KDD for Business Intelligence

Rafi Ahmad Khan, University of Kashmir, Srinagar, India

ABSTRACT:

Emergence of knowledge based economy has posed serious challenges to companies. Intelligent organizations recognize that knowledge is an intellectual asset that grow over time and when harnessed effectively, can sustain competition and innovation. Organizations can use IT for leveraging the entire organization's intellectual resources for great financial impact. Business Intelligence (BI) along with KDD (Knowledge Discovery in Databases) plays a pivotal role in leveraging the intellectual assets of companies by creating, storing and sharing that knowledge for effective decision making. Companies are now realizing the potential payoffs of KDD applications along with BI. Consequently, BI is spreading its wings to cover small, medium and large companies. This paper explores the concepts of BI, KDD, process of knowledge discovery, key levers of knowledge strategy and benefits of BI.

Keywords: *Business intelligence, Knowledge discovery in databases, KDD, Data mining, OLAP*

Introduction

Over the past few decades, the industrialized economy has been going through a transformation from being based on natural resources to being based on intellectual assets (Alavi, 2000; Tseng & James Goo, 2005). The knowledge based economy is a reality (Godin, 2006). Rapid changes in the business environment cannot be handled in traditional ways as companies are expanding and are much larger today than they used to be, fueling the need for better tools for collaboration, communication and knowledge sharing. Competing in the globalized economy and markets requires quick and effective response to customer needs and problems. For companies spread over wide geographical areas and virtual organizations, managing knowledge is critical for providing services. Companies must develop the strategies to sustain competitive advantage by leveraging their intellectual assets for optimal performance (Skyrme, may, 2002; Agrawal et al, 1996).

Knowledge And Business Intelligence (BI)

Presently, information technology assumes a major role in business, because of it pivotal role in building business intelligence in enterprises. Business Intelligence (BI) is a broad category of applications and technologies for gathering, analyzing, and providing access to huge data stored within the company's database. The term intelligence in Business Intelligence (BI) is closely related with knowledge. Knowledge refers to stored information or models used by a person or machine to interpret, predict and appropriately respond to the outside world (Fischler & Firschein, 1987). In the IT context knowledge is very distinct from data and information. Whereas data are facts, measurements and statistics, information is organized or processed data that is timely and accurate (Hoffer et al, 2002; Kankanhalli & Tan, 2005). Knowledge is information that is contextual, relevant

and actionable. Having knowledge implies that it can be exercised to solve a problem. While data, information and knowledge may all be viewed as assets of an organization, knowledge provides higher level of meaning about data and information.

Intelligence is often defined as the general mental ability to learn (acquire knowledge) and how to apply knowledge. Intelligence encompasses cognition. Cognition is the method by which people assimilate and integrate knowledge, while intelligence is both the assimilation of knowledge as well as the ability to apply such knowledge. Thus knowledge is imperative for BI system in an organization. A BI system must have capability to manage knowledge, store it in a knowledge repository and tools that can apply that knowledge for better decision making. Companies are adopting Business Intelligence (BI) systems and tools because of their capability to learn from the past and forecast the future.

Knowledge in an enterprise may originate from many different sources. They include information systems, reports, Internet, corporate databases, customers, suppliers or government agencies. Knowledge of the employees is an absolute source of information. It results from their experience and intuition.

Polanyi (1958) first conceptualized the difference between an organizations explicit and implicit knowledge. Explicit knowledge deals with more objective, rational, and technical knowledge (e.g., data, policies, procedures, software and documents). It is leaky knowledge as it can be readily documented (Alavi, 2000). Tacit knowledge is usually in the domain of subjective, cognitive and experiential learning; it is highly personal and difficult to formalize. It is also referred to as embedded knowledge (Tuggle & Goldfinger, 2004) as it typically involves expertise, know-how, trade secrets, skill set, understanding and learning, hence difficult to document. When people leave the organization, they take their knowledge with them. Consequently it has become vital for organizations to retain the valuable know-how that can so easily and quickly leave an organization. Organizations now recognize the need to capture and integrate both types of knowledge. BI is the process that transforms data into information and then into knowledge (Golfarelli et al, 2004). It has proven to be successful not only in analyzing data, but also in discovering knowledge by uncovering trends and patterns that are hidden deep within datasets. These hidden trends and patterns can be investigated to forecast future directions (Watson & Wixom, 2007). BI is spreading its wings to cover small, medium and large companies, more and more analytical tools are penetrating the market to do any kind of analysis and help to make informed decision making (Khan & Quadri, 2012).

