International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
Sep-Oct-2017 Volume 4, Issue-5
Email: editor@ijermt.org www.ijermt.org

ADVANCEMENTS IN KNOWLEDGE DISCOVERY IN DATABASES (KDD) FOR ENHANCED WEBSITE MINING: TRENDS AND FUTURE DIRECTIONS

Ganta Ramkrishna Reddy, Research Scholar, Sunrise University, Alwar

Dr. Shalini Goel, Professor, Sunrise University, Alwar

ABSTRACT

The most recent findings in AI, statistics, ML, and database science are presented in Knowledge Discovery in Databases (KDD). Data mining and knowledge discovery are fast expanding fields, and they are a part of it. Here, we went over the basics, as well as classification, clustering, and applications. Various research concerns and stages of data gathering have been emphasized. In order to extract useful information from datasets, analysts use knowledge discovery and data mining (KDD) methods. Data management platforms, e-commerce systems, and important large-scale solutions (like CRM) are just a few examples of the many vertical solutions that use data mining for business intelligence and decision support. The goal of mining is to discover new computational theories and methods that can help people make sense of the ever-increasing amounts of digital data in order to extract valuable information. Practical uses, data mining methods, and difficulties in discovering new information are all covered in this article. Knowledge Discovery in Databases and Knowledge and Data Mining are also covered in this article. One area of computer science that is seeing rapid growth is data mining and knowledge discovery, or DMKD. Its rise to prominence is attributable to the rising need for resources that facilitate the comprehension and analysis of massive data sets. Institutions such as banks, insurance firms, retail outlets, and the Internet constantly create such data. The proliferation of electronic devices like computers, scanners, digital cameras, bar codes, etc., is largely responsible for this boom. Currently, there is easy access to massive amounts of data held in various data repositories, such as databases and warehouses.

KEYWORDS KDD, Machine learning, Data Mining

INTRODUCTION

Data mining, also known as knowledge discovery in databases, is a multidisciplinary field that aims to find valuable insights inside large datasets. Data mining is the process of looking for patterns in large amounts of data that were not there before. With the proliferation of databases and the Internet, there is an urgent need for Knowledge Discovery strategies to deal with the deluge of data available online. Data visualization, optimization, statistics, machine learning, databases, pattern recognition, and high-performance computing are all fields that study the difficulties of data extraction. Data mining plays a significant role in the larger processes of Knowledge Discovery in databases. Data mining is crucial and essential due to the availability and volume of information in the modern day. A number of sources, including Brachmana et al. (1994), Fayyad et al. (1996), Maimon et al. (2000), and Reinartz et al. (2002) [1-3], have suggested several strategies to break down the KDD process into its component parts.

Knowledge Discovery (KD) is the rather complex process of extracting meaningful patterns from large databases. These patterns should be original, useful, and ultimately understandable [30]. One step of KD is data mining (DM). Data mining (DM) is only concerned with knowledge extraction from data, while information discovery (KD) covers a broad spectrum of operations, including data understanding and preparation, verification, and application of the obtained information. But the truth is that you may use DM, KD, or DMKD for anything.

Unfortunately, our data analysis, summarization, and knowledge extraction capabilities have not kept up with the rapid development of ICT. Database technology has given us the fundamentals for efficient data storage and retrieval, but humans still haven't figured out how to make sense of and draw conclusions from massive datasets. As a result, effective data mining and knowledge discovery techniques are required to handle massive amounts of data. Library and information science researchers are starting to pay attention to knowledge discovery in databases (KDD) as these institutions are seen as crucial to knowledge organization. With a focus on digital libraries, this article explains the KDD method and why it's important.

There is a pressing need for efficient ways to glean useful insights from the vast amounts of data produced by websites due to the exponential expansion of the internet. In recent years, website mining—the process of using data mining methods to discover patterns and insights within website data—has become an important field of study. Poor data quality, excessive dimensionality, and a lack of subject understanding are some of the drawbacks that conventional internet mining approaches sometimes face.

Data mining methods are used in the information Discovery in Databases (KDD) process to extract useful information from massive datasets. When it comes to marketing, healthcare, and financial data, KDD has been a lifesaver for discovering patterns and insights. Having said that, KDD's use for mining websites is quite new.The current methods for extracting knowledge from website data may be drastically altered if KDD is combined with internet mining. Website optimization, customization, and design may be informed by insights and patterns extracted from data using KDD approaches. In addition, KDD may be used to find ways to make websites more accessible, easier to use, and generally better for users.

