Business Uses of Data Mining and Data Warehousing

MIS 304 Section 04 CRN-41595

Group 4

Kirk Bishop

Joe Draskovich

Amber Hottenroth

Brandon Lee

Stephen Pesavento

Submitted: 12/9/2008

Introduction-

Data Mining is the process of collecting and sorting through massive amounts of data in order to find useful bits of information that can be used by a business. In the following pages data mining and data warehousing will be explained and the many uses of these two subjects will be expanded upon. Additionally the history of data mining will be presented. Several of the most popular software products that are used for the collection and sorting of data will also be explained. Current issues and trends in data mining will also be addressed. Lastly a conclusion and analysis of data mining will be given.

Objectives-

The primary objective of this paper will be to explore the uses of data mining and data warehousing for business application. Within the exploration of these two related topics we will define data mining and data warehousing. Included in the analysis will be the advantages and disadvantages of data mining and data warehousing. We will also show how data is collected. Additionally we will explain how data mining is useful and necessary for business and the processes by which raw data is converted into useful information by a business. In all data mining and data warehousing will be defined, advantages and disadvantages explained and a broad overview of how data mining is used for business applications.

Scope of Project-

Data mining is the practice of searching through large amounts of computerized data to find useful patterns or trends (Merriam-Webster). A consortium of individuals, which utilize data mining techniques, developed the Cross-Industry Standard Process for Data Mining (CRISP-DM). The group comprised of representatives from DaimlerChrysler, SPSS Inc., and NCR Corporation. Their derived model consisted of six major phases of data mining.

The first of these phases is the process of understanding the business. The researcher undergoing this step needs to only focus on salient components of the project. Understanding the data is another crucial step in the process of data mining. The researcher needs to ask one’s self the following questions regarding the data gathered: What is the nature of the data? What is the quality of the data? Is there missing or miscoded data? Does the data relate directly to the goals set in the prior step? In order to effectively mine data the data needs to be sufficient. If there is missing data, or low quality, or inaccurate data, the process will not be accurate in viewing, or predicting any patterns. In the step of understanding data the researcher needs to either accept or deny the data as a whole before moving forward to preparing the data.

In the data preparation step the researcher has the opportunity to disregard data that isn’t going to be used. The analyst also needs to deduce which method should be utilized for dealing with missing data problems. In this step the data also needs to be organized in such a way that the computer can analyze the data in the most efficient manner.

Step Four is the process of modeling the data gathered and organized. This step has a plethora of techniques. The best method varies depending on the application, the data, and the amount of work the analyst wants to put in. A good model clearly summarizes patterns and relationships.

Evaluation is the next process in CRISP-DM. In the evaluation process the analyst determines if their model is sufficient. The sufficiency is measured based on a technical standard as well as suitability to the business problem. The technical aspects of the project can then be determined to need enhancement, which is done in this step of the project. To determine if the collected data and model meets organizational requirements an analyst may utilize graphical diagnostics.

The last stage is deployment. If data mining is to be useful to management research findings and models should accurately guide business decisions. Models can be implemented, via computer programs, for product customization and personalization, customer acquisition and retention, credit scoring, and pricing, among an enormous amount of other applications.

In order for data mining to be successful information needs to be collected. This is a crucial aspect of data mining, which can prove to be a huge undertaking. Fortunately, many companies have been gathering this data since the birth of data mining techniques. A user can gather this data through internal resources, as well as secondary information, accrued by others. When the proper amount of data is collected, CRISP-DM then goes into effect.   (Data and Text Mining, Miller)

An enormous database of information gives rise to a system highly susceptible to noisy, missing, and inconsistent data. Data mining techniques are not just encapsulated within the boundaries of analytics; they are also used in the preprocessing of the data in order to create a more fluent, efficient system. These processes are called data cleaning, data integration, data transformations, and data reduction.

Data cleaning helps to fill missing values, smooth noisy data, identify and remove outliers (data so skewed they fall outside the expected boundaries), and resolving inconsistency. Missing values can be filled in manually but this approach is very time consuming. A global constant can be used to fill the missing data (i.e. 'unknown'), usually this makes the data misleading because the system will determine that there is a pattern concerning all the missing data. A global constant is not recommended even though it is a simple fix. Another fix is taking the average data and replacing all the missing values with that average.

