

A Survey on Web Pattern Mining



Authors

¹Prashasti Kanikar, ²Neha Mishra *

Address For correspondence:

Professor, Department of Computer Science, NMIMS University, Mumbai
Department of Computer Science, NMIMS University, Mumbai

Abstract: - Web mining can be described as the process of discovering and analyzing useful information from the World Wide Web. Web mining can be classified into three categories: web structure mining, web usage mining and web content mining. Web usage mining is basically used to find the navigational behavior of the user on the World Wide Web which can be used to personalize web pages. The web usage mining process includes pre-processing of web log data. After that frequent item discovery, frequent sequence discovery and frequent sub tree discovery steps are being carried out. Patterns are extracted by pattern analysis tools from the web log data. Various pattern discovery techniques are applied on the web log data. Various algorithms have been discovered for web usage mining. Page rank algorithm can be used to provide ranking to a web page. The ranking of a page depends upon the number of back links pointing to a particular page. Another algorithm which is most commonly used is HITS algorithm which calculates the hub pages as well as the authoritative pages. FP tree algorithm is used to generate a candidate generate and test tree which provides the sequence of the web pages in which they are accessed. Another approach is to calculate the significant interval and confidence which can be used to discover frequent patterns in web log data.

Keywords- Web Data Mining, HITS, Page Rank, Web Usage Mining, Web Structure Mining

1.1 Why Web Usage Mining

In the current scenario, most of the companies are doing their business through the E-commerce. E-

Commerce is basically a process of conducting business over the internet. So, web usage mining provides a way to collect information from the web log. Based on this information, the companies can advertise their products

on various websites depending on their behavior. A millions of users daily interact with websites around the world, large amount of data are being produced. This information is very precious for companies and websites in the field of understanding customer behavior, improving services for customers and relationships, measuring the success of marketing efforts, launching target marketing campaigns and so on.

1.2 Way to perform Web Usage Mining

In order to perform web usage mining, first visitors traffic needs to be monitored which is based on web server log files and other source of traffic data. Webmasters and system administrators uses web server log files for the purposes of getting information that how many requests fails and what kind of errors are being generated.

Web server log files are also used to record visitor's on-line behavior to getting information about the search engines that are widely used by the visitors and which browsers are most popular amongst the users. One way to collect web traffic data is web log file. The other way is to "sniff" TCP/IP packets as they cross the network, and to "plug in" to each web server. Once the web log data is collected from web log files then it may be combined with other relational databases, over which the data mining techniques are implemented.

The above is the brief explanation of how Web usage is done shown in fig 1 [3].

misniya@gmail.com * Corresponding Author Email-Id

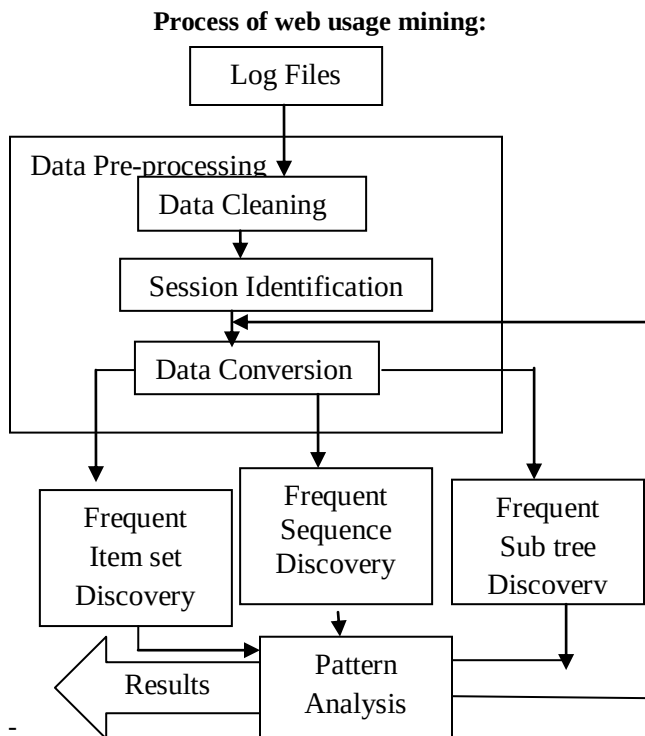


Fig 1: Process of Web Usage Mining [3]

1.3 Data Pre-processing

The web log files which are collected from various sources needs to be cleaned/filtered, by using some techniques like filtering the raw data to eliminate irrelevant items and outliers and, grouping individual page accesses into semantic units.