LITERATURE REVIEW

Gilchrist, Mark & Mooers, Deana & Skrubbeltrang, Glenn & Vachon, Francine. (2012). In today's increasingly competitive business world, organizations are using ICT to advance their business strategies and increase their competitive advantage. One technological element that is growing in popularity is knowledge discovery in databases (KDD). In this paper, we propose an analytic framework which is applied to two cases concerning KDD. The first case presents an organization at the analysis stage of a KDD project. The second one shows how a multinational company leverages its databases by mining data to discover new knowledge.

Akanmu, Semiu & Jaja, Shamsudeen. (2012). Knowledge management had been a critical focus and interest in Information Technology, especially as it affects business organizations through the implementation of business intelligence and expertise. Knowledge discovery and knowledge conversion (tacit/implicit to explicit knowledge) play important roles in these aspects; through the application of technologies in the SECI model to aid knowledge management, and identifying the sources of the expertise whether in humans or physical databases serve as the basis for expertise's knowledge management. This paper presents in detail the significances of knowledge discovery in databases (KDD) in achieving an all encompassing knowledge management strategy. This strategy must comprise of transparent and multiple interrelationships of organizational agents through shared mental maps, collaborative and distributed technologies, and solves all problem in other ways with a special focus on data mining which is also found in the KDD process. Extensive literatures were reviewed to operationalize Knowledge discovery in human and in data ware houses as its affect knowledge management, and bring to the fore the processes involved in KDD process, its applications, understanding using SECI model, possible challenges, and suggest the future research areas to solve the observed challenges.

Cao, Longbing & Zhang, Chengqi. (2007). Knowledge discovery and knowledge conversion (tacit/implicit to explicit knowledge) play important roles in these aspects; through the application of technologies in the SECI model to aid knowledge management, and identifying the sources of the expertise whether in humans or physical databases serve as the basis for expertise's knowledge management. This paper presents in detail the significances of knowledge discovery in databases (KDD) in achieving an all encompassing knowledge management strategy. This strategy must comprise of transparent and multiple interrelationships of organizational agents through shared mental maps, collaborative and distributed technologies, and solves all problem in other ways with a special focus on data mining which is also found in the KDD process. Extensive literatures were reviewed to operationalize Knowledge discovery in human and in data ware houses as its affect knowledge management, and bring to the fore the processes involved in KDD process, its applications, understanding using SECI model, possible challenges, and suggest the future research areas to solve the observed challenges.

Dhiman, Anil. (2011). Traditionally, data mining is an autonomous data-driven trial-and-error process. Its typical task is to let data tell a story disclosing hidden information, in which domain intelligence may not be necessary in targeting the demonstration of an algorithm. Often knowledge discovered is not generally interesting to business needs. Comparably, real-world applications rely on knowledge for taking effective actions. In retrospect of the evolution of KDD, this paper briefly introduces domain-driven data mining to complement traditional KDD. Domain intelligence is highlighted towards actionable knowledge discovery, which involves aspects such as domain knowledge, people, environment and evaluation. We illustrate it through mining activity patterns in social security data.

RESEARCH METHODOLOGY

Data mining techniques can be classified, discovering the knowledge and utilizing the techniques. Different features of Knowledge discovery: Accuracy, Automated learning, large amount of data, High Level Language, Interesting Results and Efficiency.

A. Association rules: It Detects sets of attributes that recurrently co-occur and rules. Among them, e.g. 80% of the people who buy biscuits, also buy sugar (70% of all shoppers buy both).

B. Sequence mining (categorical): Find sequences of events that usually occur together e.g. In a DNA set's sequences ACGTC is followed by GTCA after gap of 9, with probability of 30%. CBR or Similarity search: The objects that are within a defined distance of the queried object otherwise it will find all pairs that are within some distance of each other.

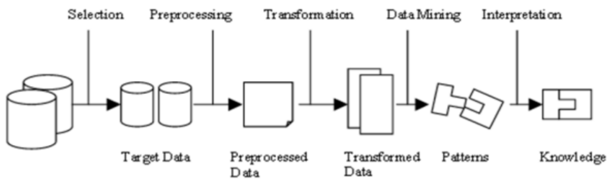


Fig1: KDP

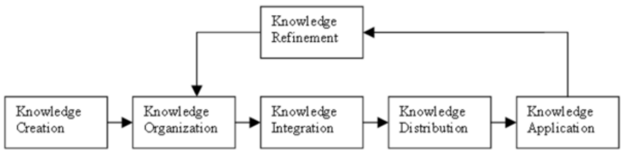


Fig 2: Knowledge discovery in database

C. Deviation detection: Discover the records that are the different from the other records, i.e., finding all outliers.

D. Classification and Regression: Assigning a new data record to one of the several predefined categories or classes. Real-valued fields can be predicted through regression, may called as supervised learning.