In business management term BI is used to be described as applications and technologies which are used to gather, provide access to and analyze data and information about an enterprise, in order to help them make better informed business decisions (Reinschmidt & Francoise, 2002; Moss & Atre, 2003; Wu et al, 2007; Jonathan, 2000). Ranjan (2008) argues that BI is the conscious, methodical transformation of data from any and all data sources into new forms to provide information that is business-driven and results-oriented. As per (Pirttimäki, 2004), BI process concept is understood as a continuous and systematic method of action by which an organization gathers, analyses, and disseminates relevant business information to business activities. Cui et al (2007) argues that BI is the way and method of improving business performance by providing powerful assistance to executive decision maker which enables them to have actionable information at hand. BI tools are viewed as technology that enhances the efficiency of business operation by providing an increased value to the enterprise information and hence the way this information is utilized. According to Zeng et al (2006)

BI is "The process of collection, treatment and diffusion of information that has an objective, the reduction of uncertainty in the making of all strategic decisions."

BI can be utilized to view not only current action, but also suggest the most suitable direction an organization should take, consequently BI can be an invaluable tool for decision–makers and managers (Dhar & Stein, 1997). However, the success of BI tools depends on the quality of data it uses. Therefore quality data, its transformation to information and extraction of knowledge from it, are essential to a successful BI implementation. As a result, it is vital to explore the techniques that can be implemented to select and analyze organizational data. Knowledge Discovery in Databases (KDD) is one process, which can be explored to ensure the highest quality of data is available for BI applications.

Knowledge Discovery In Databases (KDD)

Database and their tools provide the necessary infrastructure to store, access, and manipulate data. Data warehousing, a recently popularized term, refers to the current business trend of collecting and cleaning transactional data to make them available for online analysis and decision support. A popular approach for analysis of data warehouses is called online analytical processing (OLAP) (Agrawal et al, 1996). OLAP tools focus on providing multidimensional data analysis, which is superior to SQL (a standard data manipulation language) in computing summaries and breakdowns along many dimensions. While current OLAP tools are semi-automated and target interactive data analysis, they will also include more automated discovery components in the near future.

KDD is the automated process that relies on methods from various fields like pattern recognition, applied statistics, machine learning, neural networks etc. to find patterns from data in the data mining step of the KDD process. The phrase 'Knowledge Discovery in Databases' was coined at the first KDD workshop in 1989 by Piatetsky-Shapiro, who emphasized that knowledge is the end product of a data-driven discovery (Piatetsky-Shapiro & Frawley, 1991). KDD can be defined as the non-trivial extraction of implicit, previously unknown, and potentially useful information from databases. The declining cost of storage technology and advances in data communication technologies have enabled the companies to capture and store data with ease (Fayyad, 1996: Witten & Frank, 2005). This growth, by far exceeds human capacities to analyze the databases in order to find hidden rules or patterns within the data. Therefore, knowledge discovery becomes more and more important in databases (Lazcorreta et al, 2008).

Knowledge discovery is multidisciplinary area of research (Wu, 2004) and is apparent in almost any field; science; marketing; finance; health care; retail etc. The traditional method of turning data into knowledge relies upon manual analysis and interpretation (Witten & Frank, 2005), but as the volume of data captured increases, manual data analysis has become unrealistic in many domains. Therefore, the requirement to scale up human analysis capabilities in order to analyze the large volumes of data became imperative.

KDD is concerned with the development of methods and techniques that helps in the analysis and interpretation of huge volumes of data. At the core of the KDD process is the application of specific data mining methods for pattern discovery and extraction (Han & Kamber, 2006; Geist, 2002).

Process Of Knowledge Discovery

The KDD process model is an interactive, iterative, procedure that attempts to extract implicit, previously unknown and potentially useful knowledge from data through a scientific method. KDD process model can be divided into a number of steps.

