Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 97 (2016) 42377-42379

ISSN: 2229-712X

Categorizing Web Pages and Data Mining

Seyed Mostafa Javadi Moaghaddam and Zohreh vahedi*

Islamic Azad University of Ferdows, Department of Computer Engineering, Ferdows, South Khorasan, Iran.

ARTICLE INFO

Article history:

Received: 10 September 2015;

Received in revised form:

21 August 2016;

Accepted: 31 August 2016;

Keywords

Data mining,
Web Mining,
Classification,
E-Commerce.

ABSTRACT

The diversity of knowledge on the web has made determining communication patterns in the database and knowledge discovery among this mass of information an attractive target. The first step to achieve this goal is to classify web pages. The current machine learning techniques to classify content are initially discussed by flat and simple text documents and do not use structures such as links and headers discussed in the web optimally. Data mining is a set of techniques that allows a person to move beyond conventional data processing and help mining the information hidden in a massive volume of data. Today, in most organizations, data are collected and stored rapidly. However, using these data is not simple and they cannot be used as a single unit of the volume of data, so techniques they can be applied properly using a combination of statistics and computer science and the use of machine learning. However, in order to achieve a meaningful result of Web mining we need to have good data on our website, so effective management of Web data is crucial in web mining. The ultimate goal of this descriptive paper is that the organizations and small businesses could use data mining in their decisions and e-commerce as a future big business would enter into new future areas. There is a wide range of users considered for this technology, including for example research centers, companies operating in the field of web and database, analysts and managers of organizations, business, Web (search engines) and …

Introduction

Today, we are faced with a great deal of data. In order to use them we need knowledge discovery tools. Data Mining is used as an advanced ability to complete data and discover the required knowledge. Data mining is applied in science (astronomy), business (advertising, customer relationship management, etc.), Web (search engines) and applications. The term data mining is like coal and gold mining. Data mining extracts the data buried in warehouses as well. In fact the goal of data mining models is to create models for decision making. These models predict future behaviors based on an analysis of past behaviors (C.Clifton, 2001).

There are lots of benefits in Web mining, including database search on the web or explore patterns being used to provide useful information to users. Data and web mining was developed as independent technology in the mid-1990s. Researchers started thinking seriously about Web mining not long ago. Web mining workshops was one of the first experiences during the knowledge database discovery conference in 1999.

Stivastava and Cooley determined a classification for Web mining in which they initially divided Web mining into two parts:

- Obtaining patterns from the web data
- Obtaining web log

Then their classification was developed into three categories including Web content mining, Web usage mining and Web structure mining (David, 2001).

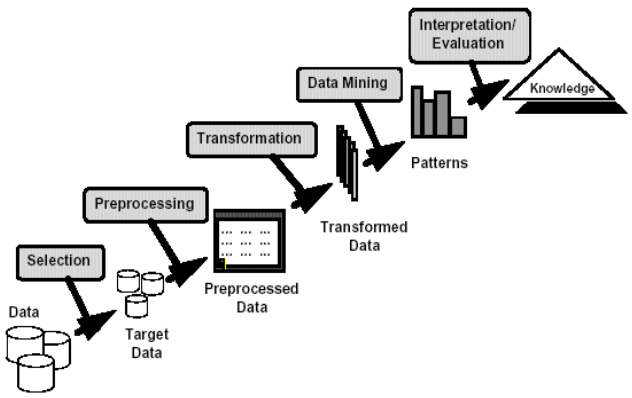
Web content mining includes text, video and… on the web which is a subset of data mining.

Web usage mining includes information exploration on access to web pages, and includes click stream analysis.

Web structure mining is about URL search and other web links to obtain their structure.

Defining Data Mining

New information and communication technologies, as well as decision support technology can affect timely, accurate and required data search of people through collection, storage, analysis, interpretation and analysis, retrieval and dissemination of information and knowledge to specific users. One of the tools used in these technologies is data mining. Data mining involves the use of advanced tools for data analysis in order to discover valid patterns, previously unknown relationships in large data sets. These tools are statistical models, mathematical algorithms and machine learning methods (algorithms that automatically improve their performance through experience). Data mining is beyond data collection and management and includes analysis and prediction. Another name for it is Knowledge Discover in Database (KDD) (J.Han, and M.Kamber, 2001).



