International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 2, Mar- Apr 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Study of Web Mining and It's Types

Sabitha S[1], Jagadeeswaran VS[2]

[1]PG Student,Department of Computer Science,Dr.N.G.P Arts and Science College,Tamil Nadu,India

2Associate professor,Department of Computer Science,Dr.N.G.P Arts and Science College,Tamil Nadu,India

ABSTRACT

Mining means extracting something useful or valuable from a base substance like "mining gold from the earth".Web mining is the application of data mining techniques to automatically discover and extracts patterns and information from the world wide web. Web structure mining categories the web pages by the use of hyperlinks and generate the information.Such as checking the similarity and relationship between different websites.It helps in solving the problem of how users are using the website.Web mining is the branch of data mining which deals with searching, filtering and extracting useful data stored in web server databases and logs.

Keywords: Web pages,Patterns,Web content mining,Web structure mining,Web usage mining.

I.INTRODUCTION

Data is growing,internet is overload with information.Businesses have become increasingly data-driven. Every decision,every insight seems to emanate from analysis of web data. Once you retrieve and analyse the datas, you can derive important insights for various aspects of your business. The main purpose of web mining is discovering useful information from the WWW and its usage patterns[1].Web content-textimage,records,etc.Webstructure-hyperlinks,tags,etc.Web usage-http logs,app server logs,etc.

II.CATEGORIES OF WEB MINING

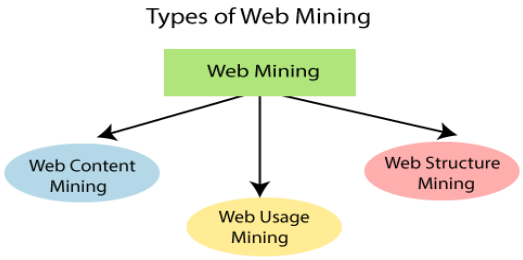

Fig.1 Types of Web mining

All of these three categories focus on the process of knowledge discovery of implicit,previously unknown and potentially useful information from the web.Each of them focuses on different mining objects of the web.

A.WEB CONTENT MINING

The process of discovering useful information from the content of web page or web document.Useful information contains Text,Image,Audio,Video.Web content mining also known as "Web Text Mining"Content data is a group of facts that a web page is designed.This type of mining performs scanning of the text,images and group of web pages.

Web Content Mining mines out just the contents of the web without any specific grouping or pattern. For actual usage of content mining, two approaches are used based on the content present i.e. Unstructured text mining approach and Semi-Structured and Structured mining approach.

[3] **a. Unstructured text mining approach**: This is also called as text data mining or text mining. The research on the mining techniques on the unstructured text data is termed as Knowledge Discovery in Texts.

b. Semi-Structured and Structured mining approach: The structured data on the web are easier to extract compared to unstructured texts. Semi-structured is a combination of the Web dealing with documents and database communities dealing with data

B.WEB USAGE MINING

It is the type of web mining activity which predicts about which pages are likely to be visited in near future based on the active user's behaviour.Such pages can be pre-fetched to reduce access time.The usage data records the users behaviour when the user browses or makes transactions on the website in order to better understand of user.Automatic discovery of patterns from one or more web servers.Web servers,web proxies and client application can easily capture web usage data.

Usage mining is done in three phase approach i.e. data preparation, pattern discovery and pattern analysis phase.

1. Data Preparation/Pre-Processing of data: The usage data is mined from the Web clients, proxy servers and servers. This data is processed identifying the users, user sessions and so on. The information is generally found in user log available in the browsers. This involves three steps i.e. Data Cleansing, Data Transformation and User/Session Identification.

a. Data Cleansing: This is process of obtaining useful information and removing the unwanted from the log data.

b. Data Transformation: Using data mining techniques like clustering, grouping together several requests. All these requests are analysed based a session period.

c. User/Session Identification: This is most difficult task in which user and session are identified from log file. This is difficult because of presence of any users on the same computer, proxy servers, dynamic addresses etc.