There are a number of variants of KDD process, such as those that have been published by Adriaans & Zantinge (1996), Brachman & Anand (1996), and Han & Kamber (2006) in addition to others, however, all variants of the KDD process remain close to as described by (Fayyad, 1996) and supported by (Roiger & Geatz, 2003). It is a well-accepted approach of the KDD process, consisting of several stages as depicted in Figure 1. The various stages are:

Data Selection: The goal of this stage is the extraction of relevant data from huge data stored in operational database, data warehouse and data marts that is relevant to the data mining analysis.

Data Preprocessing: This stage of KDD is concerned with data cleansing and preparation tasks that are essential to ensure correct results. Eliminating missing values in the data, ensuring that coded values have a uniform meaning and ensuring that no spurious data values exist are typical activities that occur during this stage.

Data Transformation: This stage is aimed at transforming the data into a two-dimensional table and eliminating unwanted fields so the results are valid.

Data Mining: The goal of the data mining stage is to analyze the data by a suitable set of algorithms in order to discover meaningful patterns and rules and produce predictive models. This is the main element of the KDD cycle.

Interpretation and Evaluation: Using discovered knowledge includes incorporating this knowledge into the performance system, taking actions based on the knowledge, or simply documenting it and reporting it to interested parties.
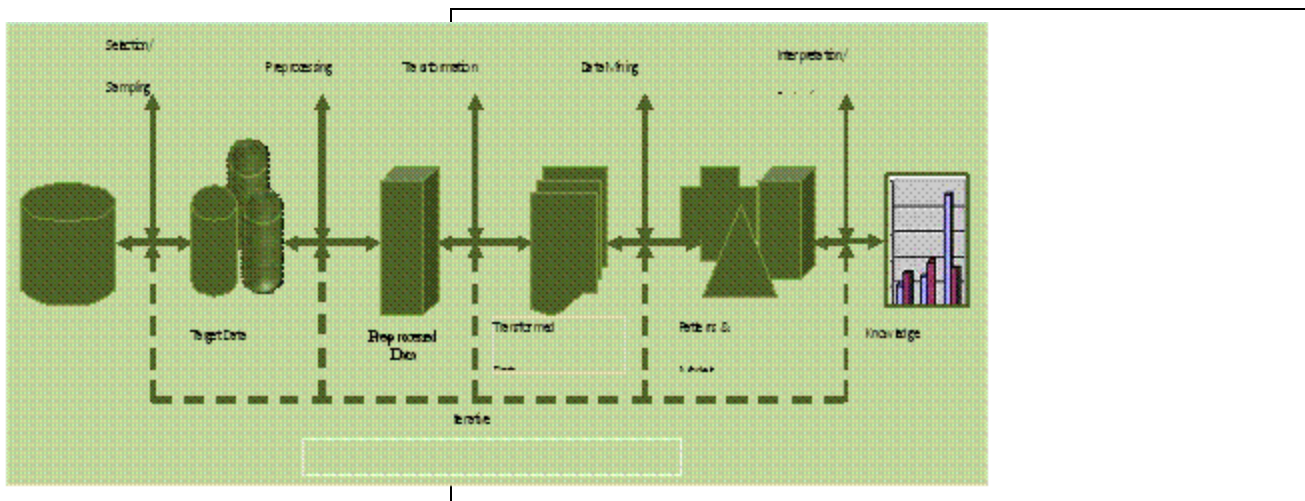
Figure 1: Knowledge Discovery Process

While data mining algorithms have the potential to yield an unlimited number of patterns hidden in the data, several of these may not be meaningful or useful. This final phase is aimed at selecting those models that are valid and worthwhile for making future business decisions. It is possible to return to any of the previous stages at this point, should the need arise. Furthermore, the evaluation can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models. The result of this process is newly acquired knowledge formerly hidden in the data. There are generally five types of information that can be obtained through data mining which include (Turban & Aronson, 2001):

- *Classification:* It infers the defining characteristics of certain group (e.g., customers who have been lost to competitors).

- *Clustering:* It identifies groups of items that share a specific characteristics (clustering differs from classification in that no predefining characteristics is given).

- *Association:* It identifies relationships between events that occur at one time (e.g., the contents of a shopping basket).

- *Sequencing:* It is similar to association, except that the relationship exists over a period of time (e.g., repeat to a supermarket or use of a financial planning product).