Data integration is the process of taking data from multiple sources and creating a single database with a standard protocol. Entity identification problems can occur in this process since different databases can use difference references and categorizations for the same entity of data. For example, customer_ID in one database and customer number can refer to the same data entity but may not be detected in the data integration process.

Transformation puts data into a format that is more advantageous to the user. Data transformation can include aggregation. For example: taking monthly sales and putting them into annual sales, if this provides a better outlook of the data being processed. Another transformation tool is generalization. An example of this is taking age and breaking it up into a broader category, such as young, middle aged, and senior.

The last process in data transformation is data reduction. Data reduction is a technique that takes a representation of the data in such a way that it doesn't retract from the integrity of the originated data. Data reduction techniques include but are not limited to: data cube aggregation, dimension reduction, data compression, and numerosity reduction (Data Mining: Concepts and Techniques, Han and Kamber).

The main function of data mining is the modeling and techniques used to discover patterns in the data. Regression modeling is the most commonly used method. The objective of regression is to predict a continuous response. Simply explained regression is the process of taking a seemingly skewed graph of data, and figuring out where the closest average line is, or a line in the direct middle of the data. Another tool used in data modeling is cluster analysis. Cluster analysis can be used to examine similarities and differences among customers. In other words, customers within a segment are more like one another than they are like customers in other segments. Clustering is the process of determining what those segments are and how they can be utilized to predict behavior. There are hundreds of techniques for modeling data. The two methods discussed above are the two most commonly used.

Data Warehousing-

Data Warehousing is the computerized warehouse where large amounts of raw data are stored. Its main function is to execute analysis and generate statistical reports. "Any Organization that is considering using a data warehousing system must decide if the benefits outweigh the cost (Exforsys Inc, 2008)."

Advantages-

There are many obvious advantages to having a data warehouse. The main advantage is that users are able access large amounts of information in order to solve different problems. Having a data warehousing will also create improved data consistency. Using consistent data, computing cost can potentially be decrease. By solving problems at a faster rate, the internal cost within a company would be reduced which will then cause productivity to increase. With an increase in productivity comes an increase in company profits. Another advantage to data warehousing is that data can be combined in one location from many different locations.

Disadvantages-

As with many advantages that data warehousing presents, there are many disadvantages that must be analyzed. Prior to imputing data into the warehouse, "it must be cleaned, loaded, or extracted", this could take up quite a bit time. Some say that this step actually accounts for 80% of the overall process. Another disadvantage is one that is highly variable. "Users who will be working with the data warehouse must be properly trained to use it" (Seifert, 2008) because if the data is not properly inputted into the warehouse, it will build invalid results. Another disadvantage of using data warehousing is that it is difficult to maintain.

Data Mining Disadvantages-

Some of the limitations of data mining are the very users that are responsible for the data that is inputted into the system. "Data Mining requires skilled technical and analytical specialist who can structure the analysis and interpret the output that is created.” (Seifert, 2008) The limitations lie not on the technology itself but are primarily data or personnel related. It is also difficult for the user to determine if the patterns and relationships hold value or significance. The validity of the patterns discovered is dependent on how they compare to ‘real world’ circumstances (Seifert, 2008).” Although behavioral connections can be observed, data mining cannot indentify key causal relationships. Many organizations have noticed unsuccessful implementations of various data warehouses because only top management would know how to input data and generate reports. “If this advanced level of analysis is reserved for the few, instead of for the masses, the full value of data mining in the organization cannot be realized (Basu).”

Advantages-

Data mining has many benefits fore companies large and small. One major benefit, and probably the most important aspect of data mining, is the fact that as a software application, it brings profitability to the company. The company is able to produce patterns and facts that were undetected before the use of data mining.

The second benefit of data mining is the fact that it saves time and money. Data mining can be carried out on existing software even if that software is being upgraded. This saves the company money from having to purchase new software in order achieve all of the benefits of data mining. Data mining also saves the company time from having to install new software.

Another benefit data mining gives to a company is the ability of providing better customer service. A company that utilizes data mining can aim at specific customers with the perfect deals or specials if they are familiar with the customers’ buying behavior.

