

Information Mining – An Overview

by

Chuck Easttom

Information mining is a general term for the process of discovering data that exists in some underlying data. This is accomplished by applying some algorithm to a body of data in order to extract meaningful information. The term is generally used synonymously with data mining. Alexander (2009) defines data mining as:

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions.

That definition emphasizes business applications, but it should be obvious that the same techniques can be applied to scientific problems. Alexander (2009) describes the problem, which data mining is meant to solve, as an overabundance of data without actionable information. Inexpensive data storage means that most organizations capture and store many different data elements about their organizational operations. However that abundance of data can make it even more difficult to extract meaningful, actionable, information. The need to be able to extract actual knowledge from data provides the impetus for data mining.

Alexander (2009) defined the process of data mining as modeling. He states that data mining makes use of modeling techniques to extract information from the underlying data. For example a business executive might create a model of a potential business scenario, and use a data mining tool to extract information based on that theoretical model. Different algorithms are not the same as different models. A model is a given scenario such as examining what happens to sales numbers if one alters certain advertising variables. Alexander's (2009) examination of data mining considers algorithms separately from the modeling being done. Alexander (2009) focuses on five different approaches to data mining. Within each of these approaches there may

be a number of different algorithms that can be utilized to facilitate the approach. The first technique is the implementation of neural networks. Neural networks can be defined as models that learn through training and resemble biological neural networks in structure. The second data mining technique is the application of decision trees. These are Tree-shaped structures that represent sets of decisions. The third technique that Alexander (2009) discusses is rule induction. This is the extraction of useful if-then rules based on statistics. Those rules are then used to extract meaningful information from the underlying data. One of the more basic, and easily implementable, techniques that Alexander (2009) discusses is called nearest neighbor. With this technique each record is classified based on the records most similar to it in an historical database. The fifth approach mentioned by Alexander (2009) is the most computationally complex. It is the implementation of genetic algorithms. These are algorithms that adapt and change. All of these are possible approaches to extracting meaningful information from a large body of data.

Ciftcioglu and Durmisevic (2001) focus on a very specific aspect of data mining. That aspect is the approach to knowledge generation one implements in data mining. They discuss that fact that one must first establish certain relations among the design information in advance. The point being that one must have some established relationship between data points before one can begin to utilize data mining effectively. They describe this aspect of data mining as knowledge management. They define knowledge in the following way:

Knowledge is the formation of implicit and explicit restrictions placed upon objects, operations and relationships along with general and specific heuristics and inference procedures involved in the situation being modeled.

Ciftcioglu and Durmisevic (2001) emphasize the importance of creating the proper heuristics when one is designing a data mining solution. They imply that the information gathered can only be as good as the design of the heuristics which are used to discover that information. During the design phase of a data mining solution, Ciftcioglu and Durmisevic (2001), state that the main task during design is the implementation of consistent and effective rules. They further state that the difference between data mining and simply selecting a limited number of records from a data source, is a matter of the complexity of the rules used to select the data. In other words, according to Ciftcioglu and Durmisevic (2001), the difference between simply executing a structured query language (SQL) select statement on a database, and implementing a data mining solution, is a matter of the complexity of the selection rules. Data mining solutions don't simply select records or data values that match specific criteria, as occurs in an SQL select statement.

Fayad, Piatetsky-Shapiro, and Smyth (1996).focus on the goal of data mining, rather than the techniques. According to their paper, the goal is simply the discovery of new knowledge that one did not previously have. This is different from simply querying a data source. In a standard database query one has a general idea of what one is going to find, and is simply gathering specific values for data. For example if you query a sales database for the number of widgets sold in Kansas during December of 2009, you know what data you are going to get back, you just don't know the specific values. The process of data mining is about finding new relationships you had not anticipated, thus it is described as knowledge discovery. Fayad, Piatetsky-Shapiro, and Smyth (1996)., cite an important application of data mining to the science of astronomy:

SKICAT, a system used by astronomers to perform image analysis, classification, and cataloging of sky objects from sky-survey images (Fayyad, Djorgovski, and Weir 1996). In its first application, the system was used to process the 3 terabytes (1012 bytes) of image data resulting from the Second Palomar Observatory Sky Survey, where it is estimated that on the order of 109 sky objects are detectable. SKICAT can outperform humans and traditional computational techniques in classifying faint sky objects

This example provided by Fayad, Piatetsky-Shapiro, and Smyth (1996), effectively demonstrates the use of knowledge discovery to discover new knowledge. By implementing a data mining solution, they were not simply able to process more data, but they were actually able to extract more meaningful information, in this case astronomically interesting objects, than traditional methods could do. This aspect of discovering knowledge is the ‘why’ behind data mining. The techniques, approaches, or algorithms are of secondary concern in the study done by Fayad, Piatetsky-Shapiro, and Smyth (1996) study of data mining. These are all implemented in order to find new knowledge.