2. Pattern Discovery: This data is analysed to find out the patterns in them.

3. Pattern Analysis: These patterns are understood to figure out the sequence of data being accessed.

C.WEB STRUCTURE MINING

The structure of a typical web graph is as follows:

Structure InformationWeb pages-as nodes.Hyperlinks-as connection between two related pages(nodes).It is a process of using the graph theory to analyse the node and connection structure of a website.It extracts patterns from hyperlinks in the web.A hyperlink is a structural component that connects the webpage to a different location.

Web structure mining is analysis of hyperlinks with in the web. This is also called as Link Mining, which is a combination of link analysis (old area of research), hypertext and web mining as well as graph mining. The tasks that are possible through link mining are [4]

1. Link-based Classification: This task is to find out the category of the web page depending on factors like word occurrence, links and anchor texts.

2. Link based Cluster Analysis: The data is segmented into groups where similar ones are grouped and others are grouped into different ones.

3. Link Type: This predicts the existence of links, type of link as well as the purpose of link.

4. Link Strength: This tells about the weights of the links.

5. Link Cardinality: This predicts the number of links between two entities.

III.METHODOLOGY

PATTERN MATCHING

Pattern matching is the act of checking a given sequence of tokens for the presence of the constituents of some pattern. In contrast to pattern recognition, the match usually has to be exact: "either it will or will not be a match." The patterns generally have the form of either sequences or tree structures. Uses of pattern matching include outputting the locations (if any) of a pattern within a token sequence, to output some component of the matched pattern, and to substitute the matching pattern with some other token sequence (i.e., search and replace)[2].

DECISION TREE

Decision tree learning is a method commonly used in data mining.The goal is to create a model that predicts the value of a target variable based on several input variables.

A decision tree is a simple representation for classifying examples. For this section, assume that all of the input features have finite discrete domains, and there is a single target feature called the "classification". Each element of the domain of the classification is called a *class*. A decision tree

or a classification tree is a tree in which each internal (non-leaf) node is labeled with an input feature.

IV.CHALLENGES IN WEB MINING

The web is too huge-Size of the web is very huge and rapidly increasing.Complexity of web pages-Web pages are very complexas compared to traditional text document.Web is dynamic information source-Information on the web is rapidly updated.Diversity of user communities-User community on the web is rapidly expanding .These users have different backgrounds, interests,and usage purposes.Relevancy of information-A particular person is generally interested in only small portion of the web,while the rest of the portion of the web contains the information that is not relevant to the user.

V. APPLICATIONS OF WEB MINING

Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.It is used for Web Searching e.g.,Google,Yahoo etc and Vertical Searching.Web mining is used to predict user behaviour.Web mining is very useful of a particular Website and e-services e.g.,landing page optimization.Helps to improve the power of web search engine by classifying the web documents.Security and Crime Investigation.

VI. CONCLUSION

Web data mining is considered as sub approach of data mining that focuses on gathering information from web. Web is a large domain that contains data in various forms i.e.: images, tables, text, videos, etc. As size of web is continuously increasing; it is becoming very challenging task to extract information. In this paper we described some basics of Web Content Mining, Web Structure Mining and Web Usage Mining. We analysed their strengths and limitations and provide comparison among them. Finally applications are discussed which specifies a fields where actually web mining is used. So we can say that this paper may be used as a reference by researchers when deciding which techniques are suitable.

REFERENCES

[1] DarshnaNavadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[2] http://en.wikipedia.org/wiki/Web_mining

[3] AbdelhakimHerrouz et al. “Overview of Web Content Mining Tools” The International Journal of Engineering and Science (IJES) (2013) Volume: 2 Issue: 6 and References there in.

[4] Miguel Gomes da Costa Júnior and Zhiguo Gong. “Web Structure Mining: An Introduction” Proceedings of IEEE International Conference on Information Acquisition (2005) Page: 590- 595 and References there in

.

