International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.37, June 2021 – www.ijais.org

Understanding the Classification of Data Mining and Web Mining

Gehad Abdallah Amran
Northeastern University/ software college /Liaoning, Shenyang, China

Hassan Faisal Aldheleai
Department of Computer Science
Aligarh Muslim University, Aligarh, India

Hussein Al-Sanabani
Sakarya University, Department of Computer Engineering, Sakarya, Turkey

ABSTRACT

The amount of data stored in databases is rapidly increasing. This creates the need for new technologies and tools to auto handle and enables humans to manage and analyze large data sets in a smart way to gather useful information. This growing need is generating a new field of research called Knowledge Discovery in Database (KDD) or Data Mining, which has cached researchers' interest in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization. Web Mining is part of data mining technology, which aims to extract interesting and useful hidden patterns and information from web documents and web activities.

Keywords

Data mining, web mining, KDD, Knowledge Discovery in Databases.

1. INTRODUCTION

The amount of data stored in databases is rapidly increasing. This creates the need for the new technologies and tools to auto handle and helps humans to manage and analyze large data sets in an intelligent way to gather useful information. This growing need is generating a new field of research called Knowledge Discovery in Database (KDD) or Data Mining, which has cached researchers' interest in many different fields including database design, statistics, pattern recognition, machine learning, and data visualization.

With the explosive growth of the Internet, most of the current search engines [1] such as Google, Yahoo and MSN offer users a single, linear list arranged by pages along with their partial content arranged by relevance to the search query. The query list form is supported by most of the search engines. The users surfing the Internet are forced to browse the long list and check addresses in sequence to determine desired results. Search engines are expected to not return only the most popular documents matching the query, rather than that, it is predicted to give comprehensive information related to the query. The web search result displayed in an ordered list limits the effectiveness of the search within the top few documents. To tackle this problem the clustering of search results into different sets of documents has been determined [2]. Once the outcome of the query is managed in clustering sets, users only need to pick the related cluster and surf for the required document.

Web Mining is part of data mining technology, which aims to extract interesting and useful hidden patterns and information from web documents and web activities. Web mining focuses on data such as web page content, user access information, hyperlinks between pages, and a variety of World Wide Web (WWW) resources to achieve intrinsic properties among data objects through machine learning, inductive learning, and statistical analysis to search for interesting and possibly useful patterns and implicit information, using the data mining method extracting a higher level of knowledge and lows of interest to users in the network using information filtering technology.

Generally speaking, web mining can be divided into three categories: web content mining, web architecture mining, and web mining. Web content mining is the process of extracting useful information from web pages or documents. Web architecture mining refers to the process of deriving knowledge from the relationship between the organizational structures of the web. Web mining is the extraction of data during a user's interaction with the web [3].

2. KNOWLEDGE DISCOVERY IN DATABASES

KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable (Comprehensible) patterns in data [4].

Valid: are the discovered patterns representatives of the data.

Novel: are the discovered patterns new to the organization.

Useful: can the organization use the discovered patterns.

Comprehensible: can the discovered patterns easily be understandable. Here, data is a set of facts (for example, cases in a database), and a pattern is an expression in some languages describing a subset of data or a model applicable to a subset. Hence, pattern extraction also specifies the fit of a model to the data; Find a structure from the data; Or, in general, any high-level description of a set of data. The term process denotes that KDD includes many steps, which include data preparation, pattern search, knowledge assessment, and refinement, all of which are repeated in multiple iterations. In a non-intuitive way, the process is not a direct calculation of predetermined quantities like computing the average value of a group of numbers [4].

3. DATAMINING

Data mining is an analysis (often large) of observational data sets to find unexpected relationships and to summarize the data in new ways that are understandable and useful to the data owner [5].

The data mining tasks depend on the type of knowledge the KDD system is looking for. Each data mining task has its specificities and follows specific steps in the discovery process. The following data mining tasks are among the most used at present in KDD / data mining applications [4] [6].

*Classification:*The task is to discover if an item from the database belongs to one of the predefined categories.

*Cluster Identification:*The goal of this task is to create descriptions of categories of data (groups), as opposed to categorization, where classes are known in advance.

*Change and Deviation Detection:*A form of knowledge that can be detected in the data set is related to certain deviations and changes in the data values concerning some expected values.

*Mining Association Rules:*This is the task of extracting the frequently used data. Finds links between two or more fields in a data record, or between sets of values in a single field. Correlations are represented as rules, such as `` From all data records containing A, B, and C, 72% also contains D and E. " This is important if the task is, for example, to discover which items consumers usually use to buy together In retail shopping ("shopping styles").

4. WEB MINING

Web mining is a type of application used to obtain data and extract knowledge from web data so that at least one or more structure or usage data (weblog) is used in the mining process (with or without other web types)Figure 1 illustrates the web mining process.

Since web mining is derived from data mining, its definition is similar to the well-known definition of data mining. However, web mining has many unique characteristics compared to data mining first, the source of web mining is web documents. Using the web as middleware in extracting databases and extracting logs, user profiles on a web server still belong to the traditional data mining category.