Data Mining can be performed in quantitative, text or multimedia data. Its applications include:

- Acossiation Rules: Patterns in which the existence of one item indicates the existence of other items,

Tele:
E-mail address: mahdidarbandi@hotmail.com

- Classification: attribution of the patterns to a small set of predefined classes through discovering some relationships between characteristics,
- Clustering: grouping customers or a set of patterns with similar characteristics,
- Prediction: discovering patterns for logical predictions about future,
- Path Analysis or sequential patterns: the patterns in which an event leads to another event.

Data mining is not a new technology but its application is generally growing significantly in various and public sectors and generally industries such as bank, insurance companies, and medical and retail stores use data mining to reduce costs, increase research and sales.

Data mining operations

In data mining, four main operations are performed including:

1. Predictive modeling
2. Database fractionation
3. Link analysis
4. Deviation detection

Data explore on the web

Data explore on the web is a main challenge in data management, web data management and machine learning.

There are a lot of data on the web that makes useful and appropriate data extraction a real challenge for the users. When users are browsing the Web, the Web can be quite quiet and users can obtain needed data quickly. The question is that how these data are transformed into information? Are the data acquired by the users the same as what they need? Which method can be used to extract the previously unknown data on the web? One simple solution is to complete data mining tools with the data itself. This is a good solution if the data is in a relational database. Therefore, one of the requirements of extracting data within a relational database is data mining tools. These tools need to develop a Web interface. For example, if a relational interface is prepared, SQL-based search tools can be connected to the relational database.

Web data mining on the Web relational database

Unfortunately, the Web is not very sincere. Most data are unstructured and semi-artificial. There are a lot of video and image data that can be complicated in case of the existence of a single connection interface for all these databases. The question is that how data are stored? We mainly focus on the extracted text, pictures, audio and video data. One of the needs of development tools is multimedia data mining and then focusing on data extraction tools such as the web data where the multimedia database are combined and then explored (Usama Fayyad, 1996).

Data mining and Visualization on the Internet

Recently, various standards have been developed by organizations such as ISO, W3C and OMG to have access to internet data. These standards include models, specific languages and architectures. One of them is XML (Extensible markup language) for the writing of Document that allows Document to be translated by the people who get it. The relationship between data mining and standards is indiscoverable such as XML. However, various technologies must work together to make data mining effective. This includes exploring on multimedia data, extraction tools to predict the tendencies and activities on the Web, as well as technologies for data management, data storage and visualization of them.

Usage Mining patterns

Other aspects of exploring the web include collecting various statistics based on traditional patterns to determine which web page is likely to be achieved. Search in this part is driven by various groups.

Web mining

Web-mining is one of the most important applications of data mining to search the world wide network to discover and extract useful patterns. Web in addition to having a vast collection of information, includes dynamic collection of links to access web pages and use of information that creates a rich set for data mining.

Web Structure Mining

Web Structure mining is about extraction of the links on the web and it is closely related to Web usage mining. Link extraction is needed to determine where the user is and the page the user can access. Web Structure mining is used in search engines such as Google, for example the links are extracted and then one of them can define the web page mentioned before. When you search a complete stream first the web pages with highest links related to that stream are listed.

Why Data Mining Web?

There are many problems with the new tools and in order to solve these problems new tools and techniques should be applied. For example data analysis is performed through various methods but today we just talk about data mining and the development and improvement of its techniques are discussed.

The data around us are on paper and in the mind of people in many cases. The clerks usually spend several years to record data and humans spend several years to identify and analyze various patterns. They finally decided to use a new way to analyze their data. However the organization of data is still been a big problem. Since the advent of computer and databases, data storage started in the computer databases. This was the first major step towards data mining. After the appearance of artificial intelligence search techniques improved. The thing that had the major role in data mining to improve the storage and retrieval of data was the database management systems.

