Towards Optimization and Validation of Different Data Mining Models to Support Knowledge Retrieval

Rohit Kumar Sahu[1], Nirbhay Kumar Kushwaha[2], Anirban Bhar[3], Shambhu Nath Saha[4]

[1,2] *B. Tech student, Department of Information Technology, Narula Institute of Technology, Kolkata, India.*
[3] *Assistant Professor, Department of Information Technology, Narula Institute of Technology, Kolkata, India.*
[4] *Associate Professor, Department of Information Technology, Narula Institute of Technology, Kolkata, India.*

ABSTRACT

Advancements in technologies various new area have emerged. There are various fields like science, engineering, health, business which collect huge amount of data every day. The necessity for effective procedures for information and knowledge location, selection, and retrieval among users has been made even more acute by the internet's rapid development and uptake. Data mining is an important and new area in technologies which manages and extracts the required information. Data mining also known as knowledge discovery in database (KDD) involves extracting interesting, explanatory, and useful information form raw data. This is main reason why the data mining is growing rapidly. This work shows the process of data mining and review its applications such as educational data Mining (EDM), finance, life science and medicine, how it can aid those who make decisions in choosing wisely. The degree and effectiveness of knowledge retrieval are considerably increased by realizing knowledge retrieval through a variety of ways, levels, and multi-modes. Our classification throughout this paper focuses exclusively on research that has been carried out in the recent period. With this classification, the focus is to show the various steps taken in the process of data mining and we present an easy or concise view of the different models that have been optimized for combining knowledge search with data mining technologies.

Keywords: *- Data mining, Knowledge discovery in database (KDD), Knowledge search, Knowledge retrieval.*

1. INTRODUCTION

Data mining can be defined as the process of extracting validated, previously unknown and actionable information form an ocean of data. Information might be static as well as dynamic and that is an important step in the pursuit of knowledge database (KDD). The main purpose of data mining is to use unthread information to make some important decisions in the field of engineering, health, business and science. Data mining tools can provide the answers to queries that were previously too time-consuming to find in the knowledge driven data environment. They set up databases for discovering hidden patterns and predicting information that experts would overlook because it deviates from what they would normally expect. Data mining popularly known as the Knowledge Discovery in Database (KDD), it is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data in database. Data mining is a step in the knowledge discovery process, even though the terms knowledge discovery in databases (or KDD) and data mining are frequently used interchangeably. Data mining is concerned with a sub-filed of statistics called exploratory data analysis and its sub-filed is Artificial intelligence is called knowledge discovery and machine learning.

2. CATEGORIZATION OF DATA MINING AND THEIR TASKS

The many types of patterns that need to be recognized in data mining operations are perceived by data mining functionality. Data mining characteristics are used to specify the kinds of patterns that will be found during data mining operations. Big data can be forecasted and characterized via data mining, which has several applications. The two main categories of data mining tasks are descriptive and predictive.

2.1 Descriptive Data Mining

Descriptive data mining is a ‘description’ of data that provides a detailed description of the data -for example it gives information about what is happening inside the data without any forethought. An association technique that uses a prior algorithm to characterize a student’s performance in finding correlation between a set, without prior consideration to understand.

This type includes the following functions: Association Rules, Clustering, Summarization, and Sequence Discovery.

2.2 Predictive Data Mining

This allows user to consider features that are not specifically available. For example, the projection of the previous monthly output as well as the market analysis in the next month. The predictive data mining is mainly focus on supervised methods used for prices forecasting, education service.

Predictive can be further characterized into four other parts: Classification, Regression, Prediction, and Time Series Analysis.

