ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

Knowledge Discovery Process: The Next Step for Knowledge Search

Ravindra Changala, D.Rajeswara Rao, T Janardhana Rao, P Kiran Kumar, Kareemunnisa

Assistant Professor, Dept of IT, Guru Nanak Institutions Technical Campus, Hyderabad, India

Professor, CSE Dept, K L University, Guntur, Andhra Pradesh, India

Associate Professor, Dept of IT, Guru Nanak Institutions Technical Campus, Hyderabad, India

Assistant Professor, Dept of CSE, Vignan Institute of Management and Technology for Women, Telengana, India

Assistant Professor, Dept of IT, Guru Nanak Institutions Technical Campus, Hyderabad, India

ABSTRACT: The process of extracting knowledge from the large volumes of data is data mining is a crucial step in KDD (Knowledge Discover from Data). Many algorithms are available to analysis data in mining. But process of getting knowledge was not explained in detail. We attempt to have a process to find new knowledge. KDP (Knowledge Discovery Process) is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non database sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyse massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analysing the application domain. Although the models usually emphasize independence from specific applications and tools, they can be broadly divided into those that take into account industrial issues and those that do not. However, the academic models, which usually are not concerned with industrial issues, can be made applicable relatively easily in the industrial setting and vice versa. We restrict our discussion to those models that have been popularized in the literature and have been used in real knowledge discovery projects.

KEYWORDS: Data mining, KDD, KDP, Process models, research issues.

I. INTRODUCTION

Knowledge Discovery and Data Mining is a very dynamic research and development area that is reaching maturity. As such it requires stable and well-defined foundations which are well understood and popularized throughout the community. This survey presents a historical overview description and future directions concerning a standard for a Knowledge Discovery and Data Mining process model It presents a motivation for use and a comprehensive comparison of several leading process models and discusses their applications to both academic and industrial problems. The main goal of this review is the consolidation of the research in this area. The survey also proposes to enhance existing models by embedding other current standards to enable automation and interoperability of the entire process. Knowledge Discovery is the most desirable end-product of computing. Finding new phenomena or enhancing our knowledge about them has a greater long-range value than optimizing production processes or inventories and is second only to task that preserve our world and our environment. It is not surprising that it is also one of the most difficult computing challenges to do well. Current technological progress permits the storage and access of large amounts of data at virtually no cost. Although many times preached the main problem in a current information-centric world remains to properly put the collected raw data to use. The true value is not in storing the data but rather in our

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

ability to extract useful reports and to find interesting trends and correlations through the use of statistical analysis and inference to support decisions and policies made by scientists and businesses.

II. KNOWLEDGE DISCOVERY PROCESS

Knowledge discovery is the process of nontrivial extraction of information from large databases, information that is implicitly present in the data, previously unknown and potentially useful for users. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

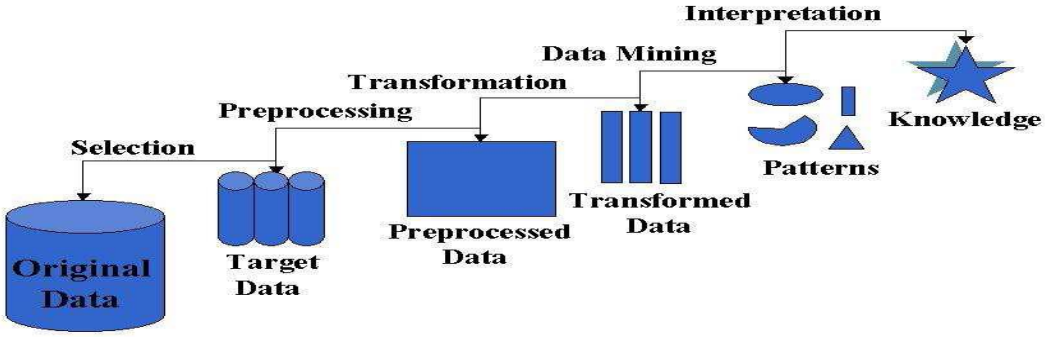

Fig 1. Steps in KDD

It consists of an iterative sequence of the following steps:

