Data Mining

Maen Caka and Steve Tipton

The amount of information in the world grows at an exceptionally fast rate. According to IBM, the world generates 2.5 quintillion (2.5×10^{18}) bytes of information each day (2012). This new information comes from a variety of sources, including social media websites, search engines, GPS devices, sensors (such as climate sensors or cameras), cell phone usage, and e-commerce. With the growth of content on the Internet, the number of users online climbed from 360 million in 2000 to 2.3 billion in 2011 (Roy 2012). When providing their statistics in 2012, IBM estimated that 90% of all data had been created in the past two years. In order to capture and analyze this constantly expanding information, the field of data mining has developed in tandem with the increase in information. In this paper, we will define the basic procedures of data mining, explore its uses and applications, and review ethical questions of privacy surrounding data mining practices.

The Basics

Data mining is one step in a larger process of information analyses known as Knowledge Discovery in Databases (KDD). When attempting to comb through data, the KDD process performs the following steps:

1) Data Selection: The first step is to understand the group of data to be analyzed. Taking the client's viewpoint into consideration and understanding the specific questions to be asked, an appropriate set of data or subset of data is chosen to undergo analysis.
2) Preprocessing: To prepare the data for analysis, the cleaning process decides how to handle statistical "noise" within the data and completes or accounts for missing data fields.
3) Transformation: At this stage, the data moves into the clusters or other patterns that will be analyzed in the data mining process. The analyst chooses the proper algorithms to perform on the data.
4) Data Mining: The data undergoes the mining process (see below).
5): Interpretation/Evaluation: Any useful information discovered through the data mining process is interpreted by the analyst and formatted into possibly actionable next steps (Fayyad 1996).

The KDD process does not start with a hypothesis to test; rather, it sifts the data to find patterns that are "unknown, valid, actionable, and understandable" (Khandar 2011). That is, we search for new and accurate descriptions of patterns in the data that follow a logical process and can be applied to newer data.

The data mining step of the KDD process applies the algorithms that parse the data to the database. In order to find the valuable information in the database, data mining uses "software techniques for finding patterns, regularities in the sets of data" (Srimani 2012). Its methods are an enlargement of statistical methods, and can include procedures such as sampling, estimation, hypothesis testing, search algorithms, modeling techniques, artificial intelligence, pattern recognition, and machine learning (Srimani 2012).

In traditional statistical analysis, researchers begin with a hypothesis and then perform tests to validate or invalidate their initial conjecture (Khandar 2011). In the descriptive wing of statistics, a theory is devised about a data set, samples are taken from the data, and tests are performed. Using the information from the test, statisticians can formulate inductive and inferential probabilities about the likelihood of future events (Zhao 2006).

The data mining approach is different from the traditional scientific method in that hypothesis testing is not its goal or approach. Because the data sets are extremely large and can be "messy," the data mining process looks for patterns within the data, which can lead to unexpected outcomes. This "serendipitous element" in the hunt reflects the "bottom-up" nature of data mining, where a large pool of information comes under scrutiny without a pre-formed hypothesis (Zhao 2006).

In fact, hypothesis testing is difficult for data mining for several reasons. First, the amount of data and the speed with which it arrives does not allow the time to devise hypotheses prior to analysis. Second, since the information gathered is usually related to social science, forming control groups to conduct hypothesis testing is difficult. Finally, data mining is less concerned with finding generalized models, than with individual outcomes and behaviors. This element of data mining proves especially useful in business applications such as customer relations and retention. Due to its break from traditional methods, data mining can sometimes be viewed in a negative light. Without a hypothesis, data mining can be called "fishing" and can seem to torture the data in order to provide results. However, as the business community adopts data mining techniques and begins to see increases in productivity and efficiency, data mining has grown from its loose beginnings into a more established and accepted field of study (Zhao 2006).

If data mining does not start with the traditional approach of hypothesis testing, what are some of its methods for discovering information?

1) Association Rule Learning: This method seeks connections between items in the data set, or associations that may signal elements' "co-occurrence" (King 2013). The analysis seeks to discover the patterns that relate the likelihood of one event happening in the data set with the likelihood of another event happening as well.
2) Classification: Before sorting the database, the analyst creates predetermined categories to divide the data. The database then sorts into these separate categories, and the analyst can attempt to determine the patterns that cause the various outcomes within the data.