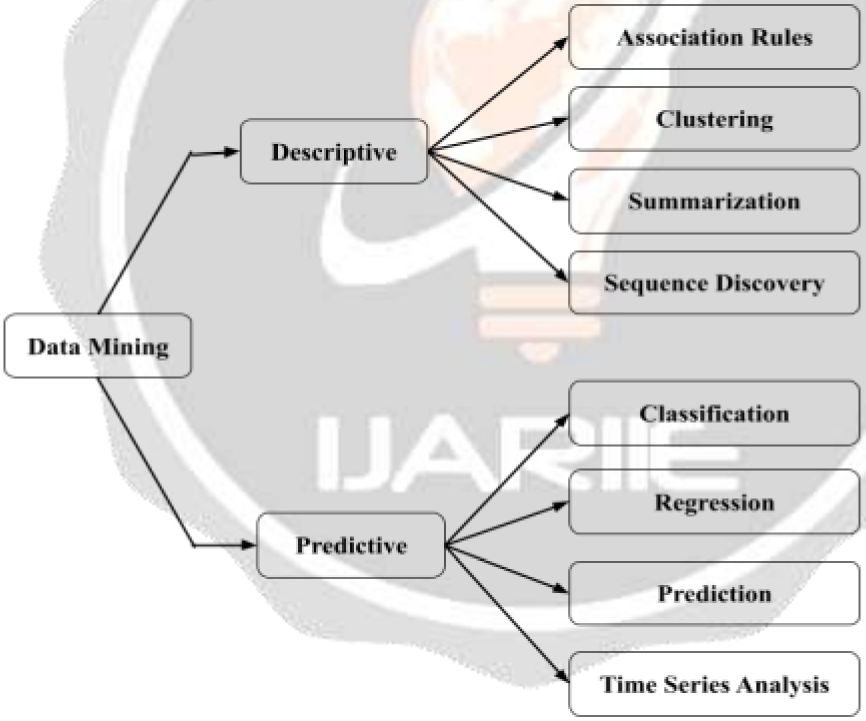

Fig -1: The data mining tasks

3. TYPES OF DATA MINING SYSTEM

A data mining system can be classified according to the various criteria and integrate techniques. the classified to following criteria:

- Classification of data mining system based on the databases mined: we can classify a data mining system according to the kind of databases mined. Database can be classified in various criteria like data model, types of data etc.
- Classification of data mining system based on the kind of knowledge mined: we can classify a data mining system according to the kind of knowledge or data mining. Knowledge mined can be classified based on functionalities such as – characterization, discrimination, classification, prediction, outlier, and evolution analysis etc.
- Classification of data mining system based on the techniques utilized: we can classify a data mining system according to the kind of techniques used. Technique utilized can be classified to the degree of user interaction involved or the methods of analysis employed.
- Classification of data mining based on the kind of application adapted: we can classify a data mining system based on the kind of applications adapted. application adapted can be classified are as finance, telecommunication, DNA, stock markets, E-Mail etc.

4. THE PROCESS

Data mining refers to extracting or mining knowledge from large amounts of data. Data mining process is a step-by-step process which cannot be completed in one step. In other words, we may not get the information you need so easily from the large amount of data. Thus, data mining should have been more aptly named as knowledge mining which emphasizes mining from large amounts of data. It is the computational process of discovering data in large amount of data. Basically, the process has evolved from the knowledge that the major objective is to create large data project to run the data mining process more efficiently.

4.1 Data Mining Process

Business Understanding

The business understanding phase mainly focuses on understanding project objective on assessment and needs, current status, set data mining goals form business point of view. In this phase we are converting knowledge into a data mining problem definition and a preliminary plan designed to achieves the objective.

Data Understanding

In the Phase the initial data collection starts with activities like, Data description, Data exploration and Validation of data quality. It is basically about with establishing the main characteristics of data which include the data structures, data quality, and identifying any interesting subsets of the data.

Data Preparation

This phase includes all the actives for the creation of the data. Data preparation is the process of collecting, combining, structuring, and organizing data. The main activities during this step select the data cleaning, data integration and data transformation. the output data of this phase is set that can be used in modeling, business intelligence, analytics, and data visualization application.

Modelling

In this phase, various modelling techniques are selected, we selected the modelling techniques, Evaluate the model built based on the modeling parameters and business objectives. Various activities carried out during this stage the modeling techniques are selected.

