International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

A Survey on Pattern Discovery of Web Usage Mining

Soundharya V[1], Ram kumar R[2], Prakash B[3], Sowndarya B[4], Prathiksha B[5]

[1,4,5]*B.Sc. Student, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore, India*

[2,3]*B.Sc. Student, Department of Computer Science, Kongunadu Arts and Science College, Coimbatore, India*

***Abstract*: In recent year, Internet technology with develop the growth of World Wide Web overreached all expectations. A web mining is application of data mining on web data and web usage mining is an important part of web mining. A huge amount of information is available in different formats and retrieving the data in that is too difficult so one possible approach to solve the problem is web usage mining. The web is rapidly begun to modernize and enlarged. In such case web mining is becoming a challenging task. Web usage mining is to understand the behavior of web site users through the process of data mining of web Access data. In this paper, we are focusing on web usage mining and description of WUM. This paper covers the basic concept of pattern discovery of web usage mining**

Keywords: web usage mining, Pattern discovery, association rules, sequential patterns, frequent episodes, maximal frequent.

I. Introduction

Mining mean extracting something valuable from baser substance. Web mining is an application of data mining techniques to discover and retrieve useful information from www. In other word, Web mining refers to information or pattern are extracting from web. Web mining enables one to discover web pages, text documents, multimedia files, images and other types of resources from web. Web mining divided into three parts web content mining, web usage mining, and web structure mining. In that Web usage mining refers to the automatic discovery analysis of pattern in clickstream and associated data collection or generated as a result of user interactions with web resource on one or more website. WUM can be divided into three inter-dependent stages are pre-processing and data collection, pattern discovery and pattern analysis.

Fig. 1. Web mining

II. Web Usage Mining and its Stages

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 click stream 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 pre-fetching and caching of pages.

Fig. 2. Pictorial representation of web mining

Web usage mining can be used for many different purposes. By keeping track of previously accessed pages, personification for a user can be achieved. These pages can be used to identify the typical browsing behaviour 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 behaviour for users. Information related to frequently accessed pages can be used for caching. Identifying common access behavior can be used to upgrade the actual design of Web pages and to make other modifications to the 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. 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:

- Pre-processing
- Data Structures
- Pattern discovery
- Pattern analysis

In the pre-processing stage, the clickstream data is cleaned and divided into asset of user transactions representing the

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

activity of each user during different visits to the site. In 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.

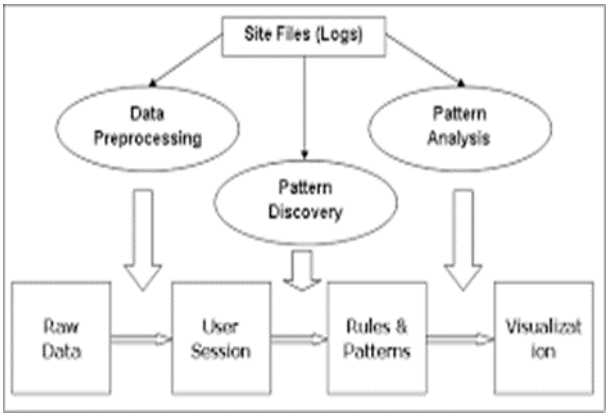


Fig. 3. Web usage mining and its stages

III. 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. 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. It focuses on applying various methods and technique developed from several fields such as data mining, statistic, pattern recognition and machine learning. The discovery patterns are usually represented as collection of pages, objects or resource that is frequently accessed by group of user with common needs. 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 pre-fetching and caching purpose. 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 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.

A. Path Analysis

Graph models are most commonly used for Path Analysis. In the graph models, a 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 and edges between trees represent the links between web sites and the edges between nodes inside a same tree represent links between documents at a website. When path analysis is used on the site as a whole, this information can offer valuable insights about navigational problems. Most graphs are involved in determining frequent traversal patterns and more frequently visited paths in a website. For Example: What paths do users traversal before they go to a particular URL? The second rule indicates an attrition rate for the site. Since many users don't browse further than four pages into the site, it is tactful to ensure that most important information for example product sample, is contained within four pages of the common site entry points. A summary report of hits and bytes transferred.

Fig. 4. Path analysis

Examples:

- 80% of clients who accessed /company/product2 did so by starting at /company and proceeding through /company/new, /company/products and /company/product1
- 78% of clients who accessed the site started from /company/products.
- 65% of clients left the site after 4 or less page references

B. Association Rules

Association rules can be used to find what pages are accessed together. Association rule generation can be used to relate pages that are most often referenced together in a single session. It predicts the association and correlation among set of items where the presence of one set of items in a transaction implies with a certain degree of confidence the presence of other items. That is, it can discover the correlations between pages that are most often referenced together in a single server session/user session.

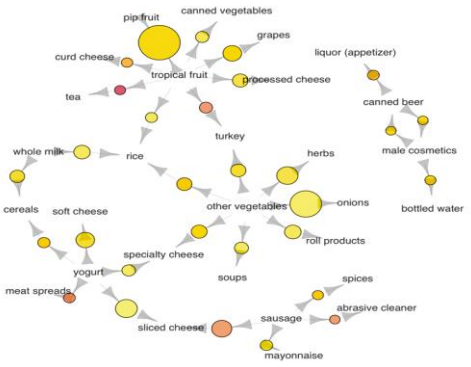


Fig. 5. Association rules

The association rules may also serve as a heuristic to fetch documents in order to reduce user-perceived latency when

International Journal of Research in Engineering, Science and Management
Volume-1, Issue-8, August 2018
www.ijresm.com
ISSN (Online): 2581-5782

loading a page from a remote site. The below image example for association rule which is related to food items.

