A Review on Knowledge Discovery from Databases

Niraj Singhal and Himanshu

Abstract Knowledge discovery is defined as the method used for discovering interesting, previously unknown and potentially useful patterns from a massive amount of data. It is an integrative area of research, including illustrative work from areas such as database technology, machine learning, and pattern recognition, extraction of valuable information, neural network, artificial intelligence, high-performance computing and data visualization. The process of finding knowledge from the data is also getting more important as the data is increasing every day. This paper discusses the process of knowledge discovery and also gives description about the challenges faced when knowledge is discovered. It also presents the work done in the related area and their comparative analysis.

Keywords Knowledge discovery · Machine learning · Data mining · Complex data · Pattern evaluation

1 Introduction

Knowledge discovery in databases (KDD) is a dynamic field of research that promises high returns in many professional and scientific domains. The corporate, government and scientific communities are being shunned by the incursion of data that is consistently stored in online databases. Analysis of this data and extracting out some meaningful pattern in a felicitous manner is impractical. The process of KDD involves searching for useful knowledge from the data gathered from various sources. The current scenario is characterized by increasing enormous amount of data and all kinds of human efforts are being generated and shelved. This vast amount of data is recorded as computer databases are managed by computer technology in an easy way. Data is being collected and assembled across a wide variety of areas at a dramatic

N. Singhal (✉)
Shobhit Institute of Engineering and Technology (Deemed to-be-University), Meerut, India
e-mail: niraj@shobhituniversity.ac.in

Himanshu
Swami VivekanandSubharti University, Meerut, India

P. K. Mallick et al. (eds.), *Electronic Systems and Intelligent Computing*, Lecture Notes in Electrical Engineering 860, https://doi.org/10.1007/978-981-16-9488-2_43

pace. New development of tools and computational theories are urgently required to help human being in withdrawing valuable information from the enormously rising proportions of digitalized data [1]. These tools and theories together constitute the main part of the dynamic field of KDD.

At abstraction levels, the field of KDD is mainly concerned with developing various strategies and methods for generating meaningful data. The major problem came across knowledge discovery is mapping low-level data into another more abstract, compact and useful form [2]. Fundamentally the process includes implementation of specific data mining techniques for pattern recognition and extracting useful knowledge.

Many techniques used for handling these tasks include cluster analysis, regression analysis, multidimensional analysis, numerical taxonomy and several other statistical methods. Many practical problems are solved by using such techniques. However, they are mainly focused on extraction of quantitative and statistical data, and as such they have some limitations. While discovering knowledge from single data source, the problem lies in that only of one type and less amount of information is obtained. So, there is a need of efficient methods to collect the vital information from multiple data source [2]. In this paper, a survey of various approaches which are useful in this area has been carried out. A comparative study of the approaches has also been presented.

2 Related Works

Knowledge discovery is referred to as the process of observing hidden designs and patterns from an enormous volume of the data sets. It includes transformation of the obtained patterns into comprehensive and easily understandable information. The domain of knowledge discovery comprises various processes that are carried out at various stages through which the basic rules of the knowledge discovery domain are made. It involves the possible analysis and interpretation of the evaluated patterns to decide what is called knowledge [3]. This includes schematic enciphering, preprocessing, sampling and projections of data before we move for data mining.

2.1 Steps Involve in Knowledge Database Discovery Process

Seven major aspects should be considered before selection of databases for their analysis. To understand the database prerequisite knowledge is required which is as follows [4]

- Cleaning of Data—The process of removing the irrelevant and noisy data from gathered data.

- Integration of Data—The process of combining adverse data collected from numerous sources into one common source.
- Selection of Data—The process in which the relevant data required for analysis is decided and retrieved from data collection.
- Data Transformation—The process of converting data to the appropriate form as required.
- Data Mining—The process of applying techniques so that potentially useful patterns can be extracted.
- Pattern Evaluation—The process of identification of unrevealed patterns to represent knowledge.
- Knowledge representation—The process in which data mining results are represented using visualization tools.

The complete knowledge discovery process is shown in Fig. 1.

