

# Growing Trends and its Relevance in Web Usage Mining

PanthaganiVijayababu

*Assistant Professor*

*Department of Computer Science and Engineering*

*GVR&S College of Engineering &technology,Budampadu ,Guntur*

**Abstract:** With the large quantity of information available on the World Wide Web (WWW), the concern of how to extract useful knowledge from the Web has gained important attention among researchers in data mining and knowledge discovery areas. Web mining is applied to reveal the importance of WebPages and to expect the web domain visits of various users. This article provides a survey of the available literature on Web usage mining and reviews the research and relevance issues in web usage mining

**Index Terms**—web removal, web content mining, web usage mining, web structure mining.

## I. INTRODUCTION

Web mining refers to the attempt of Knowledge Discovery in Data (KDD) from the web. It can be defined as the process of applying data mining techniques to extract useful knowledge from the enormous amount of information available from the web. It is frequently divided into three major areas.

## II. WEB CONTENT MINING

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting connection patterns, clustering of web documents and categorization of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited.

## III. WEB STRUCTURE MINING

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

**Hyperlinks** A hyperlink is a structural element that connects a location in a web page to a different location, also inside the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

**Manuscript Structure:** In addition, the content in a Web page can also be prepared in a tree prepared format, based on the various HTML and XML tags in the page. Mining efforts now have focused on automatically extracting document object model (DOM) structures absent of documents.

## IV. WEB USAGE MINING

Web usage mining is the function of data mining techniques to discover motivating usage patterns from web usage data, in order to appreciate and better serve the requirements of web-based applications. Usage data capture the identity or source of web users along with their browsing behavior on a web site. Web usage mining itself preserve be classified further depending going on the kind of usage data considered:

Web Server Data: User logs are composed by the web server and typically include IP address, page position and access time.

Application Server Data: Commercial application servers such as Web logic, Story Server have significant features to allow E-commerce applications to be built on top of them by little effort. A key characteristic is the ability to path various kinds of business events and log them into application server logs.

Application Level Data: New kind of events can be distinct in an application, and classification can be twisted on for them — generating histories of these events. It must not be renowned, however, that various end applications need a combination of one or more of the techniques functional in the above categories. This paper deals with Web usage mining, for which many data mining techniques such as statistical analysis, clustering, classification, association rules, sequential pattern discovery, and dependency modeling have been applied to Web server logs. Among these, clustering, which has been one of the most frequently used techniques, forms the focus of this study. This paper is organized as follows. The next section provides a review of related literature on Web usage mining. In Section 3, preprocessing steps of log data is introduced, outlining the details related to web log data preprocessing. Techniques on the web usage mining are discussed in section 4. Finally section 5 concludes the paper.

## V. LITERATURE SURVEY

S.Park and Peter et al, develop an evaluation framework in which the performances of the algorithms are compared in terms of whether the clusters (groups of Web users who follow the same Markov process) are correctly identified using a replicated clustering approach. A series of experiments is conducted to investigate whether clustering performance is affected by different sequence representations and different distance measures as well as by other factors such as number of actual Web user clusters, number of Web pages, similarity between clusters, minimum session length, number of user sessions, and number of clusters to form. A new, fuzzy ART-enhanced K means algorithm is also developed and its superior performance is demonstrated. Zhang and Xamo et al, describes a toolset that exploits web usage data mining techniques to identify customer Internet browsing patterns. These patterns are then used to underpin a personalized product recommendation system for online sales. Within the architecture, a Kohonen neural network or self-organizing map (SOM) has been trained for use both offline, to discover user group profiles, and in real-time to examine active user click stream data, make a match to a specific user group, and recommend a unique set of product browsing options appropriate to an individual user. Z. Li et al, present a novel ontology based Web usage mining framework that leverages search engine queries to improve the accuracy of unemployment rate prediction. The proposed Framework is underpinned by a domain ontology which captures unemployment related concepts and their semantic relationships to facilitate the extraction of useful prediction features from relevant search engine queries. In addition, state-of-the-art feature selection methods and data mining models such as neural networks and support vector regressions are exploited to enhance the effectiveness of unemployment rate prediction. M. Belk et al, focuses on modeling users' cognitive styles based on a pot of Web usage mining techniques on client navigation patterns and click stream data. Main aim is to inspect whether exact clustering techniques can group user of particular cognitive style by measures obtained from psychometric test and content navigation behavior. S.G.Matthewset al, presented genetic algorithm (GA)-based solution is described that uses the elastic nature of the 2-tuple linguistic illustration to discover rules that occur at the intersection of fuzzy set borders. The GA-based advance is enhanced from previous work by including a graph illustration and a better fitness function. Wang et al, introduced the concept of throughout-surfing patterns (TSP) and then present a competent method for mining the patterns. Authors propose a compact graph structure, term a path traversal chart, to record information about the navigation paths of website visitors. The graph contains the frequent surfing paths that are required for mining TSPs. Wang et al, propose a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. We made use of the cluster information generate by a self organizing diagram for pattern analysis and a fuzzy inference system to capture the chaotic movement to provide short-term (hourly) and long-term (daily) Web traffic movement predictions. Castellano et al, proposed NEWER (NEuro-fuzzy Web Recommendation), a usage-based Web advice system that exploits the possible of Computational cleverness techniques to dynamically advise interesting pages to user according to their preference. NEWER employs a neuro-fuzzy move toward in order to conclude categories of user's distribution similar interests and to determine a recommendation model as a set of fuzzy rules express the associations between user category and relevance of pages.

