ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Survey on Knowledge Discovery in Database and Challenges in KDD

S. Jagadeesh Soundappan[1], R. Sugumar[2]

Department of CSE, St. Peter's University, Avadi, Chennai, India[1]

Department of CSE, Velammal Institute of Technology, Panchetti, Chennai, India[2]

ABSTRACT: Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. The widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. In the last few years knowledge discovery tools have been used mainly in research environments, sophisticated software products are now rapidly emerging. In this paper, we provide an overview of common knowledge discovery tasks and approaches to solve these tasks and challenges of the KDD. we consider the importance of knowledge discovery to possess in order to accommodate its users effectively for the issues in which are not addressed appropriately

KEYWORDS: Knowledge discovery, Data Mining, KDP

1.INTRODUCTION

In the past two decadeshas seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. The increase in use of electronic data gathering devices such as point-of-sale or remote sensing devices has contributed to this explosion of available data Knowledge Discovery and Data mining (KDD) emerged as a rapidly growing interdisciplinary field that merges together databases, statistics, machine learning and related areas in order to extract valuable information and knowledge in large volumes of data With the rapid computerization in the past two decades, almost all organizations have collected huge amounts of data in their databases. These organizations need to understand their data and/or to discover useful knowledge as patterns and/or models from their data There is a difference in understanding the terms "Knowledge Discovery" and "Data Mining" between people from different areas contributing to this new field Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or model in data

II.KNOWLEDGE

Within literature, many definitions of knowledge can be found. In this contribution, we want to give a brief overview of some important definitions, because this term is vital for our further elaborations. In subsequent sections, we refer to the following definitions and explanations. Davenport/Prusak[1] define knowledge as follows: "Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms." Therefore, knowledge comprises both information and person-specific aspects like experiences, values, and insights. Nonaka/Takeuchi[2] understand knowledge as the flow of information in combination with intrinsic attitudes and values of the receiver. Thus, information as well as knowledge can be seen as context -specific and it can be deduced that knowledge is strongly linked to individuals within organizations.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Davenport/Prusak point out that skills based on knowledge, like to organize, to select, and to judge, probably depend more on values and intrinsic attitudes of an individual than on explicit available information and logic. Even though a differentiation between data, information, and knowledge is important, we only give a brief definition of data and information. Data characterizes several objective facts of events or processes. Thus, it is meaningless symbols and it is machine readable. Information is a flow of messages or processed data; it is data, which has an effect.
An important characteristic of knowledge and difference to information is the strong affinity of knowledge to activities Individuals act and react due to their experiences and intrinsic attitudes. Murray states: "All doing is knowing and all-knowing is doing (...) We admit knowledge whenever we observe an effective behavior in a given context." Thus, in the sense of semiotics, knowledge contains a pragmatic dimension.
Knowledge is much more than transformed information, which can be presented in the form of information or data. Sveiby [3] points out: "Explicit knowledge in the form of facts is therefore 'metaphorically speaking' only the tip of the iceberg." This statement indicates the distinction between implicit and explicit knowledge, which is often discussed in literature. This differentiation was first defined by Polanyi [4], who developed a concept for implicit knowledge, which he described as follows: "We can know more than we can tell." For further discussions of different types of knowledgerefer to Biggam [5] and Lai/Chu [6]. In particular, Nonaka/Takeuchi discuss some philosophical approaches for the definition of knowledge and classifications of different knowledge types.

A. KNOWLEDGE DISCOVERY PROCESS

Researcher in [7] – [9] 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.