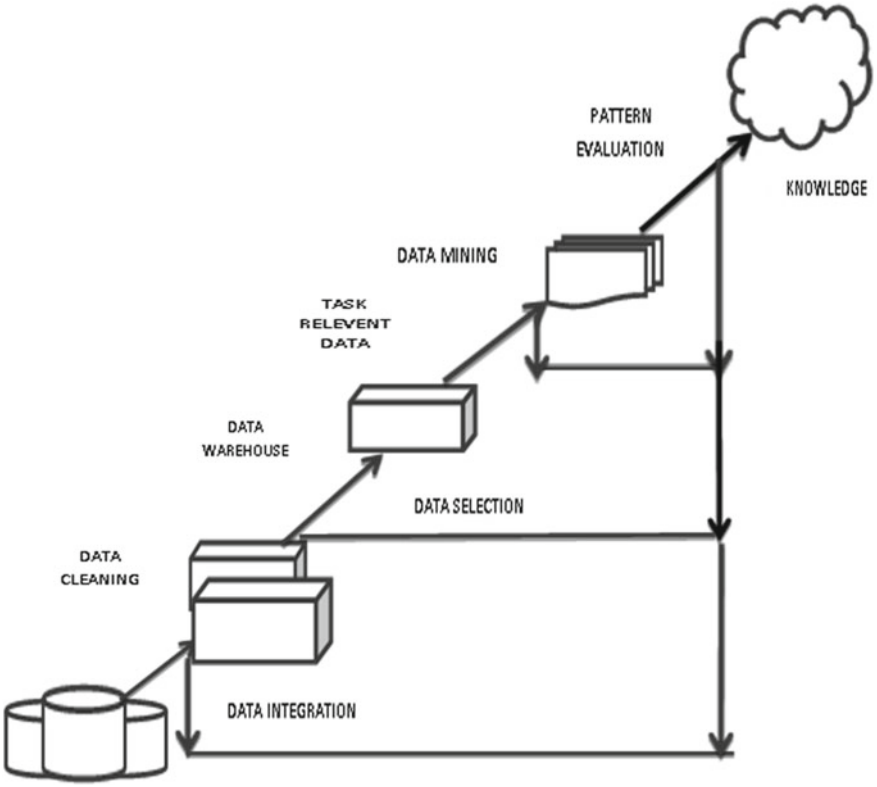

Fig. 1 Knowledge discovery process

2.2 Issues and Challenges in Knowledge Discovery

Knowledge discovery is developing into a trusted discipline; however, there are still many challenges that need to be resolved. There are some issues and challenges those are identified in knowledge discovery process [5].

- Noisy and Incomplete Data—"Data Mining is the way of acquiring information from massive volumes of data". Generally, the data collected is heterogeneous and noisy. The extensive amount of data is irregular and unreliable. Such kinds of issues may arise due to human errors or due to instruments which are used for data collection.
- Distributed Data—In this process, the data is passed through many stages. It can be easily carried out using internet or through individual systems. It is critical to unify all the data because of organizational and technical reasons.
- Complex Data—The data obtained is really heterogeneous which may include text, spatial data, audios, videos, images, words, etc. It is tough to handle such diverse kinds of data and focus on requisite vital information. Sometimes we need to create new systems and equipment's to separate crucial facts and information from the data.
- Performance—Fundamentally, the presentation of the data mining framework is dependent on efficiency of techniques and methods used. If algorithms and techniques used are insufficient, then it adversely affects the presentation of data mining.
- Scalability and Efficiency of the Algorithms—Efficient and scalable algorithms are to be used in order to extract valuable information from vast amount of data.

3 Review of Literature

Knowledge discovery covers a wide area of research. The work done in the area is as follows.

Silwattananusarn and Tuamsuk (2012) discussed the suitable methods and techniques which are needed in future to serve the requirements of data mining field as it is becoming more complex day by day [6]. According to Tomar et al. (2013) data mining is most active and likeable area of research which is capturing its attention in medical applications. [7]. Fan et al. (2014) explained knowledge discovery as the capability of obtaining useful statistics from a wide variety of datasets that because of its variability, volume and velocity [8]. Saurkar et al. (2014) described data mining as "interdisciplinary field which includes integrated databases, machine learning technique, artificial intelligence, statistical approaches etc.". The data mining technique helps in extraction of hidden information and knowledge by digging deep into the data [4]. Real time analysis of streaming data is becoming the most efficient and fastest way for obtaining useful knowledge (Bifet et al. 2014). This allows firms to respond rapidly whenever a trouble appears to ascertain the enhanced performance

[9]. Purcell (2014) stated that knowledge discovery databases consist of unstructured, semi-structured and structured data sets which cannot be handled using the traditional methods and systems. Data storage technique is used for object-based storage [10].