3) Clustering: Clustering is similar to classification, but in clustering there are no predetermined categories. The analyst groups the data by each element's individual qualities, so that items that have more in common with each other are grouped together. Eventually, these groups of similar items will form clusters, and the pattern analysis can be performed on these clusters to determine their causes.
4) Regression: With regression, the analyst attempts to find a function to define the behavior of the data set. The function will describe the outcome of a dependent variable based upon the input of an independent variable. For example, given a scatter plot of data, the analyst will try to find the line of best fit for the data. This could be a simple linear function or a more complicated function such as an exponential or logarithmic function (Manyika 2011).

These methods are only a few of the basic building blocks of data mining; additionally, with the unending and increasing volume of data streams, new methods continue to be developed and refined.

The NSA

One of the largest ethical questions surrounding data mining is its use by government agencies to collect and analyze data. In June 2013, London's *The Guardian* newspaper revealed that the National Security Agency (NSA) requires Verizon to share its information with the U.S. government. A leaked court order sustains that the information collected is "telephony metadata"—not the content of the calls, but "session identifying information," such as the outgoing and receiving numbers, call length, and detailed call-routing information. Using this metadata, the federal agencies can piece together an individual's social networks and calling patterns (Greenwald 2013).

During the summer of 2013, further articles described the expanding reach of government programs into collecting and mining data. With information provided by former NSA contractor Edward Snowden, *The Washington Post* documented that the NSA and the FBI were able to access the servers of major U.S. internet companies in order to "[extract] audio and video chats, photographs, e-mails, documents, and connection logs that enable analysts to track foreign targets." Top secret government slides show how this data mining program, under the code name PRISM, partnered with Microsoft, Yahoo, Google, Facebook, PalTalk, AOL, Skype, YouTube, and Apple to amass information and create social networks of potential terrorist suspects (Gellman 2013). Though this computer analysis was originally limited to suspects who were not American citizens, an August 2013 article in the *New York Times* disclosed government documents detailing a policy change allowing the NSA to conduct " 'large-scale graph analysis on very large sets of communications metadata without having to check foreignness' of every e-mail address, phone number or other identifier" (Risen 2013).

Controversy remains over balancing the need to ensure national security and to protect individual privacy rights. In a press release (Clapper 2013) following the newspapers' revelations, Director of National Intelligence James R.

Clapper said that “information collected under this program is among the most important and valuable foreign intelligence information we collect, and is used to protect our nation from a wide variety of threats. The unauthorized disclosure of information about this important and entirely legal program is reprehensible and risks important protections for the security of Americans.” Conversely, the American Civil Liberties Union (ACLU) filed a lawsuit challenging the constitutionality of the NSA’s programs. Jameel Jaffer, ACLU deputy legal director, stated in a press release (ACLU 2013) that the government programs are “the equivalent of requiring every American to file a daily report with the government of every location they visited, every person they talked to on the phone, the time of each call, and the length of every conversation. The program goes far beyond even the permissive limits set by the Patriot Act and represents a gross infringement of the freedom of association and the right to privacy."

Business

Companies in a wide range of industries—including retail, finance, health care, manufacturing transportation, and aerospace— are already using data mining tools and other methods to take advantage of data. If the companies use pattern recognition, statistical and mathematical techniques to sift through warehoused information, data mining helps recognize significant facts that might go unnoticed. Most companies tend to use data mining to benefit themselves by making better decisions for the company by discovering the different kinds of patterns and relationships within the data. Data mining helps the company to spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty.

The technology of data mining enables companies who use it to focus on important information within data that has been collected regarding the behavior and purchasing potential of the customers. The amount of raw data that has been stored within corporate databases is exploding. With sources ranging from sales transactions to credit card purchases, databases are now being measured in gigabytes, terabytes, and even exabytes. As an example, Best Buy uploads 10 million point-of-sale transactions to a T-Mobile parallel system with 400 processors running a centralized database. But the raw data alone doesn't provide much information. Nowadays companies need to rapidly turn terabytes of raw data into significant insights into their customers and markets to guide their marketing, investment, and management strategies.