Irrelevant items can be eliminated by checking the suffix of the URL name, which describes the format of the files. After elimination of irrelevant items, the data is integrated from all sources to form a visitor profile data.

1.4 Pattern Discovery Techniques:

1.4.1 Converting IP addresses to Domain Names

Each user connects to the internet through an IP address. Every IP address has a corresponding domain name, and these are linked through the Domain name system (DNS). DNS can convert the domain name entered by user into corresponding IP address.

It is difficult to gain any knowledge from the IP address. When the IP address is converted into the domain name, then some information may be achieved.

1.4.2 Converting File Names to Page Titles

It is easy to read the page titles than URLs, so a good system should look at these files and determine their titles.

1.4.3 Path Analysis

For path analysis, graph models are used. In graph models, graph represents some relation defined on web pages and each tree of the graph represents a website. Each node in the tree represents a web page (html document), and edges between trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site.

1.4.4 Grouping

Various kinds of information can be obtained by grouping similar information.

1.4.5 Filtering

Filtering specifies certain information from the website. For example, for a particular week, the total number of visitors for a particular website can be counted by the administrator. The filter in this example is “particular week”.

1.4.5.1 Dynamic Site Analysis / Vignette Story Server

Static HTML pages were used by traditional web sites. A number of companies allow creating a dynamic HTML web page around the database. So it offers the advantages of flexibility, centralized storage and version control. But it is also problematic in web traffic analysis because the URL that is looking very simple are replaced by a long line of parameters and cryptic ID numbers.

1.4.5.2 Cookies

The IDs which are randomly assigned to a web browser is known as the cookies. Cookies are assigned by a web server to the web browser when a browser first connects to the particular website. These cookies are independent of the IP address. These IDs are send by the browser to the server on subsequent visits so that server would be able to recognize that the same user is returned back. The cookies are a mechanism to inform server that the same visitor has returned back. Cookies only store session information. They never hold the password or credit card number. This information are collected at the time when user fills a form on the internet. It is used to track the repeat visit of the same user.

1.4.6 Association Rules

The spending habits of the users can be tracked by the association rules on some related products. For example, each item is represented in a separate URL during the transaction of a online shopper. The buying behavior of the user is stored in a log file which can be later used for knowledge discovery. The measurement between the used keywords and the products that are actually sold can be determined by the association rules.

1.4.7 Sequential Patterns

The process of discovering the inter-transaction patterns in such a way so that one item set is followed by another item in the same time stamp series.

1.4.8 Clustering

The visitors who possess the common characteristics are identified by the clustering. After getting the visitor's profiles, the number of clusters can be identified within a group of profiles. After that we can find the number of clusters that best represents the most profiles.

There are some sources to obtain user information such as web log files and online survey of customers etc. These information of the user can be used to predict the future behavior of the customer's and based upon this information the advertisement can be made for products.

1.4.9 Decision Trees

A flow chart which is essentially a set of questions or data points that ultimately draw some conclusion based on which the decisions can be made is known as decision tree. In product selection systems, decision tree systems are incorporated by many vendors. This step is needed when the user comes to a particular website to purchase some product

1.5 Web Mining applications:**1.5.1 Measuring Return of Online Advertising Campaigns**

The companies are widely using online advertisement for their product marketing. The measurement is basically done into two main areas:

1.5.1.1 Quantity: The number of impressions delivered for each ad banner and page, and the number of clicks on the ad banner determines the quantity.

1.5.1.2 Quality: Various users clicked on the ad banner but the measure of the number of the users who actually purchased shows the quality of the product.

1.5.2 Measuring Return of E-Mail Campaigns

Custom URLs or query strings are assigned to each prospect. When the prospect reads the message and clicks the URL, the Web traffic analysis program determines who the visitor is and begins the appropriate sales process.

When a user click on the URL; the unique identifier is passed to the website. By using a filter, based on the query string, it is possible to measure the best leads. When the e-mail address is linked to a customer profile, sales personnel can receive reports showing contact names, phone numbers and more.

1.5.3 Market Segmentation

Profiling system is used for market segmentation. So that web marketers can use this information to target campaigns and messages to specific group.

2. LITERATURE REVIEW

Web contains a large amount of data. Knowledge can be extracted from a large amount of data but it is very difficult task to obtain information from such a huge database. The data is available in the form of semi-structured data on the web.

