Advancing Knowledge Discovery and Data Mining

Qi Luo[1,2]

[1] School of Electrical and Information Engineering, Wuhan Institude of Technology, Wuhan, 430073, China
[2]Information Engineering School, Wuhan University of Science and Technology Zhongnan Branch, Wuhan, 430223, Hubei, China
whluo2008@gmail.com

Abstract

Knowledge discovery and data mining have become areas of growing significance because of the recent increasing demand for KDD techniques, including those used in machine learning, databases, statistics, knowledge acquisition, data visualization, and high performance computing. Knowledge discovery and data mining can be extremely beneficial for the field of Artificial Intelligence in many areas, such as industry, commerce, government, education and so on. The relation between Knowledge and Data Mining, and Knowledge Discovery in Database (KDD) process are presented in the paper. Data mining theory, Data mining tasks, Data Mining technology and Data Mining challenges are also proposed. This is an belief abstract for an invited talk at the workshop.

1. Introduction

Knowledge Discovery and Data Mining are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, AI, visualization, and high-performance and parallel computing [1]. People in business, science, medicine, academia, and government collect such data sets, and several commercial packages now offer general-purpose Knowledge Discovery and Data Mining tools [2].

An important Knowledge Discovery and Data Mining goal is to "turn data into knowledge." For example, knowledge acquired through such methods on a medical database could be published in a medical journal. Knowledge acquired from analyzing a financial or marketing database could revise business practice and influence a management school's curriculum [3] [4].

Basing on it, Data mining is the core part of the Knowledge Discovery in Database (KDD) process shown in Fig. 1. The KDD process may consist of the following steps: data selection, data cleaning, data transformation, pattern searching (data mining), and finding presentation, finding interpretation and finding evaluation. Data mining and KDD are often used interchangeably because data mining is the key to the KDD process.

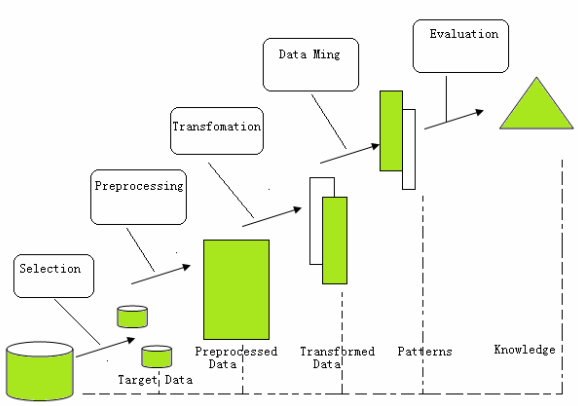

Fig. 1 Knowledge Discovery in Database (KDD) process

2. What is Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or

0-7695-3090-7/08 $25.00 © 2008 IEEE
DOI 10.1109/WKDD.2008.153

patterns among dozens of fields in large relational databases [5][6] .

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:
(1) Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
(2) No operational data, such as industry sales, forecast data, and macro economic data
(3) Meta data - data about the data itself, such as logical database design or data dictionary definitions

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into data warehouses. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

3. Data Mining Tasks

The tasks of data mining are very diverse and distinct because many patterns exist in a large database. Different methods and techniques are needed to find different kinds of patterns. Based on the patterns we are looking for, tasks in data mining can be classified into Summarization, classification, clustering, association and trend analysis[7][8] .
(1) Summarization. Summarization is the abstraction or generalization of data. A set of task-relevant data is summarized and abstracted. This results in a smaller set which gives a general overview of the data, usually with aggregate information.
(2) Classification. Classification derives a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set. In it, every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. This function or model can then classify future objects. This helps us develop a better understanding of the classes of the objects in the database.
(3) Clustering. Clustering identifies classes-also called clusters or groups-for a set of objects whose classes are unknown. The objects are so clustered that the interclass similarities are maximized and the interclass similarities are minimized. This is done based on some criteria defined on the attributes of the objects. Once the clusters are decided, the objects are labeled with their corresponding clusters. The common features for objects in a cluster are summarized to form the class description.
(4) Trend analysis. Time series data are records accumulated over time. For example, a company's sales, a customer's credit card transactions and stock prices are all time series data. Such data can be viewed as objects with an attribute time. The objects are snapshots of entities with values that change over time. Finding the patterns and regularities in the data evolutions along the dimension of time can be fascinating.