Evaluation

Evaluation is an important point in any data mining process. This is serves two purposes: predicting how well the final model will work or even whether it should be used in the future and as an integral part of many learning method, which helps to find the models that model. This phase is the model, and the modeling phase is verified in terms of achieving the business or education objectives. The various activities undertaken during this phase include evaluation of result, review process.

Development

Organize and present the knowledge gained in this phase as a model. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process. the main purpose of this model is to increase the knowledge of the data. This phase is the execution phase various tasks in phase i.e., planning deployment, plan monitoring, maintenance, production, final review, and good report. Therefore, in the deployment phase, patterns are deployed for the desired result.

4.2 Knowledge Retrieval System

Structured knowledge is congruent with human thought processes and is simple to comprehend. Sometimes people lack contextual awareness or may not know exactly what they want. Users will find it very helpful to investigate and modify the question if knowledge can be presented visually in a systematic fashion. A conceptual model of a typical knowledge retrieval system is shown in Fig. 1. The primary procedure is best explained as follows:

Knowledge Discovery
Finding information from sources using techniques like data mining, machine learning, knowledge acquisition, and others.

Query Formulation
The process of creating queries based on user requirements and input. Artificial and natural languages are both acceptable as inputs.

Knowledge Selection
Choosing from a set of potentially relevant knowledge based on the user's query and knowledge obtained from data or information sources.

Information Structure Construction
Using a variety of knowledge perspectives, domain expertise, user context, etc. to construct knowledge structures Expert systems can provide domain knowledge. User logs might reveal user preferences and backgrounds.

Exploration and Search
Investigating the knowledge architecture to gain a broad understanding and focus the search. Users can dig further by comprehending the necessary information structures.

Knowledge Structure Reorganization
If users need to investigate additional facets of a certain body of knowledge, knowledge structures may be reorganized.

Query Reformulation
If the created structures are unable to meet the user's needs, reformulating the query. Knowledge is displayed in a structured way as one of the main characteristics of knowledge retrieval systems, allowing users to gain contextual awareness of related knowledge and perform further retrieval [1].

4.3 Data Mining and Knowledge Retrieval

Information retrieval serves as the foundation for knowledge retrieval, which is being developed gradually. In order to assist semantic reasoning and retrieval, ontology-based knowledge organization may effectively express the contents of information items and the numerous semantic interactions between them [2]. An efficient method for semantic-level knowledge retrieval is provided by data mining, ontology, and semantic technology [3]. Finding undiscovered, potentially useful information and knowledge from a vast volume of data is a technique known as data mining [3],[4]. Future uses for data mining are numerous, and both the academic and corporate communities have given it considerable attention. Data mining and knowledge retrieval have essentially the same research goals. They work well together and can successfully obtain in-depth knowledge from information resources.
Fig. 2 displays the key elements of the proposed data mining and knowledge retrieval model. They are knowledge base, data mining, information retrieval, semantic analysis, and knowledge retrieval.