## VI. PREPROCESSING STEPS OF LOG DATA

One objective of web usage mining is to extract sequential usage patterns from a large collection of web logs. These patterns can be used to predict users' access patterns, to identify users' intention, and to provide timely

help for using features available on a web site. Since web log records are usually designed for debugging purposes, they need to be preprocessed before applying data mining techniques. Five preprocessing steps have been identified

1. Data Cleaning: Remove irrelevant data such as log records for images, scripts, help files, cascade style sheet, etc.
2. User Identification: To group together records for the same user. Because web logs are recorded in a sequential manner as they arrive, therefore, records for a specific user are not necessarily recorded in consecutive order rather they could be separated by records from other users.
3. Session Identification: To divide pages accessed by each user into individual sessions. A session is a sequence of pages visited by a user. We also call it as a usage sequence.
4. Path Completion: To determine if there are important accesses which are not recorded in the access log due to caching on several levels.
5. Formatting: Format the data to be readable by data mining systems.
6. Once web logs are preprocessed, useful web usage patterns may be generated by applying data mining techniques such as mining association rules, mining clusters, and mining classification rules..

## VII. TECHNIQUES APPLIED IN WEB USAGE

### *MINING*

The important data mining techniques functional in the web domain contain Association Rule, Sequential pattern detection, clustering, path analysis, classification and outlier discovery [19].

#### i) Association Rule Mining:

Forecast the association and relationship among set of items “wherever the presence of one set of objects in an operation implies (with a certain degree of assurance) the presence of extra items.

- 1) Discovers the correlations among pages that are most regularly referenced jointly in a single server session/user conference.
- 2) Provide the information:
  - a) What are the set of pages repeatedly accessed together by web users?
  - b) What page will be fetched next?
  - c) What are paths frequently accessed by web users?
- 3) Associations and correlations:
  - a) Page association as of usage data-user sessions, user transactions
  - b) Page associations from content data-similarity base on content analysis.
  - c) Page associations based on structure-link connectivity linking pages.

Advantages:

- a) Guide for web site reformation – by adding links that be linked pages often viewed collectively.
- b) Improve the system performance by pre-fetching web data.

#### ii) Sequential outline discovery:

It is applied to web access server transaction logs. The purpose is to determine sequential patterns that specify user visit patterns over assured period. That is, the order in which URLs lean to be accessed [21].

Advantages:

- a) Useful user trend can be discovered
- b) Predictions concerning stay pattern can be made
- c) To improve website steering
- d) Personalize advertisements
- e) Dynamically reshuffle link structure and adopt web site inside to individual client necessities or to provide clients by automatic recommendations that best costume customer profiles.

#### iii) Clustering:

It groups together items (users, pages, etc..) that have parallel characteristics.

- a) Page clusters: It consists of groups of pages that appear to be conceptually correlated according to users’ perception.
- b) User Cluster: It consists of groups or user that seem to be behave equally when navigating through a web site.

#### iv) Classification:

It maps a data item into one of some predetermined classes. Example: describing every user’s category via profiles. Classification algorithms be decision tree, naive Bayesian classifier, neural networks, Support Vector

Machine, K-Nearest Neighbor Classifier.

v) Path Analysis:

A technique that involve the generation of some type of graph that represents relations defined on web pages. This be able to be the physical layout of a web site into which the web pages be nodes and links among these pages are directed edges. The majority graphs are involved in formative frequent traversal patterns more frequently visited paths in a web site.

## VIII. CONCLUSION

This paper has tried to deliver a survey of the rapidly rising area of Web usage mining, which is the order of current technology. In this paper a common overview of Web usage mining is offered. Web usage mining is used in many regions. We studied various techniques for pattern discovery. We can further work on web usage mining with the combination of these techniques because we need to design algorithm, which can help to better understand the mined knowledge.

## REFERENCES

- [1] H. Blockeel, Web mining research: a survey, ACM SIGKDD Explorations Newsletter .P.L. Lanzi, Mining interesting knowledge from weblogs: a survey, Data and Knowledge Engineering.
- [2] Park, Sungjune, Nallan C. Suresh, and Bong-KeunJeong. "Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm." Data & Knowledge .
- [3] John Edwards, and Jenny Harding. "Personalised online sales using web usage data mining." Computers in Industry.