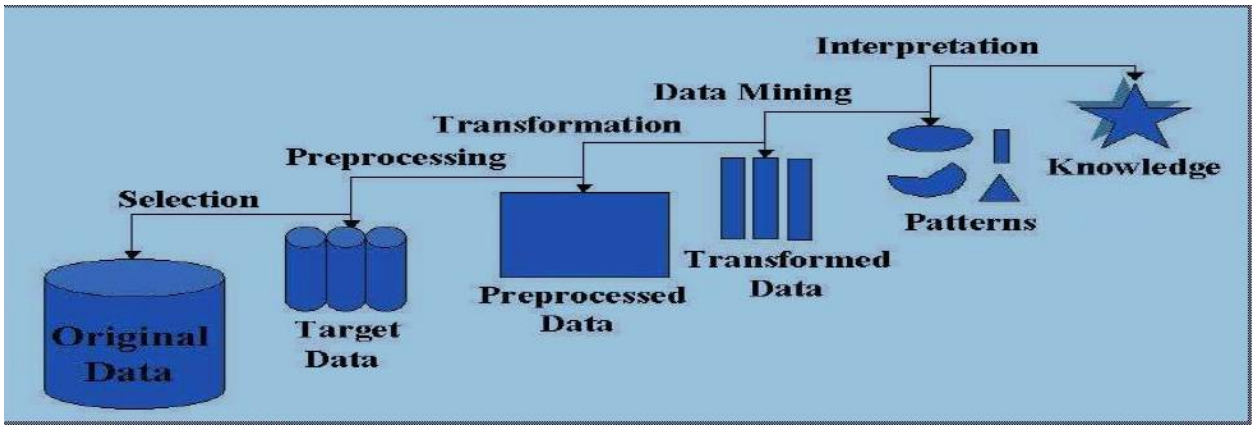

Fig 1. Steps in KDD

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)

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

III. KNOWLEDGE DISCOVERY PROCESS MODELS

Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski [10] 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 analyzing the application domain.

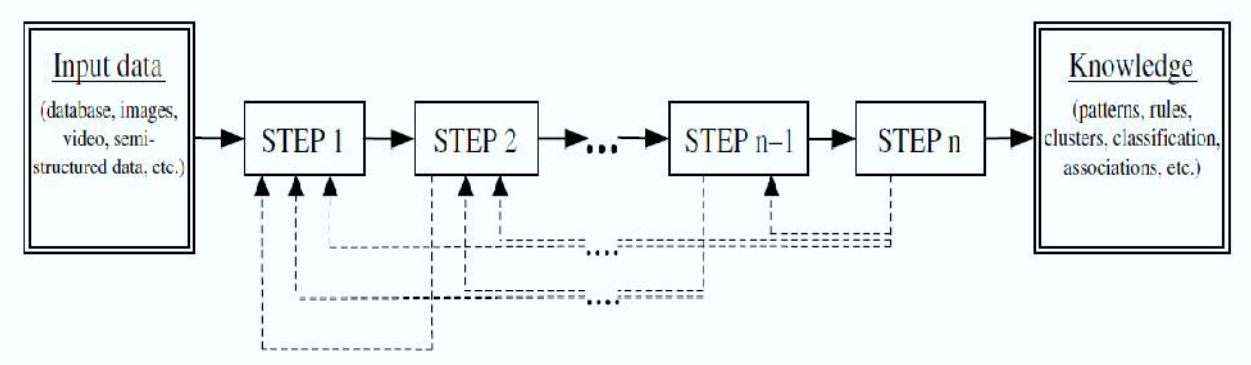

Fig 2. Sequential Structure of KDP

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.

A. 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.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

B. 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 thefive-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), andOHRA (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. 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.

IV. KDD DATA SET

Hettich, S. and Bay, S. D. [11] This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is an automatic, exploratory analysis and modelling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex datasets. DM is the core of the KDD process, involving the inferring of algorithms that explore the data, develop the model and discover previously unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction. The accessibility andabundance of data today makes knowledge discovery and DM a matter of considerable importance and necessity. Given the recent growth of the field, it is not surprising that a wide variety of methods are now available to the researchers and practitioners. No method is superior to others for all cases. The handbook of DM and knowledge discovery from data aims to organize all significant methods developed in the field into a coherent and unified catalogue; presents performance evaluation approaches and techniques; and explains with cases and software tools the use of the different methods. In that dataset, we are extracted three attacks and one normal, the attacks are DOS, Probe, and RLA.

V. CHALLENGES FOR KDD

Usama Fayyad, Gregory Piatetsky-Shapiro, andPadhraic Smyth [12]. The below mentioned were the challenges in the Knowledge Discovery in Databases

Larger databases: Databases with hundreds of fields and tables, millions of records, and multi-gigabyte size are quite commonplace, and terabyte (1012 bytes) databases are beginning to appear.

High dimensionality: Not only is there often a very large number of records in the database, but there can also be a very large number of fields (attributes, variables) so that the dimensionality of the problem is high. A high dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorial explosion manner. In addition, it increases the chances that a data-mining algorithm will find spurious patterns that are

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables.

Over-fitting: When the algorithm searches for the best parameters for one particular model using a limited set of data, it may over-fit the data, resulting in poor performance of the model on test data. Possible solutions include cross-validation, regularization, and other sophisticated statistical strategies.

