An Overview of Web Content Mining Techniques

K. R. Srinath

Associate Professor, Department of Computer Science, Pragati Mahavidyalaya Degree and PG college, Hanuman Tekdi, Koti, Hyderabad, Telangana, India

---***---

Abstract – *World Wide Web over the years has become an huge repository of data from which extracting useful information has become one of the biggest challenges. Furthermore the data is available in the form of structured, semi-structured, and unstructured formats. Much of the information on web is either semi-structured or unstructured. Extracting potential useful knowledge from these various formats has become a big area of research in the recent times. Web Content mining is a subset of Web mining which focuses on extracting useful patterns from the contents available in the web documents. This paper focuses on the various content mining techniques to be applied on the web documents.*

Key Words: Web Content mining, usage mining, structure mining, structured data, semi-structured data, clustering.

1. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data including web documents, hyperlinks between the documents, usage logs of websites etc., The world wide web is a collection of documents containing text, images, audio and video data. Existing of such diverse data poses problems to the users interacting with these documents to find the appropriate data which they are interested in. Web mining process is categorized into three categories: Web Usage mining, Web Content mining and Web Structure mining. Web content mining mines the content like text, images, audio, video, metadata, XML, HTML, hyperlinks and extracts useful information.

1.1 Overview of Web Mining

Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining [1]. The research in the field of web is classified on two aspects: the retrieval of information and the mining the information. Information extraction focuses on extracting relevant facts whereas information retrieval focus selects relevant document.

2. WEB MINING CATEGORIES

This section divides web mining into three categories depending on the type of data i.e. Web Content Mining, Web Structure Mining and Web Usage Mining.

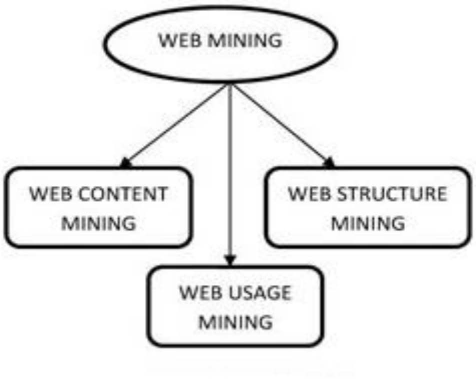

Fig-1: Web Mining Categories

2.1 Web Content Mining :

Web content mining is the mining, scanning and extraction of text, videos, graphs and pictures from web documents. It is also known as text mining. Two types of approaches are used in web content mining. The two approaches are: the *database approach* and the *agent based approach*. The database approach helps in retrieving the semi-structured data from web documents. The agent based approach searches relevant information and helps in organizing the collected information [2]. Web content mining analyzes the content of web resources. Content data correspond to collection of facts a web page was designed to convey to the users. Most of the data available on the web is unstructured data. Two different points of view of web content mining are: the *information retrieval view* and the *database view*. The main goal of content mining from information retrieval view is to improve the filtering and finding of the information to the users. The main goal of database view is to manage the web data.

2.2 Web Structure Mining :

The goal of web structure mining is to generate structural summary about web pages and web sites. It shows the relationship between the user and the web. It discovers the link structure of hyperlinks at the inter document level [3,4]. It also helps in discovering the structure of document which is used in revealing the structure the structure of web pages and it's possible to compare the web page schemes. This is further divided into two types that is based on the kind of structural information used.

a) Hyperlinks: Hyperlinks help in connecting web pages to different location either in same web page or on different web page. A hyperlink is divided into two categories i.e. intra-document hyperlink and inter-document hyperlink. Intra-document hyperlink connects different part of the same page whereas inter-document hyperlink connects two different pages.

b) Document Structure: The content within the web page can be organized in tree structure that is based on various HTML and XML tags.

2.3 Web Usage Mining:

Web usage mining is the process of finding out what users are looking for on the internet. It tries to discover useful information secondary data derived from the interaction of users while surfing web. There are three phases of web usage mining. The three phases are :

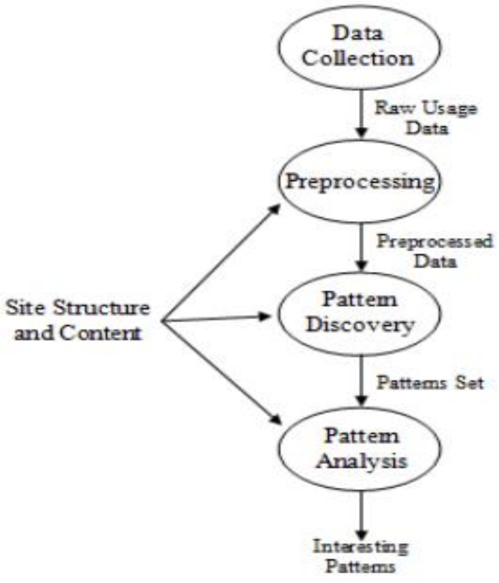

Fig-3: Phases of Web Usage Mining

a) Preprocessing: It helps in retrieving the raw data from web **resources and then processes the data.**

b) Pattern Discovery: After preprocessing the data, the data is used for discovering patterns.

c) Pattern Analysis: After discovering the pattern the pattern is analyzed and then the pattern is checked. If the pattern is correct then it is implemented on web to extract the information from web.