E. Clustering: Partitioning the dataset into subsets such as the elements of a subset share a common set of properties, with more within group similarity and less inter-group similarity which may called as unsupervised learning. Many other methods, such as Decision trees, soft computing: rough and fuzzy sets, Hidden Markov models, Time series, neural networks, Genetic algorithms, Bayesian networks.

THE DATA MINING STEP OF THE KDD PROCESS

Data mining process is used to extract information from a data set and transform it into an understandable structure for further use as shown in Fig. 4. The data mining component of the KDD process often involves repeated iterative application of particular data mining methods. It is achieved by using application domain like prior knowledge, user goals etc. to create target dataset that will be used in data mining. The knowledge discovery goals are defined by the intended use of the system. We can distinguish two types of goals: Verification, where the system is limited to verifying the user's hypothesis, and Discovery, where the system autonomously finds new patterns. In this paper we are primarily concerned with discovery-oriented data mining. The goals of prediction and description are achieved via the following primary data mining methods:

Classification: learning a function that maps (classifies) a data item into one of several predefined classes.

Regression: learning a function which maps a data item to a real-valued prediction variable and the discovery of functional relationships between variables.

Clustering: identifying a finite set of categories or clusters to describe the data. Closely related to clustering is the method of probability density estimation which consists of techniques for estimating from data the joint multi-variant probability density function of all of the variables/fields in the database.

Summarization: finding a compact description for a subset of data, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

Dependency Modeling: finding a model which describes significant dependencies between variables (e.g., learning of belief networks).

Change and Deviation Detection: discovering the most significant changes in the data from previously measured or normative values.

DATA MINING AND KNOWLEDGE DISCOVERY

DMKD was brought into attention in 1989 during the IJCAI Workshop on Knowledge Discovery in Databases (KDD) [54]. The workshops were then continued annually until 1994. In 1995, the International Conference on Knowledge Discovery and Data Mining became the most important annual event for DMKD. The framework of DMKD was outlined in two books: „Knowledge Discovery in Databases" and „Advances in Knowledge Discovery and Data Mining" [30]. DMKD conferences like ACM SIGKDD, SPIE, PKDD and SIAM, and journals like Data Mining and Knowledge Discovery Journal (1997), Journal of Knowledge and Information Systems (1999), and IEEE Transactions on Knowledge and Data Engineering (1989) have become an integral part of the DMKD field. In spite of the theoretical advances in DMKD, it is not easy to describe the current status of the field because it changes very quickly. We try here to describe the status of the DMKD field based on the web-based online research service Axiom® [6]. The Axiom service provides access to INSPEC, Compendex®, PageOne™ and the Derwent World Patents Index databases. It can find research papers using a userspecified set of keywords and a time frame. To analyze past developments, current state and future directions of the DMKD field we performed several queries that are summarized in Figures 1 and 2.

