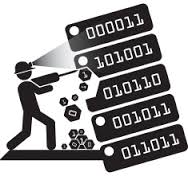
**An Overview of Data Mining**



A comprehensive look at Data Mining Architectures, designs, algorithms, and implementations

A technical paper by : Ryan T. Stearns  
 St. Cloud State University  
 Information Technology – Security

Table Of Contents



Definition……………………………………..2

History …………………………………………2,3

Implementation………………….…………3,4

Algorithms…………………………………..4

Big Data……………………………….………4

Results…………………….………………….5

Conclusion……………….…………………5

Bibliography……………….………………6

**Defining Data Mining**

One of the most prevalent terms being frequently discussed not only in computing, but also in any business that can benefit from advanced statistical analysis, is data mining. According to Jiaewi Han, data mining is the “Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amounts of data.” [2] Han continues to say that the term “data mining” is somewhat confusing because “the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.” [2] Han even suggested a more reflective name of the process would have been “Knowledge mining from data.” [2]

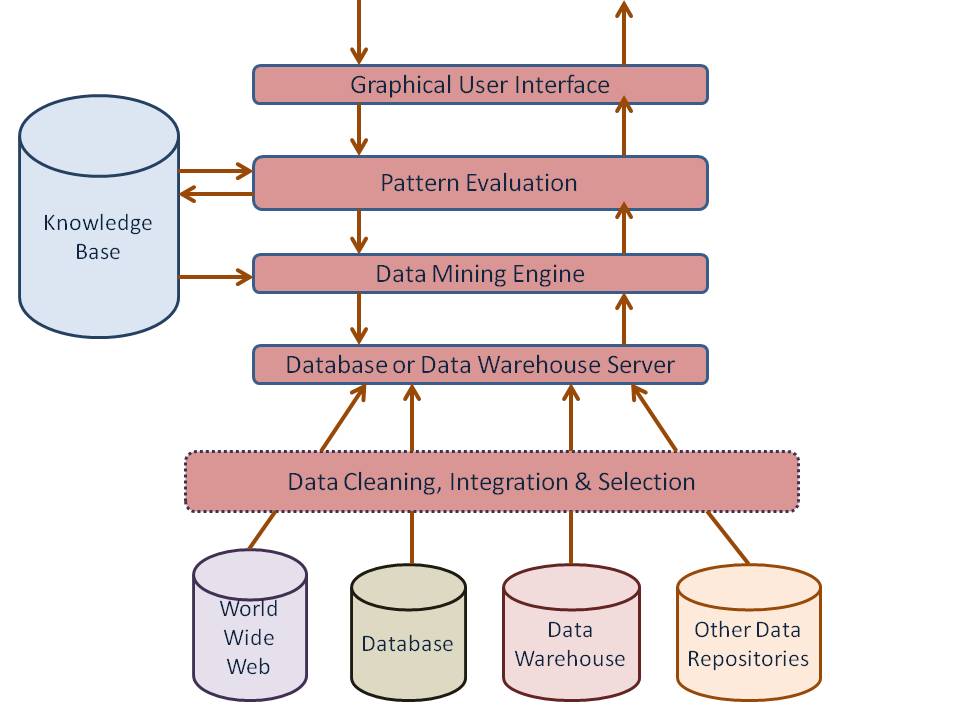
In a more technological definition, Oracle defines data mining as “the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis.” [6] Oracle continues to define it at as using “sophisticated mathematical algorithms to segment the data and evaluate the probability of future events.” [6]

**The History of Data Mining**

Theoretically, and mathematically, the theorem for data mining dates back to 1763, and was expanded upon in the 1805 when regression analysis was determined to estimate the relations among variables. [1] In 1965, the first company “specifically applying evolutionary computation to solve real-world problems” [5] was created, and was called Decision Science, Inc. By the 1970s, sophisticated database management systems were around, however, “extracting sophisticated insights from these data warehouses of multidimensional models were very limited.” [5] A very crucial book on genetic algorithms that would initiate the field of data mining study, was released in 1975 by John Henry Holland and called “Adaptation in Natural and Artificial Systems.” [5]

In the 1980’s a trademark is placed on the phrase “database mining,” which was coined by HNC for their product called DataBase Mining Workstation. It was also during this time in the 1980’s that “sophisticated algorithms can ‘learn’ relationships from data that allow subject matter experts to reason about what the relationships mean.” [5] By 1989 the term “Knowledge Discovery in Databases” (KDD) is created and a workshop of the same name is also created by Gregory Piatetsky-Shapiro. By the 1990s data mining was starting to become commonplace terminology in database communities. By 1993, Gregory Piatetsky-Shapiro, started the newsletter Knowledge Discovery Nuggets (KDnuggets). [4] It is during this period that “retail companies and the financial community are using data mining to analyze data and recognize trends to increase their customer base, predict fluctuations in interest rates, stock prices, customer demand.” [5] In 2001 the term data science, although existing since the 1960s, was introduced as an independent discipline. Presently, there is a Chief Data Scientist at the White House, and data mining is much more prevalent than ever before. Ray Li states that currently, “data mining is widespread in business, science, engineering, medicine, credit card transactions, stock market movements, national security, genome sequencing, and clinical trials,” [5] and that “terms like Big Data are now commonplace with the collection of data becoming cheaper and the proliferation of devices capable of collecting data.” [5]