3. WEB CONTENT MINING

Web content mining is the mining, scanning and extraction of text, videos, graphs and pictures from web documents. Web content mining analyzes the content of web resources. Content data correspond to collection of facts a web page was designed to convey to the users. Most of the data available on the web is unstructured data.

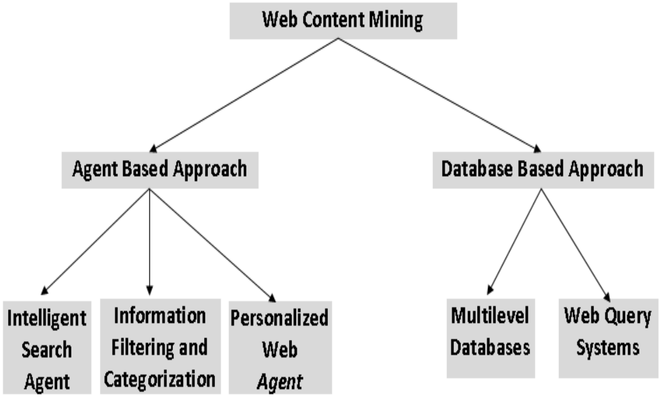

Fig-2: Web Content Mining

The automated discovery of Web-based information is difficult because of the missing structure of information sources in the World Wide Web. A help in finding information are traditional search engines such as Google, Lycos, Alta Vista or WebCrawler. But the problem is that they do not provide structural information by categorizing, filtering or interpreting documents. Two different points of view of web content mining are: the information retrieval view and the database view. The main goal of Information retrieval view is to improve the filtering and finding of the information to the users.

Two types of approaches are used in web content mining. The two approaches are: the *database approach* and the *agent based approach*. The database approach helps in retrieving the semi-structured data from web documents. The agent based approach searches relevant information and helps in organizing the collected information. Researchers have developed more intelligent tools for information retrieval, such as intelligent Web Agents or Multilevel Databases.

(i) Agent based Approach: The agent approach uses so called Web agents to collect relevant information from the World Wide Web. A Web agent is a program that visits a Web site and filters the information the user is interested in. There are three subtypes for the agent based approach: Intelligent Search Agents, Information Filtering/Categorization and the Personalized Web Agents.

(ii) Database based Approach: The database approach for Web mining tries to develop techniques for organizing semi structured data stored in the Web into more structured collections of information resources. Standard database querying mechanisms and data mining techniques can be used to analyze those collections then. The database approach can be divided into two subtypes: *Multilevel Databases* and *Web Query Systems.*

a) Multilevel Databases

Several researchers have proposed a multilevel database approach to organizing Web-based information. The main idea behind these proposals is that the lowest level of the database contains primitive semi-structured information stored in various web repositories, such as hypertext documents. At the higher level(s) meta data or generalizations are extracted from lower levels and organized in structured collections such as relational or object-oriented databases.

b) Web Query Systems

There have been many web-base query systems and languages developed recently that attempt to utilize standard database query languages such as SQL, structural information about web documents, and even natural language processing for accommodating the types of queries that are used in World Wide Web searches. We mention a few examples of these Web-base query systems here. W3QL combines structure queries, based on the organization of hypertext documents, and content queries, based on information retrieval techniques. WebLog is a logic-based query language for restructuring extracted information from Web information sources. Lorel and UnQL support querying of heterogeneous and semi-structured information on the Web using a labeled graph data model. TSIMMIS helps to extract data from heterogeneous and semi-structured information sources and correlates them to generate an integrated database representation of the extracted information.

3.1 Unstructured Data Mining

Text document is the form of unstructured data. Most of the data that is available on web is unstructured data. The research of applying data mining techniques to unstructured data is known as knowledge discovery in texts.

3.1.1 Information Extraction

To extract information from unstructured data that is present on web pattern matching is used. It traces the keywords and phrases and then finds out the connection of keywords within text. When large volume of text is there then the technique is very useful. Information extraction transforms unstructured text to more structured form. First, from extracted data the information is mined, then using different types of rules, the missed out information is found. Information extraction making incorrect predictions on data is discarded.