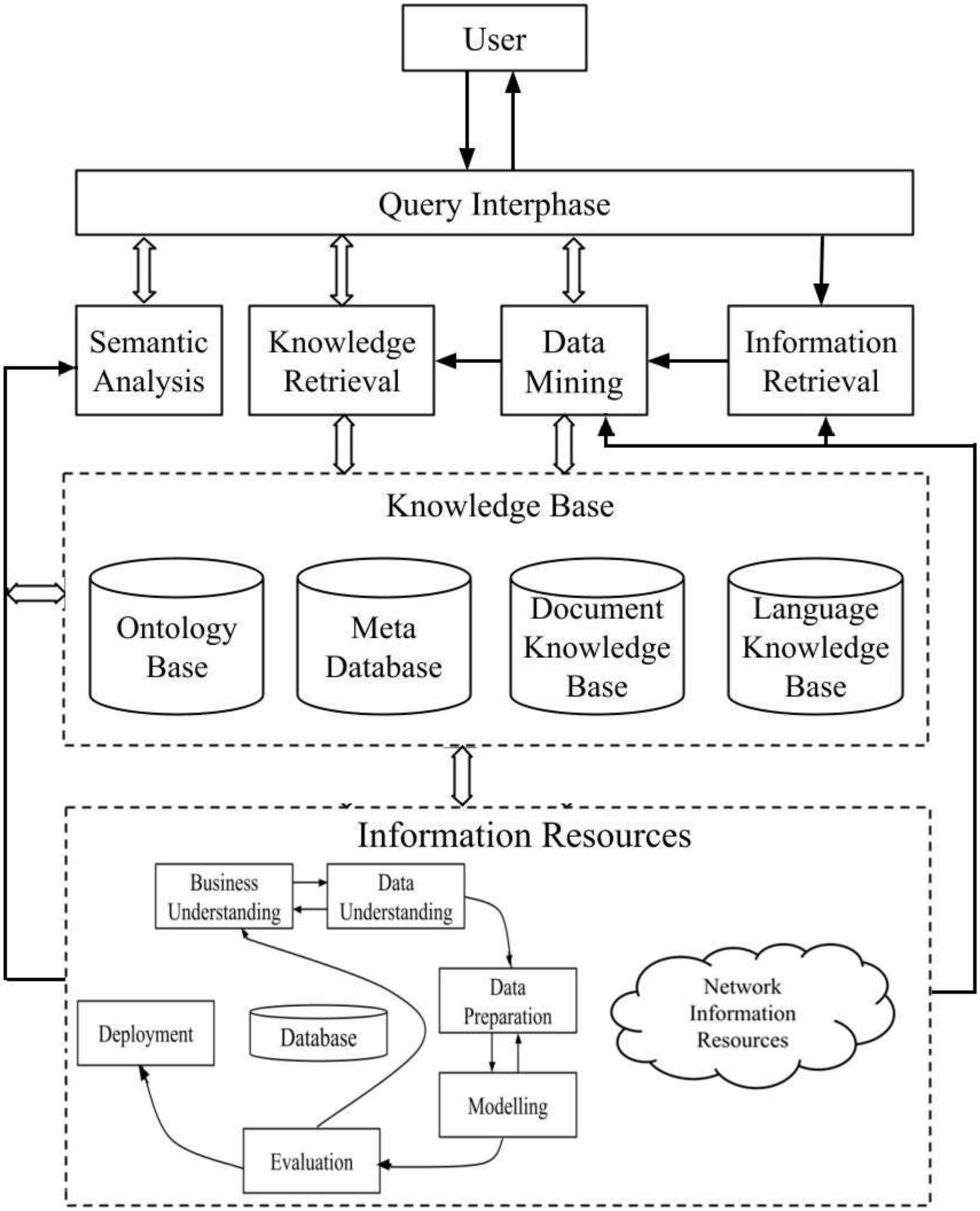

Fig -2: The data mining and knowledge retrieval model

Semantic Analysis

This section's primary job is to apply ontological knowledge to examine the semantics of user inquiries, search results, and database. The semantic query expression is then constructed using ideas and their associations. The knowledge base is also being built and enhanced after the semantic annotation and analysis of search results and information components. In the retrieval context, rich semantic information is concealed [3], [5].

Knowledge Retrieval

The method for retrieval makes it possible to mine document knowledge bases or other information sources for documents. The primary retrieval techniques include ontology association retrieval, semantic reasoning retrieval, classification retrieval based on inductive learning, concept retrieval based on association analysis, and retrieval based on association analysis. The information retrieval component serves as an adjunct mechanism for knowledge retrieval.

Data Mining

Data mining is used throughout the process of building knowledge resources, searching, and analyzing the results. Its primary duties consist of: Apply mining processing on the information retrieval findings. The processing consists of information extraction, automatic summarization, association analysis, cluster analysis, inductive learning, and classification analysis. The user may then be shown the mining findings, or they may be utilized to improve user inquiries and broaden the knowledge base. Utilize data mining as part of the search process. Direct knowledge mining from document knowledge bases, databases, and network information sources is also possible.