Data warehouses are used to consolidate data that is located in disparate databases. In other words, a data warehouse is able to store large quantities of data by specific categories, thereby making it easier for the users to retrieve, interpret, and finally sort the information. These warehouses have enabled executives and managers to work with a substantial number of transactions or other kinds of data, allowing corporations to respond faster to markets and make more informed business decisions. The drop in the price of data storage has given companies incentives to invest in data warehouses. It is expected that in the next ten years data warehouses will be used by every business around the world. Even companies who already use data warehouses need to get still more information about how to

improve knowledge of customers and markets. If companies don't continue to learn about new developments in data mining, then they will limit the potential benefits that they can derive from it.

Using massive parallel computers can allow companies to dig through high volumes of data to discover patterns about both their customers and products. Fast food chains are an example; McDonald's can analyze the factors that come into play when a person decides to order an entire meal instead of just a burger or sandwich. This information can be crucial, helping businesses to provide a wider selection of options to their customers. Apple, AT&T, and American Express are among the growing number of companies implementing data mining techniques for better sales and marketing. Doing so, these companies have been able to increase profit and gain a competitive advantage.

Conclusion

In summary, data mining is an interdisciplinary subfield of computer science. Through a computational process, patterns are discovered in large data sets and transformed into an understandable form for further use. In addition to the fields discussed here, in recent years, data mining has been used widely in science and engineering, in fields such as bioinformatics, genetics, medicine, and electrical power engineering.

References

American Civil Liberties Union, ACLU Files Lawsuit Challenging Constitutionality of NSA Phone Spying Program [Press Release]. Retrieved from https://www.aclu.org/national-security/aclu-files-lawsuit-challenging-constitutionality-nsa-phone-spying-program

Bradley, P. S., Fayyad, U. M., & Mangasarian, O. L. (1999). Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal On Computing*, 11(3), 217.

Clapper, J. R., DNI Statement on Recent Unauthorized Disclosures of Classified Information [Press Release]. Retrieved from http://www.dni.gov/index.php/newsroom/press-releases/191-press-releases-2013/868-dni-statement-on-recent-unauthorized-disclosures-of-classified-information

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

Gellman, B., & Poitras, L. (2013, June 6). U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program. *The Washington Post.* Retrieved from http://www.washingtonpost.com

Greenwald, G. (2013, June 5). NSA collecting phone records of millions of verizon customers daily. *The Guardian*. Retrieved from http://www.theguardian.com

Khandar, P. V., & Dani, S. V. (2011). Knowledge discovery and sampling techniques with data mining for identifying trends in data sets. *International Journal On Computer Science & Engineering*, 7-11.

King, B., & Satyanarayana, A. (2013, June 23-26). Teaching data mining in the era of big data. Paper presented at the 120th ASEE Annual Conference & Exhibition: Frankly, We Do Give a D*mn, Atlanta.

Manyika, J., Chui, M., Brown, B., Bughin, J., et al. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report.

Risen, J., & Poitras, L. (2013, Sept. 28). N.S.A. gathers data on social connections of U.S. citizens. *The New York Times.* Retrieved from http://www.nytimes.com

Roy, A. G., Chatterjee, A., Hossain, M. K., & Perrizo, W. (2012). Fast attribute-based table clustering using Predicate-Trees: A vertical data mining approach. *Journal Of Computational Methods In Sciences & Engineering*, 12S139-S145. doi:10.3233/JCM-2012-0444.

Srimani, P. K., & Patil, M. M. (2012). Massive data mining (MDM) on data streams using classification algorithms. *International Journal of Engineering Science & Technology*, 4(6), 2839-2848.

Zhao, C., & Luan, J. (2006). Data mining: Going beyond traditional statistics. *New Directions For Institutional Research*, 2006(131), 7-16. doi:10.1002/ir.184.

Nominating faculty: Professor Ashwin Satyanarayana, Computer Systems Technology 1204, Department of Computer Systems Technology, School of Technology and Design, New York City College of Technology, CUNY.

Cite as: Caka, M., & Tipton, S. (2014). Data mining. *City Tech Writer, 9,*25-30, Online at https://openlab.citytech.cuny.edu/city-tech-writer-sampler/