C. Frequent Episode

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.

D. Maximal Frequent Forward Sequences

One approach to mining log traversal patterns is to remove any backward traversals. Each raw session is transformed into forward reference, 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 are included only because of structure of pages. The resulting set of forward references is called maximal forward references.

E. Sequential Pattern

Sequential mining is involves data mining methods to large web data accessible to extract the sage patterns. The growing popularity experience many visitors everyday experience this from websites World Wide Web. The analysis of what the user browsed can give important for buying pattern customer, the timely and correct decisions made based on the knowledge have helped organization reaching heights in market. Sequential patterns discovery is to find the inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set. Web log files can record a set of transactions in time sequence. Using sequential pattern discovery, useful user trends can be discovered, predictions concerning visit pattern can be made, website navigation can be improved and adopt website contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups.

Example:

- 30% of clients who visited /company/products had done a search in Google, within the past week on keyword w
- 60% of clients who placed an online order in /company/product1 also placed an online order in /company/product4 within 15 days.

F. Clustering and Classification

The main purpose of clustering in web usage mining is to aggregate the similar session together. Self-organized maps, graph partitioning, ant based technique, K-means with genetic algorithms; EM-C Fuzzy means algorithms are the algorithms used for clustering the sessions.

Classification entails assigning labels to existing situations or classes; hence, the term "classification". For example, students exhibiting certain learning characteristics are classified as visual learners.

Classification is also known as "supervised learning technique" wherein machines learn from already labeled or classified data. It is highly applicable in pattern recognition, statistics, and bio metrics.

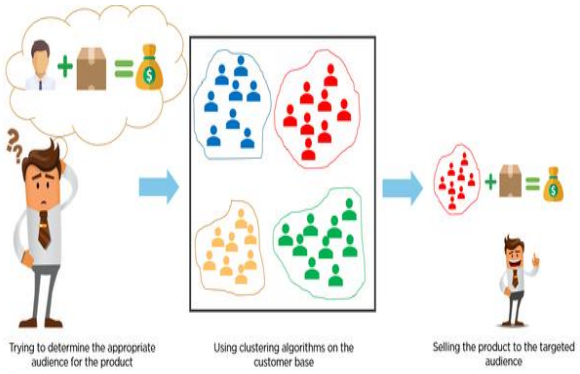
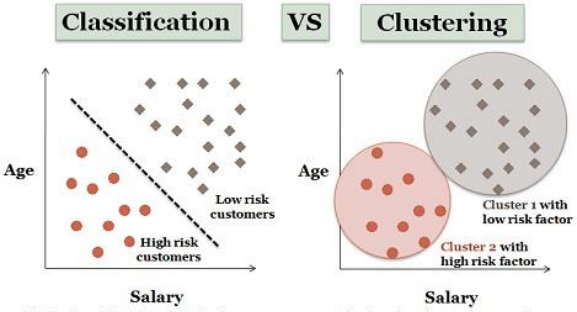


Fig. 6. Clustering and classification

Clustering and classification techniques are used in machine-learning, information retrieval, image investigation and related tasks. Notably, clustering and classification help solve global issues such as crime, poverty, and diseases through data science.



Risk classification for the loan payees on the basis of customer salary

Classification vs Clustering

Criteria	Classification	Clustering
Prior Knowledge of classes	Yes	No
Use case	Classify new sample into known classes	Suggest groups based on patterns in data
Algorithms	Decision Trees, Bayesian classifiers	K-means, Expectation Maximization
Data Needs	Labeled samples from a set of classes	Unlabeled samples

Fig. 7. Classification vs. Clustering

Goal: Clustering group's objects with the aim to narrow down relations as well as learn novel information from hidden patterns while classification seeks to determine which explicit group a certain object belongs to.

IV. Conclusion

In this paper we survey the research area of Web usage mining, focusing on the category of Web usage mining. Web mining

deals with retrieving the data from web with best output. Web usage mining is the technique to find useful and interesting information from web usage data. It is useful in e-commerce to improve structure of website, personalization, identifying customer behavior, Business Intelligence. 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] Mele, "Web usage mining for enhancing search-result delivery and helping users to find interesting web content," in ACM 6th International conference on Web search and data mining, 2013, pp. 765–770.

[2] T. T. S. Nguyen, H. Y. Lu, and J. Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," *IEEE Trans*. vol. 26, no. 10, pp. 2574–2587, 2014.

[3] Meenu, Manoj kumar, "A Survey on Pattern Discovery of Web Usage Mining," in Pattern discovery of web usage mining" in *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 3, no. 1, pp. 379-385, 2017.

[4] M. M. Sharma and A. Bala, "An approach for frequent access pattern identification in web usage mining," in *IEEE International Conference on Advances in Computing, Communications and Informatics* (ICACCI), 2014, pp. 730–735.

[5] https://cs.wmich.edu/~yang/teach/cs595/han/ch01