended for huge data to be stored in large computers which are fit with large storage capacity and high processing unit. Mainframe computers are recommended for large amounts of data.

If you are dealing with low quantity data a personal computer can be used since it does not require much space and the processing unit is relatively low. Therefore the computer to be used is dependent on the amount of data that is to be calculated. Uses of Data Mining There is no particular area where data mining cannot be used. Due to its effectiveness and its efficiency, most institutions prefer this system. This helps to save time and money.

For instance, in pharrmaceutical institutions, it helps to ensure there is no order that is made twice on something that had already been ordered before. Some companies use it to know which area to invest more due to high income and which area to invest less money. Governments in particular, use it to estimate the growth rate in the country and to know how well to distribute its resources. This also helps to monitor the growth rate of a country and how it is faring in the economic sector. Per capita income is also calculated using models that are already established (Ordones, 1998). This is very useful since the survey is not dose year after year for the government to know the number of people in the country.

This is made easy since all the government has to do is to use the laid out patterns to establish the growth rate or death rate of its people. Some initiatives like jobs for the youth are initiated when the number of unemployed people reaches an alarming stage. This becomes so useful and it helps to counter even bigger problems like crime which may be as a result of unemployment. The rate of a disease spreading in a particular area is one of the applications of data mining modeling. It helps monitoring diseases and their effects in a particular area.

For example you can monitor the number of people that are affected by the HIV virus and those that have died due to the virus. Using this system one can be able to monitor the areas which are most affected and hence they can initiate a response by visiting such areas and making them aware of the disease and giving the alternatives or to enlighten them on the preventive measures that they should apply. Such information is very useful for any government especially when the budget is being read. This helps the government to allocate fund where it is more required (Mitra, 2003). Another advantage is that governments can monitor those people who don't pay taxes and some who may be using names of people who passed on long time ago. When such database is updated every now and then it makes it very easy to apprehend criminals that may be leaving a trail of crimes as they move along.

Because of such information being shared terrorist are unable to board trains, airplanes or enter government buildings. In the advertising arena, getting the right kind of commercials using customer relations and feedback is very vital. Data mining relies heavily on customer relations because that's where all transitional data comes from are it changing trends or satisfaction with products. For instance, direct visual advertising. Researchers found out that many men do their shopping mostly on Saturdays with their toddlers.

Hence most of them usually buy diapers for the little ones and then go for beer. This two do not have anything in common but as many supermarkets found this fact to be true, they shifted the beers close to the diapers and they recorded a rise in the sale of beers as well. There is also a very lucrative method for firms and individuals to pursue increased gains over their investments. Data mining helps to reduce the risk rate and helps to give the best returns possible. For actuaries, in order to find out all the insurance and investment risks, data mining is a perfect tool for this job (Perner, 2002)

Advantages of Data Mining There is no doubt that data mining has advantages and its disadvantages: Data mining is truly useful since one has to gather information once only without having to repeat it.

a) It helps save time and resources. The only time that we are required to go out again is when we need an update in our system. b) The labor involved is less tedious and faster. c) It helps in evaluating the recent health trends and putting forth the next health or diet product. Data mining has proved to be more and more essential when it comes to making cost-cutting medicine, optimizing surgical procedures and keeping a check over pharmaceutical sales. Data mining can be used to find out the success rate and side-effects of a medicine in the market.

d) Due to modeling, we are able to minimize the time of solving problems since we can use the previous pattern to solve the problem at hand. This saves us a lot of time making it more efficient and more reliable to use. e) It is effective and it is not limited for simple data or complex data. It is both large scale and small scale applicable. f) Archives can be researched through to find out the patterns in many fraud cases, whether it is forging money or major bank frauds. This reduces the number of frauds involved in banks (Perner, 2011).

g) The assembly line is another place that can benefit from data mining. You can use it to increase performance of automated tools as well as employee efficiency along with their welfare. Disadvantages of Data Mining There are as well several disadvantages of this process. Such include the following: a) Technical knowhow that is required is expensive. b) Data manipulation should be proportional to the size of the computer. The larger the data the bigger the computer required to process the data and the smaller the computer the smaller the data c) To set up the system can be quite expensive, especially when you employ people to do the research for you.

You also have to have experts to manipulate the data and bring forth useful information. d) Projections can be inaccurate if the data is wrongly entered Conclusion Data mining is useful as we have seen in the research and it is one of the engines that drive a company or a government forward. Through data mining we have observed that we can make predictions with high level of certainty and this gives us a guide on the direction we should take. We are assured of reduced levels of crime and terrorism through data mining thus making our environs safer to live in. By use of data mining, I am able to counter the problem of fraud as this is one of my motivations as yet at the same time making the world a better place by bringing in a solution to solve the problems at hand.

Therefore data mining should be highly regarded and considered in the job market and also in government institutions to ensure effectiveness in those institutions and making them more effective in their function. Also we have observed that the merit are way beyond the demerits hence ensuring that data mining is the way to go for anyone who want progress and bring forth good returns in his/her own business. I think that data mining is worth the research as it will enable people manage organizations efficiently and help them predict the unknown or future values of any selected variables. (Zighed, 2009).