IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol-6, Special Issue-8, Oct 2018 E-ISSN: 2347-2693

A Survey on Webmining and Web Usuageminig

[1*]Ramajayam G., [2]Soundharya V., [3]Likitha M.S.

[1,2,3]Sri Krishna Adithya College of Arts and Science, Tamilnadu, India

Available online at: www.ijcseonline.org

Abstract- The World Wide Web has extremely huge amount of information and it facilitates the user to search for data by moving from one document to another. Web mining is the application of Data mining and it is the procedure of discovering and extracting fruitful information from extremely large web data. The web is rapidly began to modernize and enlarged. In such case web mining is becoming a challenging task. It has to handle different communities, different external interfaces etc. In this paper we are focusing on web mining process and one of its type, web usage mining. This paper covers the basic concept of web mining and detailed description of web usage mining.

Keywords- Data mining, web mining, web content mining, web structure mining, web usage mining.

I. INTRODUCTION

DATA MINING?: The major reason for data mining to attract a great deal of attention in information industry is due to the wide availability of enormous amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from business management, production control, and market analysis, to engineering design and science exploration. Data mining can be viewed as a result of the natural evolution of information technology.

An evolutionary path has been witnessed in the database industry in the development of the following functionalities data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining). For instance, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, data analysis and understanding has naturally become the next target.

Data mining is also known as "KNOWLEDGE DISCOVERY IN DATABASE".

Fig.1

II. WEB MINING

Data mining is to turn data into knowledge, while web mining is to apply data mining strategies to extract and uncover knowledge from web documents and services.In customer relationship management (CRM), Web mining is the combination of information congregated by standard data mining methodologies and techniques with information gathered over the World Wide Web(*Mining* means extracting something useful or valuable from a baser substance, such as mining gold from the earth.) Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign. Web mining allows you to look for patterns in data through content mining, structure mining, and usage mining. Content mining is used to examine data collected by search engines and Web spiders. Structure mining is used to examine data related to the structure of a particular Web site and usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions.

Types of web mining:

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types –

- Web content mining
- Web structure mining
- Web usage mining

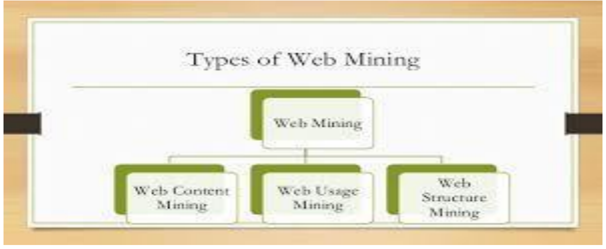


Fig.2

The below section consist of the detailed description on WEB USAGE MINING .

<u>*Web usage mining:*</u>

Web usage mining specifically performs mining on Web usage data, or web logs . The listing of page reference data is known as Web Log.It is sometimes referred to as clickstream data as each entry corresponds to the mouse click.These logs can be examined from either a client perspective or a server perspective. The information of user is detected by evaluating a client's sequence of clicks.This could be used to perform prefetching and caching of pages.

Web usage mining can be used for many different purposes.By aiding in personalization, a profile about that user could be developed , thus looking at the sequence of pages a user accesses.With site mining , the overall quality and effectiveness of the pages at the site can be evaluated.One taxonomy of web usage mining application has included [SCDT00]

- By keeping track of previously accessed pages, personalization for a user can be achieved.These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.

Needed links can be identified to improve the overall performance of future accesses by determining frequent access behavior for users.

- Information related to frequently accessed pages can be used for caching.
- Identifying common access behaviors can be used to upgrade the actual design of Web pages and to make other modifications to the in addition to modifications to the linkage structure.The behavior of the customers can be compared with that for those who do not purchase anything.This can be used to identify the changes to the overall design.Sometimes there are circumstances where many visitors never get past a particular page. That target page can be improved in an attempt to turn these visitors into customers.
- To gather business intelligence to improve sales and advertisement,web usage patterns can be used.

Web usage mining actually consists of three separate types of activities [SCDT00]:

- ➢ Preprocessing
- ➢ Data Structures
- ➢ Pattern discovery
- ➢ Pattern analysis

1.PREPROCESSING :

The Web usage log might not be in a format that is usable by mining applications. The data may need to be reformatted and cleansed, as with any data to be used in a mining application. In addition, there are some issues specifically related to the use of Web logs. Steps that are part of the preprocessing phase include cleansing , user identification , session identification , path completion , and formatting[CMS99].