With data mining, the company is able to study customers they have lost and reduce losing other customers who might have the same kind of risk.

Other advantages of using data mining in a company include the expansion of the business, improving sales effectiveness, reduction in costs, and increased marketing and sales within the company. An example of expanding a business with the help of data mining includes increasing customer base and inviting new customers by tracking characteristics of current customers. By leaning the patterns of the current customers, the business can better predict future customers. In order to minimize the costs of any company, the detection of fraud and abuse is very important. Data mining can use data patterns to better predict when there is or will be suspicious unseen costs. Knowing which promotions and special offers for which classification of customers is another important aspect in any business. With the help of data mining, customers can be sent certain promotions based on their past purchases making the company more effective in their sales.

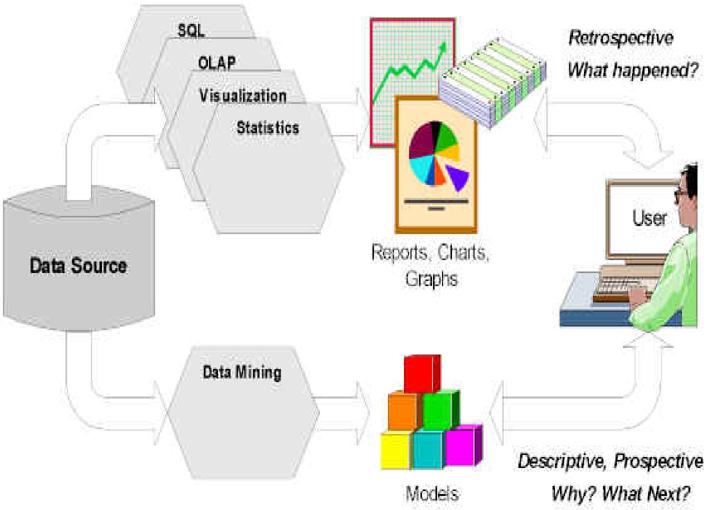
Some examples of benefits for any marketing or retail company can include the ability to forecast which customers will reply to mailing, recognize buying behavior patterns from customers, and finding similarities among customer demographics. With data mining a major advantage to a company is that it can identify cross-selling opportunities. Cross selling is the practice of up-selling a product to a person with a high probability of purchasing the product as a compliment to another item they are currently or previously purchased. Data mining also enables a company to improve customer satisfaction by tracking customer complaints as well as customer compliments. Another useful tool of data mining is the ability to determine the most profitable clients. A company can easily determine a client who spends the most money but doesn’t require too much investment to lure them into the purchase. One of the greatest benefits to company data mining is the enhancement of company marketing and sales practice. A company can easily identify marketing attempts that improved performance as well as identify attempts that were less successful.

In the banking industry, data mining can recognize stocks trading regulations from historical market information, and discover unseen connections between different financial indicators. Data mining also has the ability to identify fraudulent credit charge usage, classify the most trusted customers, and forecast customers that are likely to change their credit card company. Data mining can even assist the company in determining credit card expenses by customer groups.

In the medical industry, data mining has the benefits of distinguishing patient behavior to better see coming office visits and also spotting successful medical treatments for different health conditions.

As for the insurance and health care industry, some examples of the benefits data mining can give include claims analysis, or figuring out which medical practices are claimed together. This will make it easier and more efficient for the company to identify what is easiest and most suitable for the customer. Data mining also can calculate which customers will purchase new or different policies, classify performance patterns of risky customers and predict fraudulent behavior.

The transportation industry can also see benefits from using data mining. Some examples in the transportation industry are being able to find out the distribution schedules between outlets and examining loading trends. The company can benefit from analyzing loading trends because the company will learn and adjust themselves to working more efficiently around these loading patterns.



History-

Data mining was developed in the mid 1990's as computers changed the business landscape. Large-scale business experimented with a multitude of possible applications for the new machines. One goal they set before themselves was to utilize the massive amount of data compiled over the years. Up until the advancement of computer technology the data gathered by these companies were simply too large to properly analyze. Data mining is an analytical design process designed to explore data, usually of great quantity, in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data (Statsoft).