Web mining is the application of data mining techniques to either individual web pages, or to the entire web according to Cooley, Mobesher, and Shrivistava. (2008). With the growth of the internet and the growing number of web pages, applying data mining techniques to this environment is the next logical step for knowledge discovery. Cooley, Mobesher, and Shrivistava. (2008) provide definitions for two distinct types of web mining. The first type they call Web content mining, is the process of information discovery from sources across the World Wide Web. This is the application of data mining to scour the internet looking at the content of various websites in order to accomplish knowledge discovery. That is the type that they focus on

in their paper. However they also define a second type, which they call web usage mining.

According to Cooley, Mobesher, and Shrivistava. (2008), Web usage mining is the process of mining for user browsing and access patterns. Web usage mining could be more formally defined as the discovery of useful knowledge from the data involving the activities of users on the web, rather than the content of specific web pages.

Guan and Wong (2009). describe one method of extracting knowledge from data, that method is the KPS algorithm. KPS is an acronym for *keywords, patterns and/or samples*. The KPS algorithm mines the desired information from web pages directly using keywords, patterns and/or samples. According to Guan and Wong (2009). this algorithm is based on certain assumptions:

1. important information is always highlighted by keywords or meaningful structures since good visual effects are common practice for representing Web data;
2. common patterns exist in English, e.g. the word after "Dr." or "Mr." should be a name; and
3. similar structures or patterns appear in the Web pages of the same organization for these pages are often written by the same webmasters and thus similar styles (or even simple copies) are employed./

The second two assumptions would seem to be obviously valid given the basic structure of web pages. However the first assumption might have a flaw. Not all websites are well designed. It is entirely possible, and not at all uncommon, for someone to publish a web page that has meaningful content, worthy of data mining, but which is not well structured and would therefore not match the assumptions made by Guan and Wong (2009). It is clear that the KPS

algorithm described by Guan and Wong (2009), is most applicable to the Web mining described by Cooley, Mobesher, and Shrivistava. (2008) as web content mining.

Shrivistava, Cooley, Deshpanda, and Pan-Ning. (2000) describe web usage mining. According to their paper web usage mining consists of three phases, which they define as preprocessing, pattern discovery, and pattern analysis. The preprocessing phase consists of acquiring usage data, removing extraneous data, and then formatting the data into a useable format. The pattern discovery phase is accomplished by algorithms which search the data looking for patterns in the usage data. Finally web usage mining involves the analysis of that data to extract useful information. Shrivistava, Cooley, Deshpanda, and Pan-Ning. (2000) emphasize that all three phases are equally important and must be executed properly in order to achieve the full benefit of web usage mining.

The previously cited examples, including the use of data mining in astronomical research and in web usage mining, are traditional applications of data mining methodologies. Lee and Stolfo.(1998) describe the use of data mining methodologies to enhance network security, specifically of intrusion detection. Intrusion detection is the process of watching for anomalies in network activity that could indicate the activities of an intruder. Many networks implement network intrusion detection systems for this purpose. However one weakness of such systems is that they rely on preset rules. Lee and Stolfo. (1998) advocate systems which can periodically utilize data mining techniques on their raw data in order to extract meaningful knowledge about network activities and perhaps enhance intrusion detection efforts. Lee and Stolfo.(1998) point out that intrusion detection systems, as well as network servers, firewalls, and routers collect an enormous amount of raw data. Server and router logs are two obvious examples. Often this data is so expansive that network administrators do not even attempt to analyze it unless they believe a security incident has occurred and they are seeking more information

about that incident. What Lee and Stolfo. (1998) describe is a system which would periodically examine its own data seeking patterns and thus discovering new knowledge.

Data mining is clearly a useful technology, and there are several methods for implementing data mining principals. It can be applied to scientific data, customer data, web usage data, web content data, and even the raw data acquired by security systems. One can only assume that continued research in this area will yield additional applications of data mining techniques.

References

- Alexander, D. (2009). *Data Mining*.
Retrieved February 18 from
<http://www.laits.utexas.edu/~norman/BUS.FOR/course.mat/Alex/>
- Cooley, R., Mobesher, B., Shrivistava, J. (2008). *Web Mining: Information and Pattern Discovery on the World Wide Web*.
Retrieved February 15 2010 from
<http://www.computer.org/portal/web/csdl/doi/10.1109/TAI.1997.632303>
- Ciftcioglu, O., Durmisevic, S. (2001). *Knowledge management by information mining*.
Retrieved February 17 2010 from <http://repository.tudelft.nl/file/975142/379383>
- Fayad, U., Piatetsky-Shapiro, G., Smyth P. (1996). *From Data Mining to Knowledge Discovery in Databases*.
Retrieved February 17 2010 from <https://www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf>
- Guan, T., Wong, K. (2009). *KPS- a Web Information Mining Algorithm*
Retrieved February 19 2010 from
<http://www8.org/w8-papers/4a-search-mining/kps/kps.html>
- Lee, W., Stolfo, S. (1998). *Data Mining Approaches for Intrusion Detection*.
Retrieved February 16 2010 from
http://www.usenix.org/publications/library/proceedings/sec98/full_papers/full_papers/lee/lee_html/lee.html
- Shrivistava, J., Cooley, R., Deshpanda, M., Pan-Ning, T. (2000). *Web usage mining: discovery and applications of usage patterns from Web data*.
Retrieved February 18 2010 from <http://portal.acm.org/citation.cfm?id=846188>