**Data Mining Sample Implementation**

[9]

The World Wide Web, Database, Data Warehouse, and Other Data Repositories level “contains one or set of databases, data warehouses, spreadsheets, etc.” [9] The Database or data warehouse server level is responsible for fetching the relevant data, based on the user’s data mining request. [9] The Knowledge base level is responsible for Domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attribute values into different levels of abstraction. [9] The Data mining engine level is essential to the data mining system, ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, and evolution and deviation analysis [9]

The Pattern evaluation module component typically employs interestingness measures and interacts with the data mining so as to focus the search towards interesting patterns. It may use interestingness thresholds to filter out discovered patterns. [9] The GUI module communicates between users and the data mining system allowing the user to; specify a data mining query or task, provide info to help focus the search, perform exploratory data mining based on the intermediate data mining results, browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms. [9]

**Algorithms for Data Mining**

According to Raymond Li from KDNuggets, the same newsletter started by KDD founder Gregory Piatetsky-Shapiro, these are the 10 most popular algorithms in use for data mining. These “Top 10 most influential data mining algorithms as voted on by 3 separate panels.” These include C4.5, k-means, Support vector machines, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART. [4]

**Big Data**

A term very closely related to data mining is Big Data. Since the popularization of the internet, databases and data have been exponentially produced. SAS defines big data as, “a term that describes the large volume of data, both structured and unstructured, that inundates a business on a day-to-day basis.” [8] Similarly, IBM describes big data as “being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors, and mobile devices transmit it.” [3] With these extremely large data sets, it is no wonder that such complex assessments of the data must be conducted through data mining.

**Results from Data Mining**

As previously stated, the results of data mining are expected to produce previously unknown, and interesting patterns, such as, “groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining).” [7] Furthermore, “stored data is used to locate data in predetermined groups called Classes.” [7] Another type of data items are grouped “according to logical relationships or consumer preferences” called Clusters. [7] The data can be further analyzed to identify associations. Perhaps most unique to the results produced or interpreted is that of sequential patterns. Undoubtedly the most popular result is that “data is mined to anticipate behavior patterns and trends.” [7]

It is through this anticipation of behavior patterns, trends, and other information gained through data mining that enable business’ to “enable cost reductions, time reductions, new product development and optimized offerings, and smart decision making. It is also effecting businesses that are no realizing that instead of keeping their data discrete, “business and IT leaders must join forces to realize value from all data.” [3]

**Conclusion**

The history of data mining is quite extensive, it has long been realized that the amount of data both already available, as well as the data being produced on a daily basis needs to be further analyzed to gather valuable information. Big data sets have required complex algorithms to analyze the type of useful unknown knowledge which data mining aims to obtain. Oracle even claims that the “process flow shows that a data mining project does not stop when a particular solution is deployed. The results of data mining trigger new business questions, which in turn can be used to develop more focused models.” [6] In other words, it is safe to say that data mining has enormous benefits and is becoming a huge position in the field of the information technology discipline, and it will only continue to grow in the coming years.

Bibliography

[1] Brezany, Peter. "Data Mining Basics." *Data Warehousing Fundamentals for It Professionals* (2010): 429-60. *Data Mining and Data Warehousing*. University Wien. Web. 17 Jan. 2017.

[2] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data Mining*. Boston, MA: Addison Wesley, 2012. *Data Mining : Concepts and Techniques*. 2012. Web. 17 Jan. 2017.

[3] IBM. "Big Data." *IBM Big Data – What Is Big Data – United States*. IBM, n.d. Web. 17 Jan. 2017.

[4] Li, Raymond. "KDnuggets." *KDnuggets Analytics Big Data Data Mining and Data Science*. KDnuggets, 2015. Web. 17 Jan. 2017.

[5] Li, Ray. "The History of Data Mining." *Dataconomy - The History of Data Mining*. Dataconomy, 10 June 2016. Web. 17 Jan. 2017.

[6] Oracle. "1 What Is Data Mining?" *What Is Data Mining?* Oracle, 2017. Web. 17 Jan. 2017.

[7] Palace, Bill. "Data Mining : What Is Data Mining." *Data Mining*. UCLA, June 1996. Web. 16 Jan. 2017.

[8] SAS. "What Is Big Data and Why It Matters." *What Is Big Data? | SAS US*. SAS, n.d. Web. 17 Jan. 2017.

[9] Wideskills. "Wideskills." *Data Mining Architecture | Data Mining Tutorial by Wideskills*. Wideskills, 2015. Web. 17 Jan. 2017.