DESIDOC Journal of Library & Information Technology, Vol. 31, No. 6, November 2011, pp. 446-451

Knowledge Discovery in Databases and Libraries

Anil Kumar Dhiman

Gurukul Kangri University, Haridwar-249 404
E-mail: akvishvakarma@rediffmail.com

ABSTRACT

The advancement in information and communication technology (ICT) has outpaced our abilities to analyse, summarise, and extract knowledge from the data. Today, database technology has provided us with the basic tools for the efficient storage and lookup of large data sets, but the issue of how to help human beings to understand and analyse large bodies of data remains a difficult and unsolved problem. So, intelligent tools for automated data mining and knowledge discovery are needed to deal with enormous data. As library and information centres are considered the backbone of knowledge organisation, knowledge discovery in databases (KDD) is also getting attention of library and information scientists. This paper highlights the basics of KDD process and its importance in digital libraries.

Keywords: Databases, digital libraries, knowledge, knowledge discovery process, data mining

1. INTRODUCTION

Knowledge has been defined in a number of ways and there are a number of perspectives on its nature. There are different types of knowledge, each requiring a different management approach[1]. Traditionally, knowledge was 'justified true belief'[2], but now, 'it is the information with direction, which enables action and decisions[3]. Kafantaris[4], gave the meaning of knowledge in relation to data and information as "… data is a set of discrete objectives or facts of events but would not be instructing organisations what to do. Data when analysed, synthesised, and summarised can become information, which when compared in different situations such as connections, consequences of daily life, social interactions and thoughts and views, can become knowledge."

Earlier the knowledge were written and stored in books, but now in electronic era, it resolves in databases. Various tools, methodologies, and techniques to support knowledge discovery in many fields have been introduced[5]. Data mining came as the process of sampling, exploring, modifying, and assessing large amounts of data to uncover the unknown patterns. It is the extraction or mining of knowledge from large amount of data. Data mining is the science of finding new interesting patterns and relationship in huge amount of data. It is defined as 'the process of discovering meaningful new correlations, patterns, and trends by digging into large amount of data stored in warehouses.'

Knowledge discovery (KD) is a relatively new field where techniques borrowed from algorithms, artificial intelligence, mathematics and statistics are combined to predict and explain new knowledge hidden in volumes of raw data usually stored in databases[6]. Data mining is also taken as synonymous of knowledge discovery in databases (KDD[7)], but data mining is one of the parts of process of knowledge discovery in databases that requires intelligent technologies and the willingness to explore the possibility of hidden knowledge that resides in the data. A generally accepted definition of KDD is given by Fayyad[8], *et al*. as "…the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."

The KDD is the process of automatically searching large volume of data for patterns that can be considered knowledge about the data. This is described as deriving knowledge from the input data, which is the non-trivial extraction of implicit, unknown, and potentially useful information from the data. However, the knowledge discovery process is embedded in and constrained by the organisation's institutional context, which impacts the interpretation of novel information. The important KDD goal is to turn data into knowledge. For example, knowledge acquired through such methods on a medical

Received on 25 February 2011

database could be published in a medical journal. Knowledge acquired from analysing a financial or marketing database could revise business practice and influence a management school's curriculum. Likewise, it has also attracted the attention of library and information science professionals for getting information from large databases for the use of its customers, i.e., the users.

2. PROCESS OF KNOWLEDGE DISCOVERY IN DATABASES

A data warehouse is an ideal means for storing data, which are to be processed by a KDD system. The KDD is a way, to discover interesting and potentially useful patterns or rules in data and the core process of KDD is data mining. However, KDD is not restricted just to data mining, rather it includes data verification and preprocessing, selection of appropriate data mining techniques, verification, and exploitation of their results. A data mining process involves multiple stages, KDD involves the activities leading to actual data analysis and evaluation and deployment of the results. A KDD process thus includes data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted knowledge[9].

The KDD can be likened to the process of searching knowledge units from the databases, where huge volumes of data, particularly the documents, are available without any intended usage. The KDD process consists in processing a huge volume of data in order to extract knowledge units that are non trivial, potentially useful, significant, and reusable. The KDD process is iterative and interactive, and controlled by an expert of the data domain, called the analyst, who is in-charge of guiding the extraction process, on the base of his objectives, and of his domain knowledge. The analyst selects and interprets a subset of the units for building models that are further considered as knowledge units with a certain plausibility.