There are 3 stages to data mining: Exploration, model Building and validation, and deployment. Exploration is usually started with the preparation of the data, which typically involves cleaning data, data transformations, selecting subsets of records and preliminary feature selection (contingent on a large set of data points). The next step in stage 1 is determining the most relevant variables and to determine the complexity of the model which will be taken into account in the following stages. The next stage if model building and validation. This stage deals with making an expected model based on past data performance. There is an abundance of techniques that deal with this step:

Regression modeling is a technique, which utilizes standard statistics to data to prove or disprove a hypothesis. Each of these techniques falls into one of three categories: Clustering, classifying, and predictive.

Basically, regression modeling is comparing past performance based on a specific variable and directly comparing it to future performance using standard statistics. A second technique is visualization. In this technique a multidimensional graph is derived to help a user/analyst visually see trends. Another technique is correlation. Correlation identifies relationships between two or more variables in a data group. Another technique is variance analysis, which is a statistical analysis to identify differences in mean values between a known variable and nondependent variables. One classification technique is discriminate analysis. In this technique the data is used to identify the factors leading to membership within a grouping. Forecasting is one more useful technique, which takes past event outcomes to predict future possible outcomes.

The last stage is deployment. In this stage the model created in stage 2 is compared to future variables and outcomes are then predicted.

Software-

There are currently five major data mining software applications on the market. Microsoft Access gives a small business the ability to organize and manage data for data mining. It creates a variety of reports and forms in a database to connect with your customers and improve profitability. This software is well equipped to deal with a business growth. Another application is the FileMaker Pro. This software is a standalone solution. FileMaker Pro is in an easy-to-use format and also has an added security function which helps to safeguard your data. A third solution is the Lotus Approach. The Lotus Approach is part of the Lotus SmartSuite Package. This software program allows full management and analysis of data. This program is an alternative to Microsoft Access. The Corel Paradox is another application. Corel Paradox comes packaged in Corel WordPerfect Office. This option is designed and targeted for more advanced users but fits well with major growth companies planning an investment in technology for managing customer information. The last software is MyDatabase. It is a quick and simple solution for a company. It is ideal for one-person companies or freelancers. MyDatabase does not provide user programmability but does include a mass email function. This application is most suitable for a small business with plans to stay small.

Applications-

The concept of data mining can be used in all aspects of an organization for explanations and predictions. " Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales (Seifert, 2008)." Banks may apply the use of data mining when investigating fraudulent activity in customer's bank accounts or even when analyzing risk assessments like credit scoring. If data mining procedures are properly facilitated, organizations will have the ability to create decision models that help business decision makers predict future sales and customer incentives. Medical device companies may also apply this concept when analyzing the results of animal studies at college universities. Grocery Stores are highly dependent on its data mining techniques to analyze buying trends in the market in order to satisfy customer wants and needs. They most commonly use club card savings to input data into the data warehouse. This is an example of consumer choice modeling because “choice is fundamental to business, and models that predict consumer choice are of great value to management (Miller, 25).”

Data Mining can also be beneficial to direct database marketing because these firms only desire to contact those customers who are directly interested in that certain product. This strategy can increase profits by mailing only prospects that will most likely buy. Other marketing strategies that use data mining include market segmentation. Organizations can “identify groups or segments of customers who are more similar to one another than to customers outside their respective groups or segments (Miller, 26).”

Current Issues-

Recently data mining has played an increasing role in large companies. According to the CRS Report of Congress, there are several issues to consider related to implementation and oversight. These issues include data quality, inoperability, and mission creep. One of biggest challenges for data mining is the quality in which the data is processed. “The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data (Seifert, 2008)." For these reasons it is critical to have quality data in order to maintain accuracy and consistency of the data being analyzed. The issue of interoperability of different databases has become a popular topic among those who have attempted to implement data warehousing. "For data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously, and to help ensure the compatibility of data mining activities of different agencies (Seifert, 2008)." The third issue is mission creep, which "refers to the use of data for purposes other than that for which the data was originally collected (Seifert 2008)." Examples of mission creep have gained attention during the fight on terrorism. There has been a lot of pressure put on users who input data describing potential victims and allies while also putting more pressure on officials who analyze the data.

Research Methodology-

Although we originally expected to do almost all of our research on the Internet we were did find a significant amount of information about data mining and warehousing in books. Most of our technical and relevant information on the process of data mining came from books while information on software came from the Internet. In all both the Internet and hardcover books played an equally import role in our research of the subject.