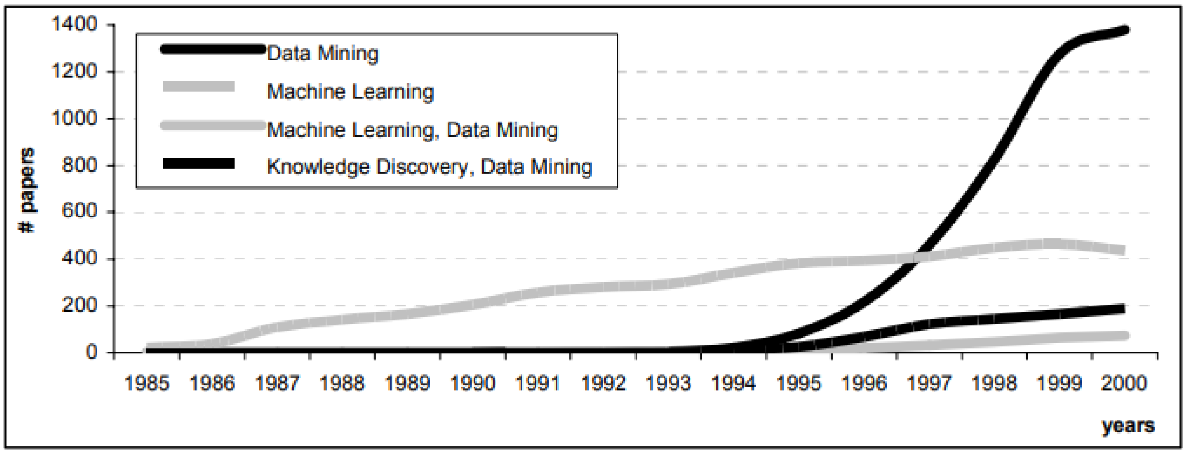
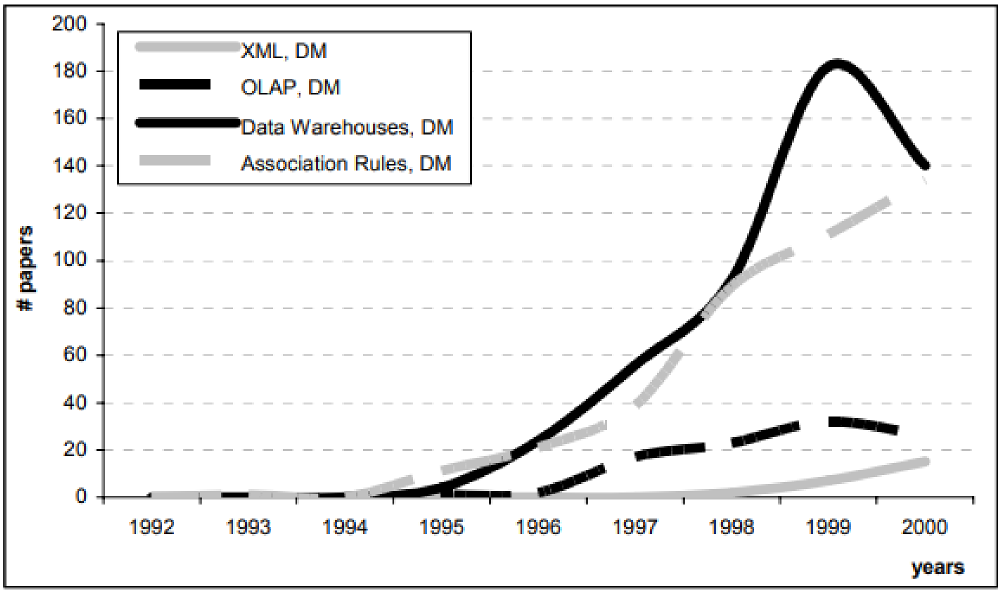


Fig. 1. Evolution of Data Mining and Data Mining and Knowledge Discovery fields

The DM revolution started in the mid 1990's. It was characterized by fast growth, as evidenced by the increase over a 5-year period of the number of DM papers from about 20 to about 1270. One of the reasons for that growth was due to the incorporation of existing tools and algorithms into the DM framework. The majority of the DM tools, e.g. machine learning (ML), were already well established. Figure 1 shows number of ML papers in the context of DM papers. The number of papers covering both ML and DM grows slowly; in 2000 there were 74 such papers, which constituted 6% of the entire DM research. The DMKD field emerged around 1995. In 2000 it constituted 15% of all DM research. This does not necessarily mean that only this percentage of the research is devoted to DMKD since some people still treat DM and DMKD as one and the same.