Web mining approaches such as web content mining, web structure mining and web usage mining are used to extract and produce useful knowledge from the web [4]. The process of web usage mining consists of three major steps: (1) data pre-processing, (2) pattern discovery, and (3) pattern analysis stages. Log files can be stored on the client browser, server side or on the proxy server. Basically, it is difficult to obtain information from all these log files. To obtain reliable results, the information from all these sources are needed. But because of the difficulty to obtain log files from all these sources, most of the algorithms are implemented on the server side log files. When the data is collected from server side log, then pre-processing steps are being carried out.

The steps of the data pre-processing are data cleaning, session identification and data transformation respectively. The pattern discovery phase involves the discovery of frequent sequences. The pattern analysis phase involves the analysis of the frequent patterns generated by the pattern discovery phase and are used for data gathering [4].

The significant intervals used for the lock and unlock operations. These significant intervals can then be used to find a relationship between the Websites. These Intervals are used to discover the frequent pattern [3].

3. VARIOUS ALGORITHMS FOR PATTERN MINING:**3.1 HITS**

HITS consisting a set of pages known as root set, Which is a brief list of web pages returned by search engine [1]. Furthermore this Root Set is get added to the pages pointed by pages in the Root Set and pages that point to pages in the Root Set. This collective set is called Base sets. HITS construct the adjacency matrix A where $A_{ij} = 1$ if there is a link from i to j and otherwise $A_{ij} = 0$. It assigns to each page i an authority weight a_i and hub weight h_i calculate as $a_i(t+1) = \sum_{j \rightarrow i} h_{jt}$, $h_{jt} = \sum_{i \rightarrow k} a_{it}$ using iteration method to get converged value a_i^* , h_i^* . a_i^* is taken to be a value of page i & hub value h_i^* [1].

The Hub score and Authority score for a node is calculated with the following algorithm [1]:

- Start with each node having a hub score and authority score of 1.
- Run the Authority Update Rule
- Run the Hub Update Rule
- Normalize the values by dividing each Hub score by square root of the sum of the squares of all

Hub scores, and dividing each Authority score by square root of the sum of the squares of all Authority scores.

- Repeat from the second step as necessary

3.2 Page rank

Page rank algorithm is implemented in Google search engine. In this algorithm, the set of pages are extracted at the crawl time rather than query time [1].

Page Rank algorithm is given by:

- $PR(A) = (1-d) + d (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$
Where, $PR(A)$ = Page Rank of a page A
 $PR(T_1)$ = Page Rank of a page T_1
 $C(T_1)$ = Number of outgoing links from the page T_1
 d = Damping factor in the range $0 < d < 1$, usually set to 0.85
The Page Rank is calculated by the ratio of the sum of the incoming links to a particular page to the outgoing link from that page [1].

3.3 Algorithm Sequence Tree

- (1) First the sequential database is read [4]. Key-value pairs are obtained from this database where key refers to the unique page name and value refers to the frequency of the web page which is accessed by the user [4].
- (2) This map is then sorted in the descending order of frequency of the keys.
- (3) After that a sequence tree is created using the following steps:

Initially the root node is set to the null value.

Read each row of sequential database.

For every row of sequential database,

Attach the sequence as the branch to the root.

Check whether the sequence is already present or not.

If yes, then

Increment the counter.

else

create a new branch with that sequence.

- (4) Include a header table with number of rows equal to sequences with value one. Each row of header table is a linked list of nodes which specify the position where nodes are present.

- (5) Mining process is performed.

For every row in the header table

If frequency < minimum support threshold

Ignore the row of header table

else

For every node in the particular linked list

If frequency < minimum support threshold

Ignore the path

else

Traverse the tree up till the root and store the path in the file as frequent sequences.

EndFor

EndFor

(6) The frequent sequences identified from the algorithm are stored in a file [4].

3.4 One pass SI and one pass AIISI Algorithm:

3.4.1 Terms and Definition of Significant Interval

The One Pass SI algorithm [4] and One Pass AIISI [5] can be applied on web log data. The time stamp is used to represent time series with an website time stamp model.

A website w is associated with a sequence of timestamp $\{T_1, T_2, \dots, T_n\}$ that describes its access over a period of time.