4. Data Mining Technology

Data mining adopted its techniques from many research areas including statistics, machine learning, association Rules, neural networks, and so on[9].
(1) Association Rules. Association rule generators are a powerful data mining technique used to search through an entire data set for rules revealing the nature and frequency of relationships or associations between data entities. The resulting associations can be used to filter the information for human analysis and possibly to define a prediction model based on observed behavior.
(2) Artificial Neural Networks are recognized in the automatic learning framework as universal approximations, with massively parallel computing character and good generalization capabilities, but also as black boxes due to the difficulty to obtain insight into the relationship learned.

(3) Statistical Techniques. These include linear regression, discriminate analysis, or statistical summarization.
(4) Machine learning (ML) is the center of the data mining concept, due to its capability to gain physical insight into a problem, and participates directly in data selection and model search steps. To address problems like classification (crisp and fuzzy decision trees), regression (regression trees), time-dependent prediction (temporal trees), and the ML field is basically concerned with the automatic design of if-then rules similar to those used by human experts. Decision tree induction: the best known ML framework was found to be able to handle large-scale problems due to its computational efficiency, to provide interpretable results, and, in particular, able to identify the most representative attributes for a given task.

3. Data Mining Application

Data mining techniques have been applied successfully in many areas from business to science to sports[10].
(1) Business applications. Many organizations now employ data mining as a secret weapon to keep or gain a competitive edge. Data mining has been used in database marketing, retail data analysis, stock selection, credit approval, etc.
(2) Science applications. Data mining techniques have been used in astronomy, molecular biology, medicine, geology and many more.
(3) Other applications. Data mining techniques have also been used in health care management, tax fraud detection, money laundering monitoring and even sports.

4. Conclusion

Successful Knowledge Discovery and Data Mining applications play an important role in data that have clearly grown to surpass raw human processing abilities. The challenges facing advances in this field are formidable. Some of these challenges include as follows:
(1) Develop new mining algorithms for classification, clustering, dependency analysis, and change and deviation detection that scale to large databases.
(2)Develop effective means for data sampling, data reduction.
(3)Develop schemes capable of mining over no homogenous data sets (including mixtures of multimedia, video, and text modalities).
(4) Develop new mining and search algorithms capable of extracting more complex relationships between fields and able to account for structure over the fields (hierarchies, sparse relations).

References

[1] Xindong Wu. “Data Mining: artificial intelligence in data analysis”. *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, 2004.pp.7..

[2] R. Evans, and D. Fisher, “Overcoming Process Delays with Decision Tree Induction,” *IEEE Expert*, Vol. 9, No. 1, 1994, pp. 60–66.

[3] S. Brin et al., “Dynamic Itemset Counting and Implication Rules for Market Basket Data,” *Proceedings of ACM SIGMOD Int’l Conf. Management of Data*, 1997, pp. 255–264.

[4] M. Pazzani, S. Mani, andW.R. Shankle, “Comprehensible Knowledge-Discovery in Databases,” *Proceeding of 19th Annual Conf. Cognitive Science Soc*.1997, pp. 596–601.

[5] R. Brachman, T. Khabaza, W. Kloesgen, G. Piatetsky-Shapiro, and E. Simoudis, Industrial Applications of Data Mining and Knowledge Discovery, *Communzcatzons of ACM*, vol. 39, no. 11.1996.

[6] Communications of The ACM, special issue on Data Mining, vol. 39, no. 11.

[7] Yongjian Fu.Data mining.*Potentials, IEEE*.Volume 16, Issue 4, 1997 pp.18 - 20

[8] Horeis, T.; Sick, B.Collaborative Knowledge Discovery & Data Mining: From Knowledge to Experience. *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*. pp.421-428.

[9] Yi Feng; Zhaohui Wu.Enhancing Reliability throughout Knowledge Discovery Process. *Proceedings of ICDM Workshops on Data Mining*, 2006. pp.754-758.

[10] C. Glymour, D. Ivladigan, D. Pregibon, and P. Smyth. “Statistica.1 Themes and Lessons for Data Mining”, *Data Mining and Knowledge Discovery*, vol. 1, no. 1, 1997.