Data cleaning (to remove noise and inconsistent data)
Data integration (where multiple data sources may be combined)
Data selection (where data relevant to the analysis task are retrieved from the database)
Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures).
Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

III. KNOWLEDGE DISCOVERY PROCESS MODELS

The **knowledge discovery process** (KDP), also called knowledge discovery in databases, seeks new knowledge in some application domain. It is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. The process generalizes to non database sources of data, although it emphasizes databases as a primary source of data. It consists of many steps (one of them is DM), each attempting to complete a particular discovery task and each accomplished by the application of a discovery method. Knowledge discovery concerns the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analysing the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

application domain.

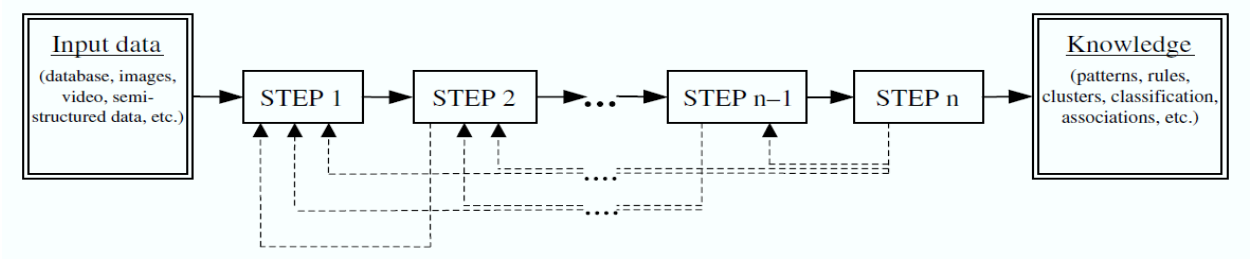

Fig 2. Sequential Structure of KDP

To formalize the knowledge discovery processes (KDPs) within a common framework, we introduce the concept of a process model. The model helps organizations to better understand the KDP and provides a roadmap to follow while planning and executing the project. This in turn results in cost and time savings, better understanding, and acceptance of the results of such projects. We need to understand that such processes are nontrivial and involve multiple steps, reviews of partial results, possibly several iterations, and interactions with the data owners. There are several reasons to structure a KDP as a standardized process model.

The KDP model consists of a set of processing steps to be followed by practitioners when executing a knowledge discovery project. The model describes procedures that are performed in each of its steps. Since the 1990s, several different KDPs have been developed. The initial efforts were led by academic research but were quickly followed by industry. The first basic structure of the model was proposed by Fayyad et al. and later improved/modified by others. The main differences between the models described here lie in the number and scope of their specific steps. A common feature of all models is the definition of inputs and outputs. Typical inputs include data in various formats, such as numerical and nominal data stored in databases or flat files; images; video; semi-structured data, such as XML or HTML; etc. The output is the generated new knowledge usually described in terms of rules, patterns, classification models, associations, trends, statistical analysis, etc.

ACADEMIC RESEARCH MODELS:

The efforts to establish a KDP model were initiated in academia. In the mid-1990s, when the DM field was being shaped, researchers started defining multistep procedures to guide users of DM tools in the complex knowledge discovery world. The main emphasis was to provide a sequence of activities that would help to execute a KDP in an arbitrary domain. The two process models developed in 1996 and 1998 are the nine-step model by Fayyad et al. and the eight-step model by Anand and Buchner.