The notion of periodicity (such as daily, weekly, monthly, etc.) is used to group the website accesses. For each website, the number of accesses at each time point can be obtained by grouping on the timestamp (or periodicity attribute). The term for the number of accesses of each website as access count (ac). Thus the time series data can be represented as $\langle w, \{T_1, a_1\}, \{T_2, a_2\}, \{T_3, a_3\}, \{T_n, a_n\} \rangle$, where T_i represents the timestamp associated with the website w and a_i represents number of accesses. a_i can be referred as the access count of the website w at Timestamp T_i . This is referred to as folding of the time-series data using periodicity and time granularity (e.g., daily on seconds, daily on minutes, weekly on minutes) [5].

3.4.2 Significant Interval

Intervals associated with a website are characterized by confidence and density [5]. The confidence of a time point can be expressed as the ratio (expressed as percentage) of the access count at that point to the periodicity of data collection (number of days or weeks (N)). When the interval consists of several time points, total access count of the interval (the sum of the access count of the points that form the interval) is used. The density of an interval is the ratio of the total access count of the interval to the interval length. A Window can be defined as a time-interval $wd [Ts, Tel]$ where $(Ts \leq Te)$, Ts is the start time and Te is the end time of the interval. An interval associated with a website is represented as $[Ts, Te, ac, l, d, c]$ where ac represents total access count of the interval [7], l denotes the length of an interval $(Te - Ts + 1)$, d indicates the density, and c represents confidence of the interval $(ac/N * 100)$ [5]. Given a time sequence T , minimum confidence $min-conf$ and interval length $max-Len$, we define the interval $wd [Ts, Tel]$ as a Significant Interval in T if:

1) Confidence(c) of $wd \geq min-conf$

2 length (l) of $wd \leq max-Len$

3) There is no window $wd' = [Ts', Te']$ in T such that $Ts' > Ts$ and $Te' < Te$ that satisfies i) and ii) [5].

3.5 Process of Significant Interval Discovery

First the Raw data is collected from various websites and then apply the preprocessing steps [5]. The Preprocessing steps are

3.5.1 Data Cleaning: The data cleaning is the process of extracting information of individual website from the raw data. The data is separated into the different website record and different views are used to show the record of each website. This process is called Data cleaning [5].

3.5.2 Data Folding: In the preprocessing phase, the input data set is combined to periodicity of interest (example, daily or weekly) [5]. This process is called Data Folding. One Pass SI [6] or One Pass AllSI [5] are used to generate the significant Interval. Those Time points satisfied the Minimum Constraint (i.e. min-conf and max Len for One Pass SI [7] or min-Conf for One Pass AllSI [5]) generate the Significant Interval are Discovered. The confidence can be defined as the percentage of the total number of access counts to the value of the periodicity. One Pass SI [6] algorithm is used to determine the significant interval for the periodicity value of one day. In the one Pass SI algorithm, the minimum confidence value is considered but there is no max length of time period. So the intervals which satisfy the minimum confidence can be considered as the part of the significant interval.

3.6 Advantages and Disadvantages of HITS algorithm [2]:

Table 1: Advantages and disadvantages

Advantages	Disadvantages
Relevant authority and hub pages.	Query time cost is expensive.
Sensitive to user query.	Irrelevant authority may be present.
Important pages are obtained.	Irrelevant hubs might be possible.
Induces web graph by finding set of pages.	Mutually reinforcing relationships between hosts.
HITS calculate authority nodes and hubness correctly.	Irrelevant pages in root cause topic drift.

3.7 Advantages and Disadvantages of Page Rank algorithm:

The strengths of Page Rank algorithm are as follows [2]:

- 1. Less Query time cost:** It will take less time to produce output. Page Rank is generated using entire web graph, so it is less susceptible to localize link spam.
- 2. More Efficient:** Page Rank computes a single measure of quality for a page at crawl time. This measure is then combined with a traditional information retrieval score at query time.

- 3. Feasibility:** The page Rank algorithm is more suitable for today's search engines because millions of search needs to be handled within a day. So it is not possible to handle them with HITS algorithm since it performs search on query time.