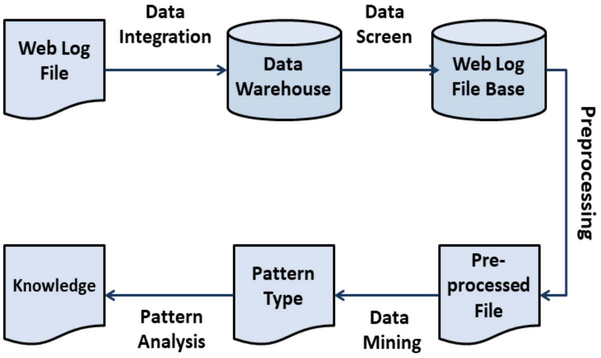

Fig 1: Web mining Process

Second, the web is a vector graph made up of document nodes and hyperlinks. Therefore, the specific style can be about the content of the documents or the structure of the web.Moreover, web documents are semi-structured or disorganized with many semantics that can be read by the machine while the source of data mining is limited to data organized in the database. As a result, some traditional data mining methods do not apply to web mining. Even if possible, it should rely on pre-processing of documents [7].

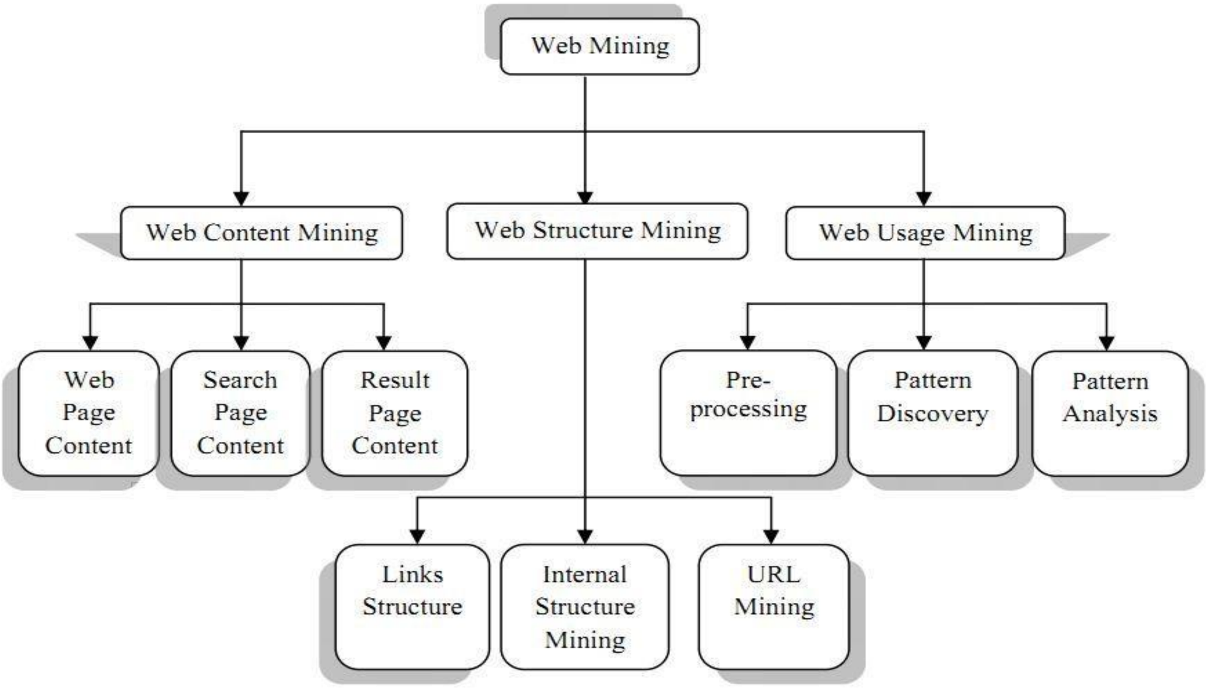

Fig 2: Web mining Taxonomy

4.1 Web Mining Categories

Generally speaking, web mining tasks can be categorized into three types: web content mining, web architecture mining, and web usage mining. All three categories focus on the process of knowledge discovery of tacit, prior unknown and potentially useful information from the web. All of them focus on different mining objects on the web. Figure 2 illustrates web mining types and their objects [8].

4.1.1 Text Mining

Looking at a specific type of categorizing task, where the objects are text documents such as articles in newspapers or scientific papers in magazines or likely abstracts of papers or might be its titles. The goal is to use a group of pre-categorized documents to categorize those that haven't yet been seen.

4.1.2 Text Categorization

Given a pre-defined classification, each document in Group C is classified into one or more suitable categories. In this way, it is not only easy for users to browse documents but also easy to search for documents by selecting the category. presently there are several text classification algorithms, the most used ones are the K-Nearest Neighbor algorithm, Naive Bayes algorithm, etc. [9].

4.1.3 Text Clustering

The difference between clustering text and categorization of text is that there is no predefined text clustering classification. Generally,, the goal of text clustering is to divide the C-document set into a group of clusters so that the similarity between the set is minimized and the similarity within the cluster is maximized. Text Clustering can be used to organize the retrieval of the outcome restored by the search engine. Then users only need to check the clusters returned by their queries, which greatly minimize the time and effort involved in screening a huge number of listed documents. Some text clustering algorithms have been introduced, all of them divided into two types. One is hierarchical clusters [10] and the other is the grouping of sections represented by the K means algorithm [11].