INDUSTRIAL MODELS:

Industrial models quickly followed academic efforts. Several different approaches were undertaken, ranging from models proposed by individuals with extensive industrial experience to models proposed by large industrial consortiums. Two representative industrial models are the five-step model by Cabena et al., with support from IBM and the industrial six-step CRISP-DM model, developed by a large consortium of European companies. The latter has become the leading industrial model, and is described in detail next. The CRISP-DM (CRoss-Industry Standard Process for Data Mining) was first established in the late 1990s by four companies: Integral Solutions Ltd. (a provider of commercial data mining solutions), NCR (a database provider), DaimlerChrysler (an automobile manufacturer), and OHRA (an insurance company). The last two companies served as data and case study sources. The development of this process model enjoys strong industrial support. It has also been supported by the ESPRIT program funded by the European Commission. The CRISP-DM Special Interest Group was created with the goal of supporting the developed process model. Currently, it includes over 300 users and tool and service providers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

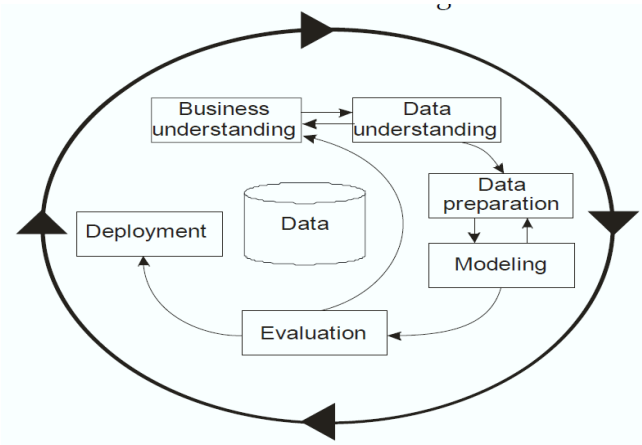

Fig 3. The CRISP-DM Process

HYBRID MODELS:

The development of academic and industrial models has led to the development of hybrid models, i.e., models that combine aspects of both. One such model is a six-step KDP model. developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research.