Standard log data consist of the following: source site , destination site, and timestamp. The source and destination sites can be listed as a URL or an IP address . The source site is identified by a user ID and the destination site is identified by a page ID . Additional data like Web browser information may also be included. The data may be changed in several ways before processing the log .For security or privacy reasons , the page address may be changed into unique page identifications . This conversion will save the storage space . The data can be cleansed by removing any irrelevant information.

The grouping of data together from the log can provide more information. All pages visited from one source could be grouped by a server to better understand the patterns of page reference from each user. In the same manner, patterns from groups of sites may be discovered . References to the same site may be identified and examined to better understand who visits this page.

The division of the log records into session is a common technique for a sever site. A session is a set of page reference from one source site during logical period. A session would be identified by a user logging into a computer, performing work, and the logging of , historically. The logical start and end of the session is represented by login and logoff. With web log data , this is harder to determine . Several approaches can be used to identify these logical periods:

- Combine all records from the same source site which will occur within a time period .
- Add records to a session if they are from the same source site and the time between two consecutive timestamps is less than a specific threshold value.

The correct identification of the actual user is the most center problem around the preprocessing activities. User identification is complicated by the use of proxy servers, client side caching, and corporate firewalls. It is difficult to track a person who is actually visiting a site. Users who access the Internet through an Internet service provider will all have the source location of that provider . It is not unique to the individual . Sometimes the same user may use different ISPs. Cookies can be used to assist in identifying a single user regardless of machine used to access the web . A *cookie* is a file that is used to maintain client-server information between access that the client makes to the server. At the client side , the cookie file is stored and sent to the server with each access.

Path completion is an attempt to add page accesses that do not exist in the log but it actually occurs . missing pages can be easily added. Algorithms are used to infer missing pages as well as to generate an approximate timestamp.

2. DATA STRUCTURES

To keep track of patterns identified during the Web usage mining process, several unique data structures have been proposed .Trie is the only possible basic alternative data structure. A *trie*is a rooted tree , where each path from the root to a leaf represents a sequence. Tires are used to store strings for pattern-matching applications. Each character in the string is stored on the edge to the node. Common prefixes of strings are shared. The problem faced in using tries for many long strings is the space required.

The compressed trie is called a *suffix tree* . A suffix tree has the following characteristics :

- Each internal node except the root has at least two children .
- A nonempty subsequence is represented at each end .
- Sibling edges begin with different symbols which represents the subsequence .

With the help of suffix tree , we can find the common subsequences among multiple sequence and also it is efficient to find any subsequence in a sequence. A suffix tree can also be constructed from a sequence in time and space linear in the length of the sequence. Many different patterns may be found when given a session of page references . The exact number of patterns depends on the exact definition of the pattern to be found .

The slight variation on the suffix tree that is used to build a suffix tree for multiple sessions is called a *generalized suffix tree*(GST).

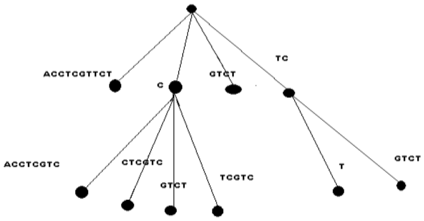


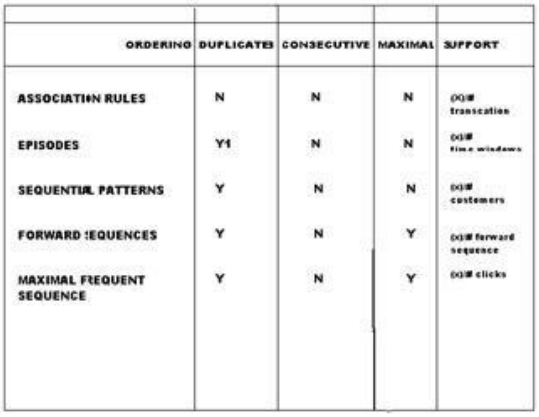
Fig.3 It is a suffix tree

3. PATTERN DISCOVERY :

Pattern discovery is the most common data mining technique used in click stream data is that of uncovering traversal patterns .A set of pages visited by the user in a session is called a traversal pattern . Other types of patterns may be discovered by web usage mining . Association rules can look at pages accessed together in one session independent of ordering . To provide a clustering of the users , similar traversal patterns may be clustered together . This is not similar from clustering of pages , which tends to identify same pages, not users .