4.2 Web Structure Mining

Link structures allow web pages to show more information than regular documents. Link numbers to page indicate page high circulation among users, and links to page indicate page topics. A frequently cited page might be a significant or a popular page. perhaps mining the web is possible using the link structures. Web structure mining faces a challenge to manage hyperlink structure within the web itself. As interest in web, mining has grown more research has increased in structure analysis, which resulted in a new field of research referred to as Link Mining, which is located at the cross-point of work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. Link mining has produced some excitement in some traditional data mining tasks. These potential link mining tasks applicable in web structure mining are as follows [12]:

4.2.1 Link-Based Classification

It's the idea of concentrating on predicting the type of the web page, depending on the words occurring on the web page, links among web pages, anchor text, HTML tags and other possible features found on the web pages.

4.2.2 Link-Based Cluster Analysis

Cluster analysis is a way of finding naturally occurring subcategories. where data is manipulated and divided into different groups, while the similar objects are assembled together in one group, the dissimilar objects are gathered in other groups.

4.2.3 Link Type

There is a broad range of tasks involved in predicting the presence of links, such as predicting the type of link between two entities or predicting the purpose of the link.

4.2.4 Link Strength

The links can be related to weights.

4.2.5 Link Cardinality

The Link Cardinality primary task is the prediction of the number of links among objects.

4.3 Web Usage Mining

Web Usage Mining (WUM) is a kind of web mining tasks that includes the auto-detection of user entry patterns from one or more web servers. WUM concentrate on technologies that can enable the prediction of the behaviour of browsing users as they interact with the web pages on the Internet. Web Usage Mining gathers the data out of the weblog records to extract the patterns of user access to the web pages. there are allots of accessible commercial projects and scientific research which analyze these patterns for different purposes. The applications built from this analysis can be categorized as personalization, system optimization, location modification, business intelligence, and usage profiling. WUM has the following phases [13].

4.3.1 Data Pre-Processing for Mining

An important and necessary step is to get the data converted into raw data for more manipulation.

4.3.2 Pattern Discovery

It is the main component of web mining. Pattern Discovery includes techniques and algorithms from many areas of research, such as data mining, machine learning, statistics, and pattern recognition. In this stage, techniques such as classification, grouping, correlation rules, and statistical analysis are used.

5. CONCLUSION

Both type of applications of data mining and web mining, technologies technique for getting out hidden and targeted data from the data's storages.

Data Mining includes the processes like Design Disclosure, Algorithm Fathoming and Data Extraction. Data Mining Applications are more related to areas like the retail industry, Financial Data Analysis, Data Analysis and Telecommunication Industry. moreover,Data Mining is usually carried out by expertise like Data Engineers.

Web mining is the application of information mining methods to extricate information from web data, counting web, reports, hyperlinks connecting archives from the data repository and extracting it from the web is referred to as web mining.

Web Mining" is carried out by the experts like various Data Analysts. Web Mining may include the processes like Algorithm Understanding, Information Extraction and Design Revelation, anyway, all these methods carried out with the aid of the Web which too on different Web Servers and Web Documents independently.

6. REFERENCES

[1] Kummamuru K., Lotlikar R., Roy S., Singal K. and Krishnapuram R., "A hierarchical monothetic document clustering algorithm for summarization and browsing search results", Proceedings of the 13th international conference on World Wide Web, ACM Press, pp. 658-665, 2004.

[2] Xuanhui W. and Cheng X., "Learn from Web Search Logs to Organize Search Results", SIGIR, July 23- 27,

Amsterdam, pp.87.94, 2007.

[3] Li Mei and Feng Cheng, "Overview of WEB Mining Technology and Its Application in E-commerce", IEEE ,2010.

[4] Fayyad, U., Piaetsky-Shapiro, G., Smyth, P. " From Data Mining to Knowledge Discovery: An Overview " In Advances In Knowledge Discovery and Data Mining , AAAI/MIT press, Cambridge mass, 1996.

[5] David Hand, Heikki M., and Padhraic S., " Principles of Data Mining ", MIT Press, 2001.

[6] Simoudis E., " Reality Check for Data Mining ", IEEE Expert , Vol.11, pp. 26- 33, 1996.

[7] Fayyad U., et al, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", Communications of the ACM, Vol. 39, No. 11, Nov, pp. 27-34, 1996.

[8] Jaideep S., Robert C., Mukund D., Pag-Ning T., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations Newsletter, 2000.

[9] Yang Y., "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval", In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994.

[10] Willet P., "Recent Trends in Hierarchical Document Clustering: a Critical Review", Information Processing and Management, Vol. 24, pp. 577-597, 1988.

[11] Rocchio J., "Document Retrieval Systems – Optimization and Evaluation", Ph.D. Thesis, Harvard University, 1966.

[12] Getoor L., "Link Mining: A New Data Mining Challenge", SIGKDD Explorations, vol. 4, issue 2, 2003.