Generally, the KDD process is aimed at extracting from large databases information units that can be interpreted as knowledge units to be reused and it is based on three major steps: (i) the data sources are prepared to be processed, (ii) then they are mined, and (iii) finally, the extracted information units are interpreted for becoming knowledge units. These units are in turn embedded within a representation formalism to be used within a knowledge-based system. Hitherto, data mining tools mostly adopt techniques from statistics[10], neural networks[11], and visualisation[12] to classify data and extract patterns[13]. But ultimately, KDD aims to enable an information system that transforms the information to knowledge through hypothesis testing and theory formation. It sets new challenges for database technology–new concepts and methods are needed for basic operations, query languages, and query processing strategies[14].

Knowledge discovery is an interactive and iterative process, which according to Soundararajan[15]*, et al*, involves the following steps:

- *Learning the Application Domain:* Includes relevant prior knowledge and goals of the application.
- *Target Data Set*: Selecting a data set or data samples on which discovery is to be performed.
- *Data Cleaning and Preprocessing:* Removing noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes and mapping missing and unknown values.
- *Data Reduction and Projection:* Includes finding useful features to represent the data and using methods to reduce the effective number of variables under consideration.
- *Choosing the Functions of Data Mining:* Selecting the data mining function based on data model such as summarisation, classification, regression, and clustering.
- *Choosing the Data Mining Algorithms:* Includes selecting the methods to be used to search for patterns in the data such as statistical algorithms, visualisation techniques, deviation trend analysis decision tree analysis etc. Here, two or more techniques can be combined depending upon the data models.
- *Data Mining:* Concerned with applying computational techniques to find patterns in data in a particular representational form or set of such representations. A Pattern that is interesting and certain enough can be treated as knowledge.
- *Interpretation:* Includes interpreting the discovered patterns and possibly returning to any of the previous steps as well as possible visualisation of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by the users.
- *Using the Discovered Knowledge:* Includes incorporating discovered knowledge into the performance system, taking action based on the knowledge or simply documenting it for management and for later use.

The KDD process can be illustrated through the following diagram (based and modified after Soundararajan[15], *et al.*) as shown in Fig. 1. So, KDD process is based on three major steps, data preparation, data mining, and interpretation of the extracted units. This

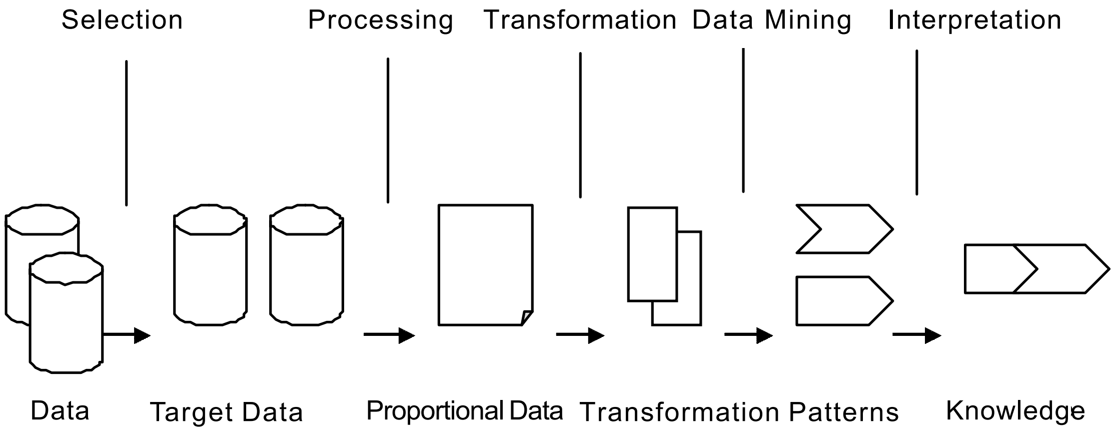

Figure 1. Knowledge discovery process.

process aims on processing a huge volume of data to extract knowledge units that can be reused either by an expert of the domain of data or by a knowledge-based system for problem-solving in the domain of data.

