Novel Pattern Classification Techniques for Web Mining

Surbhi Sharma, Anjaly Chauhan, Pankaj Sharma

Abstract: The web mining is one of the application of data mining which uses data mining techniques such as classification, clustering, association rules in order to effectively mine and extract meaningful patterns from web data. In this paper concept of web mining is discussed, process, classification, issues and applications.

I. INTRODUCTION

WWW is highly dynamic in nature as lots of updation and deletion occurs everyday. The current WWW has gained the peak of success in respect of
1. Large number of users
2. Efficient digital commerce business
3. Multidimensionality and variety of data
4. Relevant resource of information

Data mining can be called as Web mining when it is applied on web data. Web data can be web pages, web servers, web links, web documents etc. The term Web mining is defined as the process of finding the hidden patterns from the web, this pattern may be used for e-tailers to leverage their online customers data by understanding and predicting the behavior of their customers in order to cater their needs and provide best service to them. Generally we can mine three Kinds of knowledge from web dataviz. Web usage mining, Web content mining and Web structure mining

Web mining uses various techniques such as classification, clustering, association rules to discover these kinds of knowledge. Thus, web mining simply moves the surplus data environment to relevant information where we get valuable information which is beneficial in this current scenario.

II. NEED OF WEB MINING

Web mining is the application of data mining techniques to extract the useful patterns from the web. As we know, the user (who demands valuable information) business people (who provides service to the customer) both are associated with web data and face some problems while dealing with web data like Problem faced by user :

a) Analyzing unstructured data is very tedious job.
b) Extracting relevant information- Users use searching tools to find specific information but today’s search tool is not efficient to index all relevant pages so it leads to less accuracy of search result.

Discoverd knowledge

Collection of Data: The first step of web mining is finding the source of the web data i.e server logs, client side browsers, proxy server in order to gather data for pattern discovery and interpretation.

Selection and Pre-processing: Selection step is responsible for extracting task relevant data from collected web sources. Pre-processing is one of the important step in web mining process which is directly related to quality of discovered patterns. It is required to preprocess the web data as it is highly noisy which leads to poor quality information. It involves various sub tasks:

1. Data cleaning: It refers to cleaning of data by removing irrelevant references and fields, useless files such as .jpg, .mp3 etc. It also cleans log files with unsuccessful, HTTP states code under 200.
2. Data integration: It combines the data from multiple server into a coherent form by resolving different representation of web data.
3. Data reduction: It obtains reduced representation of task relevant data which is smaller in volume but yet produces some analytical results.
4. User identification: It is responsible to identify each distinct user by IP address, cookies etc.

I. PROBLEMS FACED BY INFORMATION PROVIDER:

It may be possible that provider is unable to find valuable information regarding need of customer, increasing efficiency of web data in order to provide better services etc.

III. WEB MINING PROCESS

Web mining process comprises the following task to discover the interesting pattern from the web[5].
1. Collection of data
2. Selection and preprocessing
3. Pattern Discovery
4. Pattern analysis

Revised Manuscript Received on May 05, 2019.
Surbhi Sharma, Assistant Professor ABES Engineering College Ghaziabad
Anjaly Chauhan, Assistant Professor ABES Engineering College Ghaziabad
Pankaj Sharma, Professor ABES Engineering College Ghaziabad

Revised Manuscript Received on May 05, 2019.
Surbhi Sharma, Assistant Professor ABES Engineering College Ghaziabad
Anjaly Chauhan, Assistant Professor ABES Engineering College Ghaziabad
Pankaj Sharma, Professor ABES Engineering College Ghaziabad
5. Sessionization: It involves grouping of different activities of a single user.
6. Path completion: The process of reconstructing the user’s navigation path by appending missed page request due to proxy server, browsers back button is pressed and local caching.

**Pattern Discovery**: In this task various machine learning techniques are used to discover the interesting pattern from web data[11].

**Pattern Analysis**: It examines the discovered pattern and interpret it and check whether the pattern is relevant or not using validation methods[11].