A variety of traversal patterns have been examined . It differs from how the patterns are usually defined .The difference between the different types of patterns can be described by the following features :

- Duplicate page reference (backward traversals and refreshes / reloads) may or may not be allowed .
- A pattern may be composed alternatively of any pages referenced in the same session, or only of contiguous page references .
- It is not definite for the pattern of references to be maximal in the session. A frequent pattern is maximal if it has no subpattern that is also frequent.



ORDERING	DUPLICATES	CONSECUTIVE	MAXIMAL	SUPPORT
ASSOCIATION RULES	N	N	N	(x)# transcation
EPISODES	Y1	N	N	(x)# time windows
SEQUENTIAL PATTERNS	Y	N	N	(x)# customers
FORWARD SEQUENCES	Y	N	Y	(x)# forward sequence
MAXIMAL FREQUENT SEQUENCE	Y	N	Y	(x)# clicks

Fig.4

Serial episodes are ordered , parallel episodes are not , and general episodes are partially ordered.

Pattern found using different combinations of these three properties may be used to discover difficult features and thus may be used for different purposes. Knowledge of contiguous page references and thus for prefetching and caching purposes . Knowledge of backward transversal often followed can be used to improve the design of a set of web pages by adding new links to shorten future traversals . The maximal property is used primarily to reduce the number of meaningful patterns discovered . The use of such performance improvements as user side caching may actually alter the sequences visited by a user and impact any mining of the web log data at the server side.

ASSOCIATION RULES :

Association rules can be used to find what pages are accessed together . Here we are really finding large itemsets.

SEQUENTIAL PATTERNS :

A *sequential pattern*(as applied to web usage mining) is defined as an ordered set of pages that satisfies a given support and is maximal i.e., it has no subsequent that is also frequent).

FREQUENT EPISODES :

An *episode* is a partially ordered set of pages.

A *serial episode* is a episode in which the events are totally ordered.

A *parallel episode* is a set of events where there need not be any particular ordering.

A *general episode* is one where the events satisfy some partial order.

MAXIMAL FREQUENT FORWARD SEQUENCES :
One approach to mining log traversal patterns is to remove any backward traversals. Each raw session is transformed into *forwardreference* , from which the traversal patterns are mined using improved level wise algorithms . The "real" access patterns made to get to the really used pages would not include backward references . Backward references is included only because of structure of pages . The resulting set of forward references are called *maximal forward references*.

4. PATTERN ANALYSIS :
Once patterns have been discovered , they must be analyzed to determine how that information can be used . Some of the generated patterns may be deleted and determined not to be of interest .

Recent work has proposed examing web logs not only to find frequent types of traversal patterns , but also to identify patterns that are of interest because of their uniqueness or statistical properties . Patterns that are found are not necessary to have contiguous page references . A web mining query language ,*MINT,* facilitates the statement of interesting properties .

The idea of a sequence is expanded to the concept called g-sequence . A g-sequence is a vector that consists the pages visited and the wild cards . The events need not be contiguous with the use of wild cards . More complicated g-sequences can indicate specific constrains on the number of events that replace the wild card . Selection of patterns that satisfy a g-sequence template are accomplished with MINT.
Two patterns are comparable if their g-sequences have at least the first n pages the same . Here *n* is supplied by the user.

The goal of this work is to increase the number of customers . Noncustomer patterns with no comparable customer patterns indicate that some changes to the link or the web page designs may be in order . The project proposes rules and the use of a proxy server to dynamically change the link structures of pages .

III. CONCLUSIONS

In this paper we survey the research area of Web mining, focusing on the category of Web usage mining. We also discussed the different algorithms used in web structure mining. Web mining deals with retrieving the data from web with best output. The web structure mining also deals with many algorithms that lead to fetch the data from any website. In general web structure mining is that retrieves the data from website for online user in effective manner. Since this is a vast area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

REFERENCES

[1]. https://cs.wmich.edu/~yang/teach/cs595/han/ch01
[2]. https://searchcrm.techtarget.com/definition/Web-mining
[3]. http://cyberartsweb.org/cpace/ht/lanman/wsm1.htm
[4]. https://pdfs.semanticscholar.org/5b6d/a3ba6338326facfab93d53927cc300953547