3.1.2 Topic Tracking

The documents relating to the users interest are predicted by checking the documents visited by the user and by studying the user profiles. For example The topic tracking is applied by yahoo, user gives a keyword and if anything related to keyword pops then the user is informed about that. Many fields like medical field and educational field use this technique for finding the recent developments in their respective fields. The disadvantage of the technique is that when we search for our topic then it may provide us with information which is not related to our topic.

3.1.3 Summarization

The technique summarizes the document by maintaining the important points. It helps the user to decide whether to read the topic or not. The summarization technique uses two methods that is the extractive method and the abstractive method. The extractive method selects a subset of phrases, sentences and words to form the summary from the original text. The abstractive method builds an internal semantic representation and then uses natural language generation technique to create the summary. This summary may contain words which are not present in the original document.

3.1.4 Categorization

This technique identifies the main theme by placing the documents in a predefined set of group. The technique counts the number of words in the document and this decides the main topic. According to the topic the rank is given to the document. The documents with majority contents on particular topic are given first rank. This technique helps in providing customer support to the industries and business.

3.1.5 Clustering

Clustering is the process of grouping similar web documents together in such a way that the documents belonging to the same group are similar and those belonging to different groups are dissimilar. In this grouping of documents is not done on the basis of predefined topics. It is done on fly basis. Some documents may appear in different group. As a result useful documents are not omitted from search results. This technique helps user to select the topic of interest.

3.1.6 Information Visualization

This technique uses feature extraction and key term indexing. Similarity between the documents are found out through visualization. Large textual materials are represented as visual maps or hierarchy where browsing facility is allowed. It helps in visually analyzing the content.

3.2 Structured Data Mining

Structured Data mining techniques are used to extract structured data from web pages [5]. Data in the form of list, tables and tree are the examples of structured data. The advantage of structured data is that it is easy to extract as compared to unstructured data.

3.2.1 Web Crawler

A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. The major search engines on the Web all have such a program, which is also known as a "spider" or a "bot." Search engines use crawlers frequently to collect information about what is available on public web pages. There are two types of crawlers. They are internal and external web crawler. Internal web crawler crawls through internal pages of the website and the external crawler crawls through unknown websites.

3.2.2 Page Content Mining

This technique focuses on classification of web pages by comparing the page content rank given by traditional search engines.

3.2.3 Wrapper Generation

The information is provided by the wrapper generator on the capability of sources. Web pages are ranked by traditional search engines. By using the page rank value the web pages are retrieved according to the query.

3.3 Semi-Structured Data Mining

Semi-structured data arises when source does not impose rigid structure on data. If we want to extract data from web page and populate that data in database.

3.3.1 Object Exchange Model

The relevant information is extracted from semi-structured and is collected in a group of useful information and is then stored in Object Exchange Model (OEM). This helps the user to accurately understand the structure of the information that is available on web.

3.3.2 Top down Extraction

Top down extraction technique focuses on extracting Complex objects from web sources and decomposing them into less complex objects until atomic objects are extracted.

3.3.3 Web Data Extraction Language

This technique converts web data to structured data. This structured data is then delivered to end users. The data is stored in the form of tables.

4. CONCLUSION

In this paper we have studied an overview of how to mine the data available on web. We further discussed about the various types of web mining in the form of web content mining, web structure mining and web usage mining. Large amount of data is maintained by the web sources and that data can be clearly extracted by web mining techniques when the techniques are used accurately according to the requirements of the user. Web content mining has been proved very useful in the areas of e-commerce applications, business world, social networks etc. the problems associated with finding the right information using search engines has turned out to be a big challenge. Web content mining solves the problem and helps the users in fulfilling their needs. Even though many different techniques to mine the diverse types of data on web are available there is a need to further improve the efficiency and effectiveness of retrieving the desired information from web.

REFERENCES

[1] Kosala and Blockeel, “Web mining research: A survey,” SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, vol. 2, 2000.

[2] Hussein, Mohamed-K, and Mohamed-H, Mousa. "An Effective Web Mining Algorithm using Link Analysis." (IJCSIT) International Journal of Computer Science and Information Technologies 1.3 (2010)

[3] Hussein, Mohamed-K, and Mohamed-H, Mousa. "An Effective Web Mining Algorithm using Link Analysis." (IJCSIT) International Journal of Computer Science and Information Technologies 1.3 (2010).

[4] Manoj Pandia, Subhendu Kumar Pani and Sanjay Kumar Padhi. "A Review of Trends in Research on Web Mining." International Journal of Instrumentation, Control and Automation, Volume 1, 2011.

[5] Pol and Kshitija. "A Survey on Web Content Mining and extraction of Structured and Semi-structured data." Emerging Trends in Engineering and Technology, 2008, ICETET'08, First International Conference ,IEEE, 2008.