### IV. CLASSIFICATION OF WEB MINING

1. Web usage mining
2. Web content mining
3. Web structure mining

1. **Web usage mining**: It is a process of retrieving useful knowledge from web server logs, user queries, database logs, client side cookies and user profiles to analyze the web user’s behavior[5] like finding out what users are looking on internet. It is called as web usage analysis or web log mining or click stream analysis.

   The source of data for web usage mining can be:
   1. Server data: It captures all the user logs like time and date of request for particular web page. Client Data: It includes browser’s history
   2. Proxy data: It lies in between client and server level which records the data of group of users accessing huge group of web server.

2. **Web content mining**: It is defined as the discovery of interesting patterns from web documents which are either structured(data in tables or database generated HTML pages or unstructured(text data)/semi structured(HTML tags) but mostly unstructured.

   The web content data consist of web documents which comprises text, video/audio, image, meta data, hyperlinks etc. it simply examines the result of web searching as well as the content of web pages.

3. **Web structure mining**: It is the link structure of the hyperlinks at inter level and intra level. It uses the graph theory to analyze the node and connection structure of the hyperlinks at inter level and intra level. Intra level analysis focus on the links within the page itself while inter page level analyses on the links between the web pages[1].

### V. ISSUES IN WEB MINING

WWW is considered as enormous amount of source of high dimensional and dynamic data which continuously growing. Undoubtedly it fulfills the needs of user in terms of business, communication and so on.

But due to its high dimensionality, dynamic behavior limited query interface[5].

It is very difficult to extract useful information from web which may create various issues such as-
1. The size of web data sets (Multitera bytes)-This abundant unstructured web data leads to challenging situation for both users and information provider.
2. Difficult to mine on single server so it needs large number of server.
3. Need proper organization of hardware and software to extract information from such data sets.
4. Complexity in extracting relevant information
5. Difficult to mine time series and sequence data.

### VI. APPLICATION OF WEB MINING

For the last few years web mining gained high popularity in various application fields such as e-commerce, e-business, Business intelligence system[8].

---

**Techniques of web content Mining**

Some of the important techniques are:

1. **Unstructured Text Mining**: It is related to text mining because most of the contents are unstructured(text). It is also called as KDT(Knowledge discovery in text). Under this technique the data is searched and returned but it is not necessary that the retrieved data is relevant. So we have to use some tools to get relevant data from it[2].
2. **Page Content Mining**: It is structured data extraction technique which classify the pages based on page rank[2].
3. **Semi structured web content**: It extracts the information from semi-structured data which are represented in the rigid structured(Eg. HTML)[2].
4. **Multimedia Data mining**: It focus on pattern discovery, rule evaluation and knowledge acquisition from multimedia data such as image, text, audio/video which are very difficult to access by query. Multimedia mining has become boon to the research field as large amount of data has been accumulated due to highly dependency on internet but extracting valuable information according to the need of user is such a hard task so Multimedia mining rescue this problem[2].

---

Published By:
Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: G1029587S19/19©BEIESP 144
Web mining plays a vital role in research area of e-service and business intelligently which leads to provide better services for customers. In this section we discuss some of the applications of Web Mining as follow

1. **E-Business** - Web mining supports e-business by improving customer support, marketing strategies and sales operations.
2. **Security and Crime Investigation** - Web mining techniques such as classification and clustering are also used in field of cyber crimes such as internet fraud, cyber terrorism etc in order to protect the user against such cybercrimes.
3. **E-learning** - To improve e-learning environments web usage mining techniques comes into play.
4. **E-Commerce** - In order to success in today’s highly competitive global environment business users demand business answers like which product is not doing well in market, such answers can be given by web mining techniques which helps the organizations to take strategic decision and make them successful in present and coming future.
5. **Web Tracking** - It is responsible for tracking the individual’s behavior across all sites user visits. Web structure mining simply records the users access pattern and customer buying pattern or his interest which thereby gives a clear picture of his interest to the markets.

**VII. CONCLUSION**

In this paper we addressed a detailed survey on web mining concepts. In particular we focused on web mining process and classification. In the later section this paper described applications of web mining. As we know web mining becomes an emerging field of research and extracting a valuable information is still a challenging task so a lot of work can be done under this field.

**REFERENCES**