Outputs, methods and techniques of web data mining:

What can be our expected outputs? What methods and techniques are used? What are data mining techniques?

Many books such as: [MITC 97], [BERR 97] and [ADRI 96] have discussed this issue. Linoff and Berry have provided an excellent discussion on outputs, methods and data mining techniques.

Data mining outputs refer to data mining task and type. The results are among the conclusions expected to be obtained by data mining. Data mining output includes classification, clustering, forecasting, research and related groups. It should be noted that the terms are not standardized.

Data mining techniques use algorithms to use data mining. There are differences between the techniques and outputs. For example, a set of data mining techniques are used to analyze the portfolio but a portfolio analysis is an application itself and this is everything required to determine the factors which can be bought in a supermarket. Therefore, we expect a standard to be developed for the vocabulary to remove these differences.

Assessment and prediction are the two other tasks in data mining. In a type of estimates based on patterns sent by the person and the person's age, we can estimate how much he gains or how many children he has. Under the task of predicting we can predict the future behavior of some values. For example, based on a person's level of education, his current job and tendency in the industry one prediction could be how much he gains in 2005. Another example is that depending on the observed pattern in the newspaper articles one can predict future event definitely.

Classification is one of the tasks of data mining which is

confused with categorization. While the categorization of the entities is based on the predefined values of properties, classification similar to records is not based on predetermined values. When we categorize some people, in fact we have predefined classes that are based on the values of some properties. In case of classification, we do not have these predefined classes and classify the data by data analysis instead.

Data mining Vs. Data mining web

The information discussed in this part is used for data mining, web data mining and multimedia data mining. So the question is that what is the most interesting point in web data mining? Since there are many data on the Web including multimedia and -structured data, web data mining is more general than data mining and its related factors. As far as the differences are concerned the most important challenge of data mining is to get the intended data for search. For example finding, gathering and managing data in the web might to be easy, so we need to find data source and organize it. We need a technology for wed data management. Also we need to use the information management techniques for the data existing on the web. In addition we should perform conceptual data mining as well. We need to consider that web data mining is not restricted to content data mining but it also includes structure and usage data mining as well.

Processes and techniques are interdependent, we consider all data mentioned in the previous section applicable to web data mining. We need to get the intended data, search it, and generate help projects so that we could perform a large-scale data mining. Also the techniques such as neural networks, decision and analysis tree are used as well.

Conclusion

Today, with the development of databases and the huge volume of data stored in these systems, new tools are required to process the stored data and transform them into useful information base on which vital decisions would be made in the organizations to achieve greater profits.

Therefore, data mining has access to some tools that discover useful information or useful patterns (logical relationships between data) among the high volume of data semi-automatically, with minimal user intervention.

Among the most important user of data mining include retail stores, insurance companies. In data mining process models and algorithms such as neural networks, selection trees, inference rule and genetic algorithms are used that using techniques such as predictive modeling, database fractionation, link analysis and deviation detection we could discover useful patterns with minimal user intervention.

As a result, the main purpose of data mining is to discover the knowledge hidden in data that exist in huge data banks and in order to access these huge data banks first an integrated environment of data called data mining database is required. Then the required data are searched and some conversions are performed on them and finally of knowledge discovery known as data mining the intended patterns are found by the tools used in data mining and the last stage of knowledge discovery the quite understandable result is presented to the user.

References

1. Wikipedia the free encyclopedia
2. Monthly Tadbit practical training devise No. 156
3. Mehrizi, Haeri, AA, "Data Mining: Concepts, Methods and Applications" (2003) Economic and Social Statistics MA thesis, Faculty of Economics, University of Allameh Tabatabai.
4. Zafaryan, Reza and Zafaryan, Qasim, "a review of data mining" (2001) Journal of industries, No. 29
5. Shah Samand, P., "Data Mining in Customer Relationship Management" (2005), No. 156 Tadbir journal.
6. Goudarzi, HR, Translator "What is data mining?" selective publication of statistical material, the Statistical Centre of Iran (52).
7. Jamali, Arman - Electronic City, The context of arrival to Cybernetics Age competition (http/www.editorial.com).