- *Forecasting:* It estimates future values based on patterns within large sets of data(e.g., demand forecasting)

Data Mining constitutes one step in the KDD process. It is in data mining step that the actual search for patterns of interest is performed. It is important at this stage to choose the appropriate data mining algorithm (like neural networks, linear/logistic regression, association rules, etc.) for the data mining task. The data mining task itself can be a classification task, linear regression analysis, rule formation, or cluster analysis (Imberman & Susan, Dec 2001).

The extracted information/knowledge obtained by applying BI tools must be stored in knowledge repository for future use and sharing within the organization.

Key Levers Of Knowledge Strategy

Companies are increasingly recognizing the contribution of knowledge in the form of intellectual capital or knowledge base of the firm and the value that can be derived from it. Skyrme (2002) has given seven key levers that will have the maximum impact on a knowledge-based strategy. Briefly the seven levers are:

- *Customer knowledge*: repeatedly cited in surveys as the most important knowledge an organization needs to capture and exploit; Business Intelligence (BI) helps in retaining existing and identifying potential customers and suppliers.

- *Knowledge-enhanced products and services*: adding value by surrounding the product with additional information, such as personal preferences when booking travel; BI helps to add value by using the captured knowledge and collaboration.

- *Knowledge in people:* people-focused programs aim to continually improve workforce skills through development: BI helps in skills development through collaboration technologies and eLearning.

- *Organizational memory*: knowing what an organization knows, over space and time e.g. sharing best practice or recording lessons learned: BI through collaboration technology helps in sharing practices.

- *Knowledge in processes:* capturing the knowledge of the best professionals and embedding their good practices into the recommended procedures: BI helps to capture implicit knowledge from experts.

- *Knowledge in relationships:* creating forums and other mechanisms to have intimate sharing of knowledge with suppliers, customers and partners; sharing of knowledge is a common practice BI

- *Knowledge assets* – the intellectual capital focus; BI not only focuses on the intellectual capital of company but also assists to perceiving corporate competencies of competitors.

Benefits

The major benefit of BI with knowledge discovery tools is the ability to provide accurate information when needed, including real-time view of corporate performance. Thomson (2004) on the basis of survey reported the following major benefits of BI:

- Faster, more accurate reporting (81 percent)

- Improved decision making(78 percent)

- Improved customer services (56 percent)

- Increased revenue (49 percent)

- Many of the benefits of BI are intangible. That's why, according to Eckerson (2003), so many executives do not insist on rigorous cost justification of BI Projects.

Conclusion

The continuous shift towards a knowledge-based economy has brought to the fore the issue of how knowledge is created, assimilated and used to obtain economic returns. Knowledge embodied in intellectual assets (*e.g.* human capital, R&D, patents, software, documents etc.) is becoming essential for organizations economic performance and growth. In this new environment, companies need to be able to earn economic returns from both developing and using intellectual assets. KDD is a new generation of computational techniques and tools that support the extraction of useful knowledge from the rapidly growing volumes of data. The organizations are now realizing the potential payoffs of KDD applications along with BI. By bringing together a set of diverse fields, KDD along with BI creates fertile ground for the development of new tools and techniques for managing, analyzing, and creating value from the flood of data facing modern business world. It is the responsibility of researchers, academicians and practitioners in this field to ensure that users understand the potential contributions of KDD and BI for creating value and gaining competitive advantage.

References

Adriaans, P., & Zantinge, D. (1996). *Data Mining.* Harrow, England: Addison-Wesley.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, I. (1996). Fast discovery of association rules. In U. Fayyad, Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, *In Advances in Knowledge Discovery and Data Mining.* London: AAAI/MIT Press.

Alavi, M. (2000). Managing Organizational Knowledge. In R. ZMUD, *In Framing the Domain of IT Management.* Pinaflex Educational Resources.

Brachman, R., & Anand, T. (1996). The process of knowledge discovery in databases: a human-centered approach. In *Advances in Knowledge Discovery and Data Mining* (pp. 37–58). AAAI Press.

Cui, Z., Damiani, E., & Leida, M. (2007). Benefits of Ontologies in Real Time Data Access. *Digital Ecosystems and Technologies Conference, DEST '07.pp. 392-397.*

Dhar, V., & Stein, R. (1997). *Intelligent decision support methods.* Upper Saddle River, NJ: Prentice Hall.

Eckerson, W. (2003). *smart Companies in 21st Century: The Secrets of Creating sucessful BI solutions.* Seatle: The Data Warehousing Institute.