Reddi and Indira (2014) explained that a combination of heterogeneous, homogenous, unstructured, semi-structured data is known as big data. It also suggested a model for shifting and handling of vast quantity of data over the network [11]. Ibrahim et al. (2014) said that due to the presence of partitioning skew a huge amount of data transfer occurs which causes negligence on the reduce input among different data junctions and also develop a novel algorithm named LEEN [12]. Gamache et al. (2015) discussed the idea of linking various text mining techniques to convert the unstructured data in the forms of texts into structured data in the forms of numbers so that various statistical and mathematical algorithms can be applied [13]. Baker et al. (2015) primarily deal with the development of techniques which can be used for analysis and discovery of novel and useful information [14]. Soni (2015) discussed the prediction of future sales and trends based on patterns related to customer's behaviour. This helps in increasing profits by assisting policymakers in decision-making [15].

Kaplan and vakili (2015) proposed a version to generate a text primarily-based degree of understanding recombination that they ultimately comprise as an impartial variable into their econometric version [16]. Kumar and Chatterjee (2016) focused on clarifying the relationship between techniques applicable for data mining and knowledge discovery and also discussed the data mining techniques, specialized methods for certain type of data and field [2]. Angus (2018) figured out document similarity measures in order to explore the link between search distance and firm performance [17]. Hariri et al. (2019) discussed about the capability in creating and managing information that has been a dominant factor in the growing era of technology [18].

Sankari and Shraddha (2019) introduced the application of data mining techniques on information generated from educational settings. The usage of educational data mining and analysis of data about learners and their contexts is the key to successful inference model of educational data [19]. Kumar and Basha (2020) have discussed the methods of accessibility of high volume of text-based data that needs to be examined for retrieving information [20]. Roozbahani and Rajabzadeh (2020) focused on past and current status of researches on big data in the medical and science-related areas [3]. Abdualgalil and Abraham (2020) focused on machine learning for knowledge discovery in big data. According to him machine learning needs to be more explorable, so that interacting with various kinds of data will become easier for a learner [1]. Lauw and Wong (2020) have discussed original research results, current new ideas and advanced experiences from all knowledge discovery-related areas such as data mining, machine learning, artificial intelligence, decision-making systems and other emerging applications [21].

4 Comparative Study

All the approaches discussed in earlier section use different techniques. These also differ on various parameters like technique used, database type, accuracy, sensitivity, specificity and fidelity. A comparative analysis of all approaches discussed is presented in Table 1.

Table 1 A comparative analysis

Name of authors	Technique used	Database type	Accuracy	Sensitivity	Specificity	Fidelity
Silwattananusarn et al. (2012)	Machine learning	Single	Yes	Yes	No	Yes
Tomar et al. (2013)	Machine learning	Multiple	Yes	–	No	No
Fan et al. (2014)	Classification	Single	Yes	Yes	Yes	No
Saurkar et al. (2014)	Clustering	Single	Yes	–	Yes	Yes
Bifet et al. (2014)	Machine learning	–	No	Yes	Yes	No
Purcell (2014)	–	Single	Yes	No	Yes	–
Reddi et al. (2014)	Classification	Single	Yes	Yes	No	Yes
Ibrahim et al. (2014)	OLAP	Single	Yes	–	Yes	No
Miner et al. (2015)	–	Single	No	Yes	Yes	No
Baker et al. (2015)	Machine learning	Multiple	Yes	Yes	Yes	No
Soni et al. (2015)	–	Single	Yes	–	Yes	Yes
Kaplan et al. (2015)	–	–	No	Yes	No	Yes
Ajay Kumar et al. (2016)	Machine learning	Single	Yes	–	No	No
Rüdiger et al. (2017)	Classification	Single	Yes	–	Yes	Yes
Angus (2018)	Clustering	Single	Yes	No	Yes	No
Hariri et al. (2019)	Machine learning	–	Yes	No	Yes	–
Sankari et al. (2019)	Machine learning	Single	Yes	Yes	No	Yes
Kumar et al. (2020)	Regression	Multiple	Yes	–	No	No