3. KNOWLEDGE DISCOVERY IN DATABASES IN MEETING USERS' DEMANDS

Knowledge discovery refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining involves a collection of tools and techniques for finding useful patterns relating the fields of very large databases. The newest form of data mining is the linguistic summarisation of data which aims at a computer-generated verbal description of the knowledge implicit in a database often in the form of 'if –then' rules that resemble fuzzy knowledge granules. Text mining is to extract patterns from textual documents. A text mining technique typically involves text parsing and analysis to transform each unstructured document into an appropriate set of features and subsequently applies one or more data mining techniques for extracting patterns. Finally, patterns are interpreted for getting knowledge through KDD process.

Several KDD packages offer means to visualise representations[16], but none have shown that users prefer such visualisation tools over textual representations. Furthermore, it is not clear whether 3D visualisations of learned models provide benefits over 2D visualisations[17]. Now various commercial software are also available which offer general purpose KDD tools. But there are some problems in meeting out user demands that may be due to some technical drawbacks in databases.

The first point relevant to KDD are that data repositories which tend to be very large. In practice, knowledge integration will begin to span not only disparate data models in a single archive, but disparate archives in disparate database management systems. The second point is the extremely short phase of data, which are collected cyclically. Data discovery must accommodate collection cycles that may be unknown – as for example identifying the cycles of shifts in major geological faults, or that may shift from cycle to cycle in both time and space. Instead, knowledge discovery researchers can turn attention to problems of reasoning and modeling on very short temporal cycles. Infrastructure issues that need to be included, for example – the development of real-time data mining, and to utilise knowledge discovery tools to guide correlation of discovered data patterns across time, determination of temporal drift, validation of data trends across temporal discontinuities, and so forth. And a third point of relevance to KDD applies to a characteristic of the data foundation rather than of the data. The emergence of the internet has supported development of data clearing houses, digital libraries, and online repositories wherein one does not access data, but points to data.

It is paradoxical that as increasing amounts of digital data become available via the internet, it becomes increasingly difficult to locate, retrieve, and analyse. This is due to the fact that the internet lacks a comprehensive catalogue or index[18]. Without a coordinating infrastructure, many data sources and services available today remain essentially inaccessible. Currently, over three million websites are online[19], and yet even the best search engines can locate only one third of the accessible pages. Hence data mining tools need to be established to locate environmental data sources in the 'gray literature' areas of the internet.

Such data sources include but are not limited to field data collected in developing countries, much localised community data sites such as inner city neighborhood and community activist sites, and similar data sources not known to or known by conventional doorways into the geospatial data infrastructure. This type of knowledge discovery treats the entire internet as a very large, decentralised data repository, and provides a venue for contributions to a global information infrastructure.

4. KNOWLEDGE DISCOVERY IN DIGITAL LIBRARIES

There have been seen various additions to the services in traditional libraries in the recent past, especially in the first decade of 21st century. The addition to the traditional library services meant that librarians and users had to acquire additional knowledge and skills to use computers and find the relevant information. This is very important, because since then, the pace of change of information technology and its implementation in libraries has become faster and the request for new and advanced competencies of librarians and users have put in focus many times. These basic knowledge requirements put certain amount of unwanted pressure on users and of course, on the librarians. Knowledge management is a systematic process of acquiring, organising, sustaining, applying, sharing, and renewing both tacit and explicit knowledge to enhance the organisational performance, increase organisational adaptability, increase values of existing products and services and/or create new knowledge intensive products, processes and services[20-23]. As the digital libraries become more knowledge-conscious, knowledge discovery and data mining have become essential for knowledge management. The information management system should have data mining capabilities for helping the knowledge discovery process, and subsequently, to evolve as a knowledge management system from the existing information. Knowledge is then organised by indexing knowledge elements, filtering based on content, and establishing linkages and relationships among the elements. Subsequently, this knowledge is made available to users for supporting their decision-making process. Keeling and Hornby[24], notes that the information age is linked with life-long learning and technological skills of librarians needs to be updated. Librarians live in a competitive environment, and they are advised to stay focused and relevant while applying knowledge management principles together with information technology and communication tools in the libraries to facilitate the rapidly changing environment.