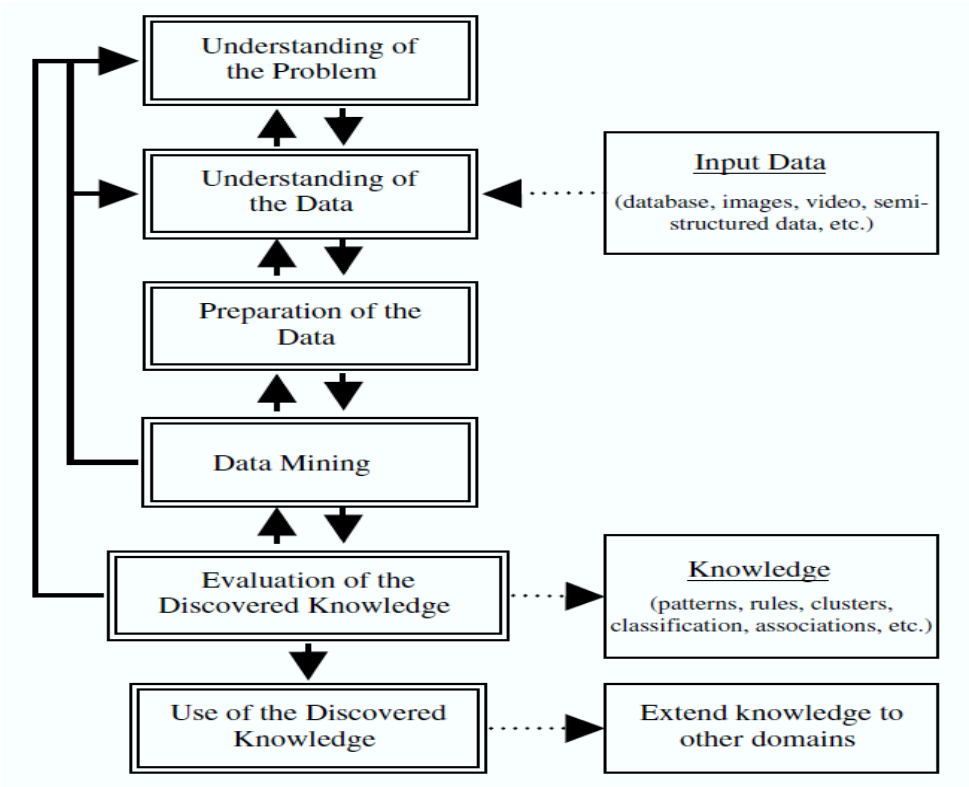

Figure 4. The six-step KDP model.(Source: Pal, N.R., Jain, L.C., (Eds.) 2005. Advanced Techniques in Knowledge Discovery and Data Mining, Springer Verlag.)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

IV. COMPARISON OF THE MODELS

Table 1. Comparison of DM & KD process models and methodologies (Kurgan & Musilek,2006)

Model	**Fayyad et al.**	**Cabena et al.**	**Anand & Buchner**	**CRISP-DM**	**Cios et al.**
No of Steps	9	5	8	6	6
Steps	Developing and Understanding of the Application Domain	Business Objectives Determination	Human Resource Identification	Business Understanding	Understanding the Data
			Problem Specification		
	Creating a Target Data Set	Data Preparation	Data Prospecting	Data Understanding	Understanding the Data
	Data Cleaning and Pre-processing		Domain Knowledge Elicitation		
	Data Reduction and Projection		Methodology Identification	Data Preparation	Preparation of the data
	Choosing the DM Task		Data Preprocessing		
	Choosing the DM Algorithm				
	DM	DM	Pattern Discovery	Modeling	DM
	Interpreting Mined Patterns	Domain Knowledge Elicitation	Knowledge Post processing	Evaluation	Evaluation of the Discovered Knowledge
	Consolidating Discovered Knowledge	Assimilation of Knowledge		Deployment	Using the Discovered Knowledge

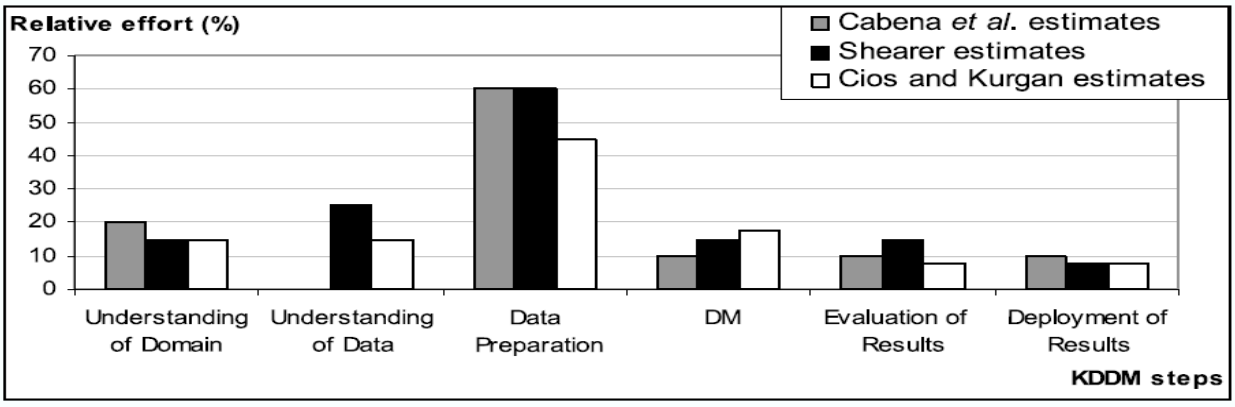

Figure 5. Relative effort spent on specific steps in the KDDM process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

V. CONCLUSION AND FUTURE WORK

This survey has provided an overview of the state-of-the-art in developing one of the most important standards, the KDDM process model. The description and comprehensive comparison of several main models has been provided, along with discussion of issues associated with their implementation. The goal of this survey has been to consolidate research in this area, to inform users about different models and how to select the appropriate model, and to develop improved models that are based on previous experiences. After analysing the SE process models, we have developed a joint model based on two standards to compare, process by process and activity by activity, the modus operandi in SE and DM & KD.