Assessing statistical significance:A problem (related to over-fitting) occurs when the system is searching over many possible models. For example, if a system tests N models at the 0.001 significance level then on average with purely random data, N/1000 of these models will be accepted as significant. This point is frequently missed by many initial attempts at KDD. One way to deal with this problem is to use methods that adjust the test statistic as a function of the search.

Changing data and knowledge: Rapidly changing (non-stationary) data may make previously discovered patterns invalid. In addition, the variables measured in a given application database may be modified, deleted, or augmented with new measurements over time. Possible solutions include incremental methods for updating the patterns and treating change as an opportunity for discovery by using it to cue the search for patterns of change only.

Missing and noisy data: This problem is especially acute in business databases. U.S. census data reportedly has error rates of up to 20%. Important attributes may be missing if the database was not designed with discovery in mind. Possible solutions include more sophisticated statistical strategies to identify hidden variables and dependencies.

Complex relationships between fields: Hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the contents of a database will require algorithms that can effectively utilize such information. Historically, data mining algorithms have been developed for simple attribute-value records, although new techniques for deriving relations between variables are being developed.

Understandability of patterns: In many applications it is important to make the discoveries more understandable by humans. Possible solutions include graphical representations, rule structuring with directed a-cyclic graphs, natural language generation, and techniques for visualization of data and knowledge.

User interaction and prior knowledge: Many current KDD methods and tools are not truly interactive and cannot easily incorporate prior knowledge about a problem except in simple ways: The use of domain knowledge is important in all of the steps of the KDD process.

Integration with other systems:A stand-alone discovery system may not be very useful. Typical integration issues include integration with a DBMS (e.g., via a query interface), integration with spreadsheets and visualization tools, and accommodating real-time sensor readings.

VI. CONCLUSION

This study supports the understanding of the role of knowledge discovery methods. The presented methods enable the identification and generation of knowledge, and support companies' ability to explore and assess their knowledge bases. Exploring different methods of knowledge discovery will potentially be an important research field of organizational knowledge management as the importance of the resource knowledge as a success factor increases continuously. Data cleaning, transformation and clustering are important tasks in many contexts such as data warehousing, data mining and data integration. The current approaches are time consuming and frustrating due to long-running, non-interactive operations, poor coupling between analysis and transformation, and complex transformation interfaces that offer require user programming.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

REFERENCES

[1] Davenport, Thomas H.; Prusak, Laurence: Working Knowledge – How OrganizationsManage What They Know; Harvard Business School Press, Boston, Massachusetts,1998.
[2]Nonaka, Ikujiro; Takeuchi, Hirotaka: The Knowledge-Creating Company – How Japanese Companies Create the Dynamics of Innovation; Oxford University Press,New York, Oxford, 1995.
[3] Sveiby, Karl Erik: The new organizational wealth: Managing & measuring knowledgebasedassets; Berrett-Koehler Publishers, San Francisco, CA, 1997.
[4] Polanyi, Michael: The Tacit Dimension, 1966; in: Prusak, Laurence: Knowledge inOrganizations, Butterworth-Heinemann, Boston, 1997.
[5] Biggam, J.: Defining Knowledge: An Epistemological Foundation for KnowledgeManagement, Proceedings of the 34th Annual Hawaii International Conference onSystem Sciences, IEEE Computer Society Press, Los Alamitos, January 2001.
[6] Lai, Hsiangchu; Chu, Tsai-hsin: Knowledge Management: A Review of TheoreticalFrameworks and Industrial Cases, Proceedings of the 33rd Annual HawaiiInternational Conference on System Sciences, IEEE Computer Society Press, LosAlamitos, January 2000.
[7] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
[8] Osmar R. Zaïane, CMPUT690 Principles of Knowledge Discovery in Databases1999
[9] LUKASZ A. KURGAN1 and PETR MUSILEK2 "A survey of Knowledge Discovery and Data Mining process models" The Knowledge Engineering Review, Vol. 21:1, 1–24. 2006
[10] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski “A Knowledge Discovery Approach"2007
[11] Hettich, S. and Bay, S. D. "The UCI KDD Archive" Irvine, CA. KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. 1999
[12] Usama Fayyad, Gregory Piatetsky-Shapiro, andPadhraic Smyth "From Data Mining to Knowledge Discovery in Databases"AI Magazine Volume 17 Number 3 1996