3.8 Drawbacks of Page Rank algorithm

The following are the disadvantages of Page Rank [2]:

- 1. Rank Sinks:** The Rank sinks problem occurs when in a network pages get in infinite link cycles as shown in the figure 2 below:

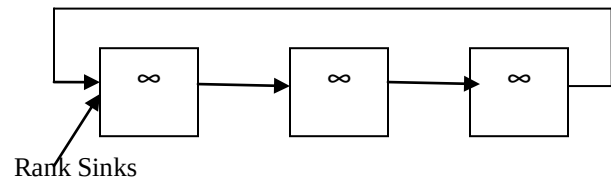


Fig 2: Illustration of Rank Sink [2]

- 2. Spider Traps:** A group of pages is a spider trap if there are no links from within the group to outside the group.
- 3. Dangling Links:** This occurs when a page contains a link such that the hypertext points to a page with no outgoing links. Such a link is known as Dangling Link.
- 4. Dead Ends:** The pages having no outbound link are known as Dead ends.
- 5. Circular References:** If circle references are present in the website, then it will reduce the front page's Page Rank. The figure 3 shown below illustrates the case of circular references.

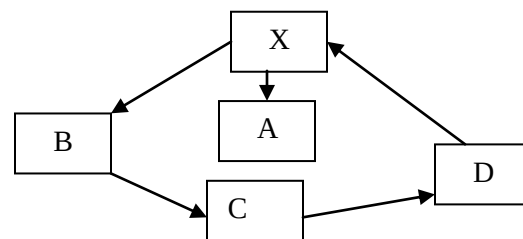


Fig 3: Illustration of circular references [2]

- 1. Effect of additional pages:** By adding more pages in the website, the ranking of front page can be increased but it will decrease the ranking of other pages as well. The solution is to swap links with websites which have high Page Rank value. The simplest way to do this is to make a page with high Page Rank and link it to your front page.
- 8. Page Rank score of a page ignores whether or not the page is relevant to the query at hand.**

3.9 Advantages of sequence tree Algorithm [4]:

- There are only two passes of the data set..
- It actually compresses data set.

- Candidate generation is not performed in this algorithm.
- This algorithm is much faster than apriori.

3.10 Disadvantages of Sequence tree algorithm [4]:

- It might be possible that the size of the sequence tree would be so large so that it may not fit into the memory.
- Sequence tree is expensive to build.
- Trade off: Takes time to build, but once it built, frequent item sets are read off easily.
- Time is wasted as the only pruning that can be done is on single items.
- Support can only be calculated once the entire data set is added to the FP tree.

4. CONCLUSION

It is a revolution that the Internet has grown from a simple search tool to a gold mine. Companies find a new and better way to do business: E-commerce through the Internet. However, E-business cannot just build a web site and then sit back and reap the benefits, which, in most cases, is fruitless. Companies have to implement Web mining systems to understand their customers' profiles, and to identify their own strength and weakness of their E-marketing efforts on the web through continuous improvements. Internet is a gold mine, but only for those companies who realize the importance of Web mining and adopt a Web mining strategy now. So Page Rank algorithm can be used to calculate the ranking of the web page but less reliable as compared to HITS (Hyperlink Induced Topic Search) because it gives more authoritative web pages.

Hubs contain authoritative pages which provide more reliable information. Frequent pattern tree algorithm can be used to determine the pattern of the web usage in which it ignores the branches which are less than the threshold values. For frequent pattern discovery, frequent interval is calculated which gives the confidence of the web page. Based on the confidence value, the web usage pattern of a user can be determined.

REFERENCES

- [1] Pawan Singh, Amit Kumar and Prashast, “*Studies on Research and Development in Web Mining*”, International Journal of Computer Applications (0975 – 8887) Volume 44– No.9, April 2012
- [2] Nidhi Grover and Ritika Wason, “*Comparative Analysis Of Page Rank And HITS Algorithms*”, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 8, October 2012
- [3] Renata Ivancsy and Istvan Vajk, “*Frequent Pattern Mining in Web Log Data*”, Department of Automation and Applied Informatics, and HAS-BUTE Control Research Group, Vol. 3, No. 1, 2006
- [4] Rajashree Shettar, “*Sequential pattern mining from web log data*”, International Journal of Engineering Science and Advanced Technology, Volume-2, Issue-2, 204 – 208
- [5] Dr. Kanak Saxena¹ and Mr. Rahul Shukla, “*Significant Interval and frequent Pattern Discovery in Web Log Data*”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 3, January 2010
- [6] Sagar Savla, Sharma Chakravarthy, “*A Single Pass Algorithm for Discovering Significant Intervals In Time series data*”, International Journal of Data Warehousing and Mining, Volume 3, Issue 3 , IGI Global
- [7] Sagar Savla and Sharma chakravarthy, “*An Efficient single pass approach to frequent episode discovery in sequence data*”, IET 4th International Conference 2008