The proposed process model includes all the activities covered by CRISP-DM, but distributed across process groups that conform to engineering standards established by a field with over 40 years' experience, i.e. software engineering. The model is not complete, as the need for the processes, tasks and/or activities set out in IEEE 1074 or ISO 12207 and not covered by CRISP-DM has been stated but they have yet to be adapted and specified in detail. Additionally, this general outline needs to be further researched. First, the elements that CRISP-DM has been found not to cover at all or only in part would have to be specified and adapted from their SE counterpart. Second, the possible life cycle for DM would have to be examined and specified. Third, the process model specifies that what to do but not how to do it.

A methodology is what specifies the "how to" part. Therefore, the different methodologies that are being used for each process would need to be examined and adapted to the model. Finally, a methodology is associated with a series of tools and techniques. DM has already developed many such tools (like Clementine or the neural network techniques), but tools that are well-established in SE (e.g. configuration management techniques) are missing. It remains to be seen how they can be adapted to DM and KD processes.

REFERENCES

1. Anand, S., and Buchner, A. 1998. Decision Support Using Data Mining. Financial Times Pitman Publishers, London
2. Anand, S., Hughes, P., and Bell, D. 1998. A data mining methodology for cross-sales. Knowledge Based Systems Journal, 10:449–461
3. Brachman, R., and Anand, T. 1996. The process of knowledge discovery in databases: a human-centered approach. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining 37–58, AAAI Press
4. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., and Zanasi, A. 1998. Discovering Data Mining: From Concepts to Implementation, Prentice Hall Saddle River, New Jersey.
5. Cios, K., Teresinska, A., Konieczna, S., Potocka, J., and Sharma, S. 2000. Diagnosing myocardial perfusion from SPECT bull's-eye maps – a knowledge discovery approach. IEEE Engineering in Medicine and Biology Magazine, special issue on Medical Data Mining and Knowledge Discovery, 19(4):17–25.
6. Cios, K., and Kurgan, L. 2005. Trends in data mining and knowledge discovery. In Pal, N.R., and Jain L.C. (Eds.), Advanced Techniques in Knowledge Discovery and Data Mining, 1–26, Springer Verlag,London.
7. Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.), 1996. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge.

BIOGRAPHY

Ravindra Changala received Master Degree in Information Technology (IT) form Jawaharlal Nehru Technological University, Hyderabad (JNTUH). His research interest includes Big Data, Data Mining and Information and Communication Technologies. Presently working as an Asst.Prof in IT Department, Guru Nanak Institutions Technical Campus.

D. Rajeswara Rao is a Professor in CSE Department, K L University. He has Published couple of papers in Journals and conferences. He is guiding couple of Research scholars, PG and UG students. His research interest includes Data Mining, Big Data and soft computing technologies.

Janardhan Rao Tenneti did his double received Master Degrees in Information Technology (IT) form Jawaharlal Nehru Technological University, Hyderabad (JNTUH) and VRMF, Tamilnadu. His research interest includes Ad hoc networks and Information and Communication Technologies. He published couple of Research Paper in various

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 5, May 2015

National and International Conference and Journals. Presently working as an Associate Professor in IT Department, Guru Nanak Institutions Technical Campus.

P Kiran Kumar received Master Degree in Computer Science and Engineering (CSE) form Jawaharlal Nehru Technological University, Hyderabad (JNTUH). His research interest includes Computer Networks and Data Mining. Presently working as an Asst.Prof in CSE Department, Vignan Institute of Technology and Science.

Kareemunnisa received Master Degree in Computer Science and Engineering (CSE) form Jawaharlal Nehru Technological University, Hyderabad (JNTUH). Her research interest includes Computer Networks and Data Mining. Presently working as an Asst.Prof in IT Department, Guru Nanak Institutions Technical Campus.