Analysis & Results-

Now that we know how data mining works we can analyze trends and current issues affecting the data mining industry. These trends shape the uses and development of data mining and break the process down into pieces that can be improved. Since data mining is fairly new current issues pertain to the use of new techniques in various fields as well as recent legal implications for business and government use.

Trends-

There are four major trends that have the power to affect data mining and its benefits for a company. The first trend is the data trend. During the last two decades, the amount of information and data collected has grown extremely. While this is happening, the number of analysts to study and evaluate this data has been constant. With this comes the difficultly to sift through all of the information that is being produced.

Another trend is the hardware trend. The ability to analyze a very large load of data requires a very large quantity of memory available on the computer. The speed of the computer also can affect the performance of the outcome. This brings up the fact that the equipment being used needs to be up to date and have to ability to store and analyze the amount of information needed to produce results.

Next is the scientific computing trend. Data mining serves a significant role in connecting the three modes of science: theory, experiment, and simulation. When theory and experiment result in very large data sets, the simulation step is going to be that much harder to analyze due to the magnitude of data.

Lastly, business trend is the final trend that can have an effect on data mining. In today’s business world, organizations need to be at the optimum level of performance. They need to be faster, more efficient, cut costs, and produce higher customer satisfaction, or the company will not stand a chance of survival. Data mining brings along the benefits of foreseeing the risks, opportunities, and benefits for the company.

Data mining is a fundamental technology for any company in today's tough business world.

Current Development-

Data mining is a constantly developing process that is being used in new ways and in new fields all of the time. There is high potential for data mining to help various industries with due to its use with prediction theory as well as the heavy statistical analysis in clinical testing. The techniques used in data mining have recently been shown to be helpful by giving more extensive and accurate information to clinical testers and researchers.

In the field of clinical testing for diabetes new methodology is being discovered at the University of Virginia. One of these is based on the 'false discovery rate' statistical significance that controls the error rate of false-positives in large biological data analysis. These new approaches to genomic data mining can help researchers learn co expression patterns of multiple genes as well as prediction of subclasses of human diseases (Biotech Business Week, March 2008).

Data mining has also recently discovered information about diabetes that employers should be interested in. Due to the possibilities of data mining health plan sponsors can require more transparency of information and choose health care providers based on price and outcome instead of more arbitrary standards. For example, one company's health plan that serviced over 3,500 employees spent 12% of total cost on diabetes alone. Further analysis of this company showed that only 15% of diabetes patients received an A1c hemoglobin test – a test that manages blood sugar levels. These levels are so important that studies show that 1 percent point drop in A1c reduces risk of eye, nerve, and kidney related damages by 40% (Employee Benefit News, February 2008).

The new & upcoming development of data mining is part of the constant process of devising new statistical information that can help us further understand the real world. Right now one of the greatest places for the emergence of that development is in the medical field where there is motivation, money, and extremely large data warehouses.

Government Regulation-

Recently in the media there is a lot of talk about the government's involvement in data-mining; whether it's regulating what data warehouses can hold and be used for, or the government itself using information it's not supposed to. Since data mining is still a relatively new tool and people are just now becoming aware of the capabilities of it, the regulations of how it can be used are just now being set in place. One example of this is the use of data warehousing and mining as an anti-terrorism tool, and the possibilities that the information that such a tool would violate privacy rights.

On January 10, 2007 Ms. Leslie Harris, Executive Director of the Center for Democracy and Technology spoke to Congress about data mining, the possibilities it entailed, and the risks that it held. Ms. Harris spoke of the dangers of incorrectly identifying someone as a terrorist and the problem that would cause. She warned Congress of the risks of data mining and the need for closer oversight and control. Since data mining is a process that is still being developed to its fullest extent the government needs to be open to the possibilities of the new technology – both what it can give to a society as well as what it can take away. One of the main concerns Ms. Harris summarizes in her conclusion is the blurring of the difference between subject and pattern-based searches (CQ Congressional Testimony, 2007).