Knowledge Base

The ontology base, often referred to as the concept base, includes all of the high-level abstract concepts of knowledge resources, the relationships between these concepts, and the description of the basic knowledge patterns. All of the metadata descriptions of information resources are contained in the meta database. It serves as the building block for mining and retrieving knowledge. The primary resources for knowledge retrieval are the Documents Knowledge Base, which includes the semantic description of information items and their relationships. The storage of linguistic information needed by the system, namely dictionaries, grammar, and semantic knowledge, is done using language knowledge bases. This information is utilized to enable semantic analysis and process.

All of the metadata descriptions of information resources are contained in the meta database. It serves as the building block for mining and retrieving knowledge. The primary resources for knowledge retrieval are the Documents Knowledge Base, which includes the semantic description of information items and their relationships. The storage of linguistic information needed by the system, namely dictionaries, grammar, and semantic knowledge, is done using language knowledge bases. This information is utilized to enable semantic analysis and process.

5. FUTURE TRENDS

Data mining and knowledge retrieval have become prominent computer science disciplines thanks to the widespread success of their many applications, and they have showed promise for future advancements. Future application fields and ever-evolving technology present data mining with new opportunities and difficulties; typical future developments in data mining include the incorporation of knowledge discovery like:

- Standardization
- Data preprocessing
- Knowledge mining
- Complex objects of data
- Computing resources
- Web mining
- Scientific Computing
- Business data

The term "data mining" comes from the analogy between mining a mountain for a vein of valuable ore and looking for useful business information in a vast database, such as locating related products in terabytes of store scanner data. In both cases, locating the value requires either combing through a vast amount of information or carefully probing it. Given large, high-quality datasets, data mining technologies can create new business prospects by offering the following features:

- Automated prediction of trends and behaviors
- Automated discovery of previously unknown patterns
- Artificial neural networks
- Decision trees
- Genetic algorithms
- Nearest neighbor method
- Rule induction

6. CONCLUSIONS

We have covered a thorough analysis of data mining in this paper, covering a range of topics including jobs, approaches and applications. Clarifying the connection between knowledge discovery and data mining is a fundamental goal. We gave an overview of the KDD procedure and the fundamentals of data mining. By using data mining techniques, users will be able to extract useful information from virtually integrated data. These methods have a wide range of applications in sectors like biomedicine, retail, and telecommunications. These technologies forecast future trends and behaviors, enabling companies to be proactive and deliver information in a way that is simple for people to understand. The fundamental approaches to data mining have been the emphasis. Information extraction and natural language processing are some of the techniques. We can manage vast amounts of data more easily thanks to data mining techniques. The majority of data mining tasks can be completed manually by a human, but doing so would take a long time.

7. REFERENCES

[1]. Yiyu Yao, Yi Zeng, Ning Zhong, Xiangji Huang. “Knowledge Retrieval (KR)”, In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, Silicon Valley, USA, November 2-5, 2007, 729-735.

[2]. K.Z.Gao,Z.Q.Bao,X.Q.Li, "Study on Two-layer Knowledge Retrieval Technology in Conceptual Design," Grey Systems and Intelligent Services, 2007, GSIS 2007, pp.1523-1527, 2007.

[3]. YanHao, Yu-feng Zhang, “Research on Knowledge Retrieval by Leveraging Data Mining Techniques”, 2010 International Conference on Future Information Technology and Management Engineering.

[4]. K.Z.Gao,Z.Q.Bao,X.Q.Li, "Study on Two-layer Knowledge Retrieval Technology in Conceptual Design," Grey Systems and Intelligent Services, 2007, GSIS 2007, pp.1523-1527, 2007.

[5]. L.M.Shao,S.G.Zhang,X.S.Suo, "Research on Ontology Knowledge Retrieval to the Expert Consultation of Greenhouse, "Communication Technology, 2006, ICCT '06,pp.1-4,2006.