Fayyad, U. M. (1996). Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE Expert*(11(5)), 20–25.

Fischler, M. A., & Firschein, O. (1987). *Intelligence: The Eye, The Brain anf The Computer.* Addison-Wesley.

Geist, I. (2002). A framework for data mining and KDD. *2002 ACM symposium on applied computing.* Madrid, Spain.

Godin, B. (2006). The Knowledge-Based Economy:Conceptual Framework or Buzzword? *The Journal of Technology Transfer*, 17-30.

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: what's next in business intelligence? *DOLAP '04: Proceedings of the 7th ACM international workshop on Data warehousing and OLAP.* New York.

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques, Second Edition.* Morgan Kaufmann.

Hoffer, J., Prescort, M., & McFadden, F. (2002). *Modern Database Management 6th ed.* Prentice Hall.

Imberman, & Susan, P. (Dec 2001). Effective Use Of The KDD Process And Data Mining For Computer Performance Professionals. *Proceedings of CMG 2001.*

Jonathan, W. (2000). Business Intelligence: What is Business Intelligence? *DM Review*.

Kankanhalli, A., & Tan, B. C. (2005). Knowledge Management Metrics: A Review and Directions for Future Research. *International Journal of Knowledge Management vol, 1*(2), 20-32.

Khan, R. A., & Quadri, S. M. (2012). Business Intelligence: An Integrated Approach. *The Business Intelligence Journal (BIJ), 5*(1), 64-70.

Lazcorreta, E., Botella, F., & Fernández-Caballero, A. (2008). Towards Personalised Recommendation by Two-Step Modified Apriori Data Mining Algorithm. *Expert Systems with Applications, 35*(3), 1422-1429.

Moss, L., & Atre, S. (2003). *Business Intelligence Roadmap: The Complete Lifecycle for Decision-Support Applications.* Boston: Addison-Wesley.

Piatetsky-Shapiro, G., & Frawley, W. (1991). *Knowledge Discovery in Databases.* Menlo Park, Calif: AAAI Press.

Pirttimäki, V. (2004). The Roles of Internal and External Information in Business Intelligence. *Frontiers of E-Business Research*.

Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy.* Chicago: University of Chicago Press.

Ranjan, J. (2008). Business justification with business intelligence. *The Journal of Information and Knowledge Management Systems, 38*(4), 461-475.

Reinschmidt, J., & Francoise, A. (2002). *Business Intelligence Certification Guide.* IBM, International Technical Support Organization.

Roiger, R. J., & Geatz, M. W. (2003). *Data Mining a Tutorial Based Primer.* San Francisco: Addison-Wesley.

Skyrme, D. J. (2002). Business value from knowledge management. *Conference Mobilising Knowledge for Business Performance.* London.

Thomson, O. (2004, Oct). *siness Intelligence Success, Lessons learned.* Retrieved from www.technologyevaluation.com

Tseng, C. Y., & James Goo, Y. J. (2005). Intellectual capital and corporate value in an emerging economy: empirical study of Taiwanese manufacturers. *R&D Management*, 187–201.

Tuggle, F. D., & Goldfinger, W. E. (2004). A Methodology for Mining Embedded Knowledge from Process Maps. *Human Systems Management, 23*(1).

Turban, E., & Aronson, J. E. (2001). *Decision Support Systems & Intelligent Systems, 2nd ed.* India: Pearson Education Inc .

Watson , H. J., & Wixom, B. (2007). The Current State of Business Intelligence. *IEEE Computer*, *40*, pp. 96-99.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann.

Wu, L., Barash, G., & Bartolini, C. (2007). A Service-oriented Architecture for Business Intelligence. *Service-Oriented Computing and Applications SOCA '07, IEEE International Conference.* 279-285.

Wu, X. (2004). Data Mining: Artificial Intelligence in Data Analysis. *Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology.*

Zeng, L., Xu, L., Shi, Z., Wang, M., & Wu, W. (October 8-11, 2006). Techniques, process, and enterprise solutions of business intelligence. *IEEE Conference on Systems, Man, and cybernetics*, *6*, p. 4722. Taipei, Taiwan.

Contact the Author:

Rafi Ahmad Khan, The Business School, University of Kashmir, Srinagar, J&K, India; Email: mca_rafi@yahoo.com