Digital libraries have a much greater capacity for knowledge management. Their current architecture should include classification and thesaurus—the vocabulary control and knowledge-organising tools, which serves three purposes in a traditional library-the description, organisation, and retrieval of information. For more effective and efficient exploration, the networked information should be pre-arranged together with vigorous improvement of search techniques. Data mining involves techniques for machine learning, pattern recognition, statistics, linguistics, and visualisation with the rapid expansion of full-text and bibliographic databases on the web. Internet navigation has become a serious concern to libraries. Data mining is more suitable to the libraries that purchase access to full-text databases rather than physical materials and bibliographic databases. Full-text databases are considered one of the most important resources in the digital libraries, especially the hybrid libraries which contain print and offline learning materials together. Metadata – the data about the data itself can play important role in data mining in these databases for the users. Whereas, in a digital library, electronic documents should be accessible on 24x7 basis; scholars should be able to find needed documents using search techniques; and there should be the possibility of free exchange of scholarly ideas and different levels of editing for authors should be provided, which are to be expected to be fulfilled by data mining and a step further, by knowledge discovery in libraries. Mudhol & Gowda[25], have listed the following benefits of data mining, and of course of the knowledge discovery in the digital libraries:

- Users can apply the techniques to measure the use patterns and reuse patterns of databases and software
- In the area of bibliometrics, to discover patterns in unidentified knowledge areas
- Data mining (knowledge discovery) techniques can be applied to text analysis tasks such as discipline extractions
- Data mining (knowledge discovery) techniques can be incorporated for information retrieval process for better browsing and searching.

Data mining is considered to be an important step in the process of knowledge discovery that emphasises the cleaning, warehousing, and mining of knowledge in databases. It is a form of artificial intelligence that uses automated processes to find information. Although its use in libraries is limited, data mining has been used successfully for several years in the scientific, medical, and business communities for tracking behaviour of individuals and groups, processing medical information and for a number of other applications.

But as stated earlier, the KDD is a step ahead of data mining process. Classification and thesauri, that contain condensed intelligence can be used in organising networked information, especially metadata, to facilitate the information resources' usability and catalyse the digital library into knowledge management centre in KDD process. Classification and thesaurus can be merged into a concept network and the metadata can be distributed into the nodes of the network according to their subjects. The abstract concept node substantiated with the related metadata records becomes a knowledge node. This forms a consistent knowledge network that is not only a framework for resource organisation, but also a structure for knowledge navigation, retrieval, and learning. The bibliographic data are one of the most important resources of library, which may be useful for knowledge discovery process. Based on the subject indexing, the bibliographic data can be combined with the classification and thesaurus to form a knowledge structure, which provides a skeleton for organisation of bibliographic data. Corpus knowledge can be formed when new terms can be extracted automatically from the bibliographic data to update the classification and thesaurus. Such a knowledge network provides the user with an opportunity for navigation, searching, and learning in digital libraries.

5. CONCLUSIONS

The KDD technology is emerging as an empowering tool in the development of the next generation database and information systems through its abilities to extract new, insightful information embedded within large heterogeneous databases and to formulate knowledge. Knowledge discovery in databases is rapidly growing and its development is driven by strong research interests as well as urgent practical, social, and economical needs.

Digital libraries are also expected to extend their current services of modern automated libraries and present creative environment for different multidisciplinary projects[26]. There is no doubt that today's KDD tools provide value to organisations that collect and analyse their data and it is expected that more knowledge discovery tools than simply creating accurate models as in machine learning, statistics, and pattern recognition will be evolved in future. Thus, digital librarians will be able to fully utilise the benefits of data mining and knowledge discovery in mining information from the huge world of information, particularly over the net to their users is library and information centres.

REFERENCES

1. Alavi, M. & Leidner, D.E. Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly,* 2001, **25**(1), 107-36.

2. Nonaka, I. A dynamic theory of knowledge creation. *Organization Science,* 1994, **5**(1), 14-37.

3. Beccerra-Fernandez, I.; Gonzalez, A. & Sabherwal, R. Knowledge management: Challenges, solutions, and technologies. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

4. Kafantaris, Yasmin. The importance of tacit knowledge. *Managing Information,* 2002, **9**(9), 54-55.

5. Fayyad, U.M. & Simoudis, E. Tutorial on knowledge discovery and data mining. *In* IJCAI-95. Montreal, Quebec, Canada, 21 August 1995. www-aig.jpl.nasa.gov/public/kdd95/tutorials/IJCAI95-tutorial.html (accessed on 20.1.011).

6. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. & Uthurusamy, R. Preface. Advances in knowledge discovery and data mining. AAAI Press, California, 1996. 560 p.

7. Raza, Khalid. Application of data mining in bioinformatics**.** *Ind. J. Comp. Sci. Engg.,* 2002, **1**(2), 14-8.

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

9. Fayyad, U. Editorial. Data Mining Knowl. *Discovery,* 1997, **1**, 5-10.

10. Glymour, C.; Madigan, D.; Pregibon, D. & Smyth, P. Statistical inference and data mining. *Communications of ACM*, 1996, **39**(11), 35-41.

11. Lu, H.; Setiono, R. & Liu, H. Effective data mining using neural networks. *IEEE Trans. Knowl. Data Engg.,* 1996, **8**(6), 957-61.

12. Lee, H. & Ong, H. Visualization support for data mining. *IEEE Expert,* 1996, **11**(5), 69-75.

13. Dhiman, A.K. Data mining and its use in libraries. *In* CALIBER-2003: Mapping technology on libraries and people, edited by T.A.V. Murthy. INFLIBNET, Ahmedabad, 2003. pp. 568-74.

14. Lmielinski, T. & Mannila, H. A database perspective on knowledge discovery. *Communications of ACM,* 1996, **39**(11), 58-64.

15. Soundararajan, E.; Joseph, J.V.M.; Jayakumar, C. & Somasekharan, M. Knowledge discovery tools and techniques, 2005. library.igcar.gov.in/readit-2005/conpro/km/s3-1.pdf (accessed on 20.1.2011).

16. Brunk, C.; Kelly, J. & Kohavi, R. Mine-set: An integrated system for data mining. *In* 3rd International Conference on Knowledge Discovery and Data Mining. AAAI Press, California, 1997.

17. Sebrechts, M.; *et al.* Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. *In* 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, 1999.

18. Buttenfield, B.P. Looking forward: Geographic information services and libraries in the future. *Cartography GIS,* 1998, **25**(3), 161-71.

19. National Research Council. Data foundation for the national spatial data infrastructure. National Research Council Mapping Science Committee. National Academy Press, Washington (DC), 1999.

20. Dhiman, A.K. Knowledge management system for knowledge management in IT era. *Ind. J. Info. Lib. Soc.,* 2003, **16**(1-2), 1-8.

21. Dhiman, A.K. Knowledge management: Its implications in library & information centers. *In* 4th National Conference of CGLA: Library Vision 2020: Profession & Education, edited by V.P. Singh, *et al.* CGLA, Dehradun, 2009. pp. 3-11.

22. Dhiman, A.K. & Rani, Yashoda. Library management: A manual book for effective management. Ess Ess Publications, New Delhi, 2004.

23. Dhiman, A.K. & Sharma, H. Knowledge management for librarians. Ess Ess Publications, New Delhi, 2009.

24. Keeling, Carole & Hornby, Susan. Knowledge management in the networked public libraries. *Managing Information,* 1999, **6**(8), 27-29.

25. Mudhol, Mahesh & Gowda, Purushothama. Data mining in the process of knowledge discovery in digital libraries. *In* PLANNER–2003. INFLIBNET, Ahemdabad, 2004. pp. 164-67.

26. Dhiman, A.K. & Rani, Yashoda. Digital libraries. Ess Ess Publications, New Delhi, 2011.

About the Author

Dr Anil Kumar Dhiman holds MA, MSc, MLIS, BEd, PGDCA and PhD (Botany) and PhD (Library & Information Science). He has 118 papers and 28 books to his credit in Library & Information Science and Botany. He has presented 31 papers in seminars and conferences of national and international reputes including one in Nepal during February–March 2011. He has more than two-decades working experience in the library profession. Presently, he is Information Scientist at Gurukul Kangri University, Haridwar. He is guide of few students for their PhD in Library Science. He is awarded with the *APSI Young Scientist Award* and *Gold Medal* in 1999, *USSHLE-IJILIS-PBDBBS Award* in Library & Information Science in 2003, and *Glory of India International Award* and *Gold Medal* in 2006.