(continued)

Table 1 (continued)

Name of authors	Technique used	Database type	Accuracy	Sensitivity	Specificity	Fidelity
Roozbahani et al. (2020)	Machine learning	Multiple	Yes	Yes	Yes	No
Abdualgalil et al. (2020)	–	Single	Yes	–	Yes	Yes
Lauw et al. (2020)	Machine learning	Multiple	No	Yes	Yes	No

5 Conclusion

The knowledge discovery process primarily aims at finding out the exact information from the large datasets. The implementation of knowledge database discovery methods and techniques will help users to extricate meaningful information from virtually accumulated large amount of data. For industries like telecommunication, retail, biomedical, etc., such techniques are used widely. These techniques are proved to be helpful in predicting future trends and allow business activities proactive, dynamic and present valuable and useful knowledge which is simply understandable to human being. This paper provides an outline of the knowledge database discovery process. It presents a detailed study of knowledge discovery with various studies like steps, principle and challenging issues. A primary goal of this paper is to elucidate the relation between knowledge discovery and data mining. It also defines the knowledge database discovery process and important data mining techniques.

References

1. Abdualgali and Abraham (2020) Efficient machine learning algorithms for knowledge discovery in big data. J Int J Adv Sci Technol 29(5):3880–3889
2. Kumar A, Chattterjee I (2016) Knowledge discovery-techniques and application. Int J Comput Sci Inf Technol 7(1):321–322
3. Soleimani F, RajabzadehGhatar A (2019) Knowledge discovery from a more than a decade studies on healthcare big data systems: a scientometrics study. J Big Data 6(8):2–15
4. Saurkar AV (2014) A review paper on various data mining techniques. Int J Adv Res Comput Sci Softw Eng 4(11):437–442
5. Armour F, Kaisler S (2013) Big data: issues and challenges moving forward. In: Proceedings of The IEEE 46th annual hawaii international conference on system sciences, vol 7, pp 995–1004
6. Silwattananusarn T (2012) Data mining and its applications for knowledge management: a literature review from 2007 to 2012. Int J Data Min Knowl Manag Process 2(5):13–24
7. Tomar D, Agarwal S (2013) A survey on data mining approaches for healthcare. Int J Bio-Sci Bio-Technol 5(5):241–266
8. Fan W, Bifet A (2014) Mining big data: current status, and forecast to the future. Artic ACM SIGKDD Explor Newsl 14(2):1–5
9. Bifet (2014) Big data analytics: a text mining-based literature analysis. In: 29th international conference on data engineering

10. Purcell B (2014) The emergence of “big data” technology and analysis. Int J Technol Res 6(10):1–7
11. Indira KK, Reddi D (2014) Different technique to transfer big data: survey. IEEE Trans 5(12):2348–2355
12. Ibrahim (2014) Handling partitioning skew in mapreduce using LEEN. ACM 51:107–113
13. Gamache M (2015) The impact of CEO regulatory focuses on firm acquisitions. J Acad Manag 58(4):1261–1282
14. Baker (2015) Big data in education: new efficiencies for recruitment, learning, and retention of students and donors. J Sci Direct Elsevier 8(5):25–48
15. Soni S (2016) A literature review on data mining and it’s techniques. Int J Comput Sci Inf Secur 14(11):437–442
16. Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. J Strateg Manag J 36(10):1435–1457
17. Angus (2019) Problematic search distance and entrepreneurial performance. J Strateg Manag J 40(5):2011–2023
18. Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities and challenges. J Big Data 6(44):1–16
19. Sankari and Shraddha (2019) A review on research areas in educational data mining and learning analytics. J Int J Sci Technol 8(12):319–323
20. Kumar GR, Basha SR, Rao SB (2020) A Summarization on text mining techniques for information extracting from applications and issues. J Mech Contin Math Sci 40(5):324–332
21. Lauw HW, Wong RC-W (2020) Advance in knowledge discovery and data mining. In: Proceedings, part I: 24th Pacific-Asia conference