Unfortunately over the last couple years there has been a debate in Congress about the legality of the government's counterterrorism program, in which the government has grabbed millions of medical records through tapped fiber-optic lines. James Dempsey, vice president of the Center for Democracy & Technology, sees no reason for medical records to be used, as they are not nearly as helpful as travel records, communication records, and other information (Modern Healthcare, 2008). The information pulled by the government results in a "digital footprint" of individual lives, which causes privacy concerns. In addition, there is fear that the data mining that is used relies on incomplete data sets with untested and poorly focused initiatives for counterterrorism.

Conclusion-

The data mining industry is quickly growing and developing due to the large amount of information that it can supply. While setup costs restricts data mining to large companies and corporations the information provided is very valuable, due to the forecasting abilities of various techniques used. However to make best use of benefits of data mining the size and various information of a data warehouse has to be maximized.

While development for basic business and marketing purposes has already been researched, there is still growth in other industries. Since each separate industry needs to analyze a set of data differently each industry is creating a unique set of techniques to analyze data. The bottom line is that we are still learning the full potential of data mining and how to take advantage of the technology.

Recommendations-

Data mining and data warehousing provide companies and industries with great insight into customers and other aspects of many different businesses. If a company has the ability to implement a data mining program it should most likely do so as the benefits the company could reap from it could be extremely valuable. Overall data mining can be beneficial to a company immediately and also provide a predictive element that could pay off in the future as well.

Future Research-

Possible future research could include the negative impact that data mining and data warehousing have. The research could focus on the invasion of people's privacy from the storing and unauthorized use of personal information, buying habits and other useful information. The research could explain the financial costs involved with the exploitation of personal information that was gathered through unscrupulous methods. Other negative aspects that could be researched further might include identity theft, credit card fraud, and other related fraudulent activities that are results of certain types of data mining. Additional research could be conducted into the government's collection of peoples' information, travels, and phone calls. Research into the government's use of data mining could prove to be extremely sensitive and closely related to domestic security. Overall there are many possible areas of future research that could be conducted but they all seem to be of a negative nature.

Works Cited

"A brief history of data mining." Data-Mining-Software. <http://www.data-mining-software.com/data_mining_history.htm>.

Conn, J. (2008, October 13). Who's peeking? as the federal government continues to wage the war on terror, electronic medical records could be fair game. Modern Healthcare, 26.

"Data Mining Improves Decision Making." 2008. SPSS, A Predictive Analytics Company. 20 Nov. 2008 <http://www.spss.com/data_mining/>.

DATA MINING; Reports from university of Virginia, department of public health sciences add new data to research in data mining. (2008, March 17). Biotech Business Week, 2090.

"Data Mining: The Benefit." Data-Mining-Software. 2008. 15 Nov. 2008 <http://www.data-mining-software.net/data-mining-business-benefits.shtm>.

Graettinger, Tim. "Digging Up Dollars with Data Mining- An Executive's Guide." The Data Administration Newsletter. 1 Sept. 1999. 20 Nov. 2008 <http://www.tdan.com/view-articles/5263/>.

Greenfield, Larry. "Getting Started with Learning About Data Warehousing." 2008. LGI Systems Incorporated. <http://www.dwinfocenter.org/getstart.html>.

Grossman, Robert, Kasif Simon, Reagan Moore, David Rock, and Jeff Ullman. "Data Mining Research: Opportunities and Challenges." RGrossman. 1 Jan. 1998. <http://www.rgrossman.com/epapers/dmr-v8-4-5.htm>.

Grossman, Robert, Kasif Simon, Reagan Moore, David Rock, and Jeff Ullman. "Data Mining Research: Opportunities and Challenges." RGrossman. 1 Jan. 1998. <http://www.rgrossman.com/epapers/dmr-v8-4-5.htm>.

Kaplan, E. A. (February 1, 2008). Using data mining to deal with diabetes. Employee Benefit News.

"LyzaSoft." 2008. <http://www.lyzasoft.com>.

Palace, Bill, comp. "Data Mining." 1 June 1996. UCLA. 20 Nov. 2008 <http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm>.

Reed, Michael. "A definition of Data Warehousing." Intranet Journal. 2008. 15 Nov. 2008 <http://www.intranetjournal.com/features/datawarehousing.html>.

U.S. Government Data Mining. (2007, January 10). CQ Congressional Testimony,