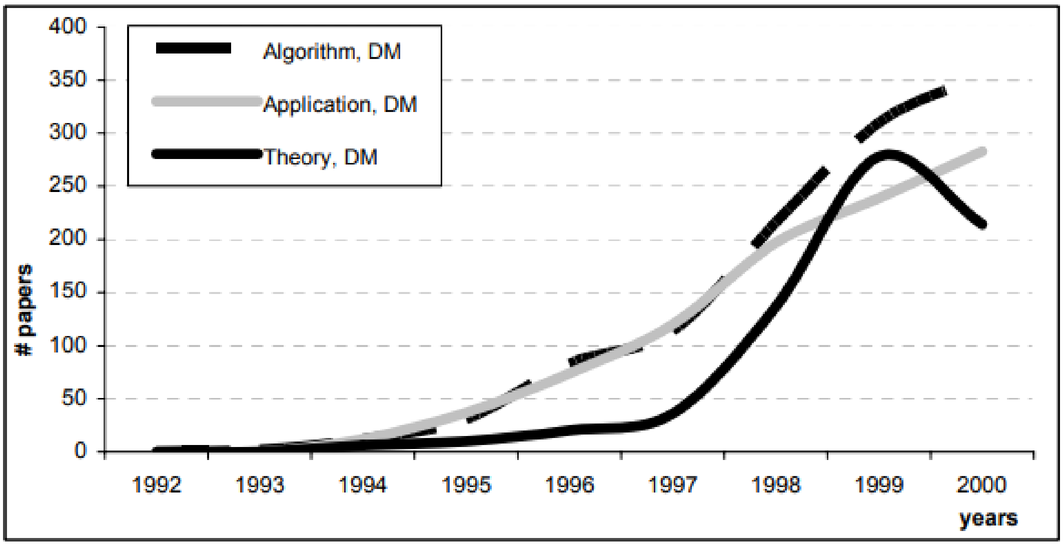


Fig. 2. a) Trends in Data Mining b) Data Mining theory and applications

The trends in DMKD over the last few years include OLAP, data warehousing, association rules, high performance DMKD systems, visualization techniques, and applications of DM. The first three trends are summarized in Figure 2a. The research interest in association rules follows a pattern generally similar to that of the DM field. On the other hand, the research in OLAP (On-Line Analytical Processing) and data warehouses initially was growing, getting maximum attention around 1999. Our observation is that some of the trends that initially had the greatest impact on the DM field began to decline because the majority of the issues concerned with those areas may have been solved, and thus the attention shifted towards new areas and applications. Recently, new trends have emerged that have great potential to benefit the DMKD field, like XML and XML-related technologies, database products that incorporate DM tools, and new developments in the design and implementation of the DMKD process. Among these, XML technology may have the greatest influence on DMKD since it helps to tie DM with other technologies like databases or e-commerce. XML can also help to standardize the I/O procedures of the DM tools, which in turn will help to consolidate the DM market and help to carry out the DMKD process. Figure 2a shows that XML has gained increasing interest, and being a very young concept, the interest in this technology can explode within a very short time.

CONCLUSION

In this last section of the Knowledge discovery, we defined a few key words. We set out to define data mining and knowledge discovery and make that relationship very clear. Here is a rundown of the KDD procedure and some fundamental data mining techniques. Data mining encompasses a wide range of methodologies, each tailored to a certain kind of data and industry. The goal of any KD method may be better understood with an understanding of knowledge discovery and model induction. We believe that this article will help get everyone on the same page about KDD's overarching objectives and methodology. We are optimistic that it will contribute to a deeper comprehension of the many methods within the interdisciplinary realm of KD.

It may be necessary to use a more conventional approach when developing and implementing a new DMKD system [37]. It entails creating and deploying next-gen DM systems that can manage massive volumes of complicated data, illustrate the findings, and mine diverse sources of information including multimedia data [76, 77]. Creating more intuitive interfaces for these technologies should be a priority throughout their development. Because of this, the items will be more widely accepted, especially by smaller and mediumsized businesses, whose employees may not have extensive technical training. The perspective of the user about the uniqueness, understandability, and simplicity of the information provided by the DMKD process is another crucial problem to address. For the next generation of DMKD tools to be more effective, we need to account for human cognitive processes and understand how individuals absorb new information.

REFERENCES

1. Gilchrist, Mark & Mooers, Deana & Skrubbeltrang, Glenn & Vachon, Francine. (2012). Knowledge Discovery in Databases for Competitive Advantage. Journal of Management and Strategy. 3. 2. 10.5430/jms.v3n2p2.

2. Akanmu, Semiu & Jaja, Shamsudeen. (2012). Knowledge Discovery in Database: A knowledge management strategic approach.

3. Cao, Longbing & Zhang, Chengqi. (2007). The Evolution of KDD: towards Domain-Driven Data Mining.. IJPRAI. 21. 677-692. 10.1142/S0218001407005612.

4. Dhiman, Anil. (2011). Knowledge Discovery in Databases and Libraries. DESIDOC Journal of Library & Information Technology. 31. 10.14429/djlit.31.6.1319.

5. Brachman, R., and Anand, T., The Process of Knowledge Discovery in Databases: A human-centered Approach, In Fayyad, U.M., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R., (Eds), Advances in Knowledge Discovery and Data Mining, AAAi/MIT Press, 1996

6. Bradley, P., Fayyad, U., and Reina, C., Scaling Clustering Algorithms to Large Databases, Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, Menlo Park, California, pp. 9-15, 1998

7. Bray, T., Paoli, J., and Maler E., Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, http://www.w3.org/TR/2000/REC-xml-20001006, October 2000

8. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., Classification and Regression Trees, Wadsworth Int. Group, Belmont, California, USA, 1984