MAT JOURNALS

Journal of Web Development and Web Designing
Volume 2 Issue 1

Mining User Interests from User Search by Using Web Log Data

K. Srinivasa Rao[1], Dr. A. Ramesh Babu[2], Dr. M. Krishnamurthy[3]
[1]Hindustan University, Chennai, India
[2]S&H-Maths, Hindustan University, Chennai, India
[3] Dept. of CSE, KCG College of Technology, Chennai, India
E-mail: srinu532@gmail.com, sh@hindustanuniv.ac.in, mkrish@kcgcollege.com

Abstract

Web Usage Mining (WUM) is a kind of data mining method that can be used to discover user access patterns from Web log data. A lot of work has been done already about this area and the obtained results are used in different applications such as recommending the Web usage patterns, personalization, system improvement and business intelligence. WUM includes three phases that are called preprocessing, pattern discovery and pattern analysis. There square measure totally different techniques for WUM that have their own benefits and downsides. We tend to initial describe a way for extracting a worldwide linguistics illustration of a pursuit question log then show, however, we are able to use it to semantically extract the user interests. During this paper extraction of users interest from journal knowledge will be done, that square measure supported visit time and visit density which might be get from an analysis of internet users journal knowledge.

Keywords: *User interests, web log, frequent patterns*

INTRODUCTION

The Web could be a large, varied, dynamic and principally unstructured knowledge repository that provides unbelievable quantity information, and additionally will increase the quality of a way to manage the data from the various perceptions of users, internet service suppliers, business analysts etc. [1–5]. Web mining is split into 3 areas: online page mining (WCM), internet structure mining (WSM), and internet usage mining (WUM). Online page mining could be a method of reading info from texts, images, audio, video, or structured records like lists and tables and scripts. Internet structure mining could be a method of discovering structure info from linkages of web content (inter page structure/hyper link structure). The web usage or log mining is outlined because the method of extracting fascinating patterns from the log knowledge. The log knowledge is consists of matter knowledge and is delineate in customary format (common log format or extended log format). The main goal of internet usage mining is to capture, model and examine |the online log knowledge in such some way that it inevitably determines the usage behavior of web user [6–10].

LITERATURER SURVEY

A. Automatic Identification of User Goals in Web Search Based on the Web query Assigned by the consumer's evaluation the goal, the intention identification is used to give a boost to best of search results.

B. Query-Sets

Document representation model (DRM) relies on the implicit shopper suggestions. Implicit shopper suggestions are mean that the feedback from web logs. Record illustration model is got from computer program queries. the first perform of this DRM is to get the easier results utilizing non-supervised tasks the same as agglomeration and labeling received from computer program queries. Customers are stirred for file illustration. Set on the clicked queries the period offer the simpler

selection of characteristic from the user's issue of read.

C. Learn from Web Search Logs to Organize Search Results

Search results of the effective organization square measure imperative to enhance the utility of the computer programmed. Cluster the search outcome is that the nice methodology to arrange the search outcome. Use the cluster of search results users finds the report quickly. There square measure 2 faults of this methodology are: 1. The clusters do not depends on the fascinating points of users. 2. The cluster labels would not be informative, so as that the identification of correct clusters is tough.

D. Generating Query Substitutions

Query substitution generates the new question to Substitute the user's designed question. This technique makes use of amendment focused on question substitution. The different queries and also the phrases square measure intently regarding the first queries and also the phrases. Question substitution is distinction with question enlargement and question relaxation, the question enlargement by method of pseudo-relevance feedback that is rate and end in aimless approach.

WEB USAGE MINING TECHNIQUES

Web usage mining is the "Applying data mining techniques to web data repositories to extract patterns". Data mining techniques that are commonly used includes association rules, sequential pattern, clustering, and classification.

Association rules are used to find the relationship between attributes from the item set. In web usage mining item set is set of items. Rules are applied to discern pages which are often looked together in order to reveal associations between guidelines to web designers for reorganizing Websites. Sequential pattern is used to discover sequential navigational pattern for user session. Using this approach, useful users' trends can be discovered, and forecast concerning visit patterns can be made. Clustering is a technique to group together items that have similar features. In Web usage domain, there are two clustering groups, user clusters and page clusters. Page clustering generates the group of pages that are considered to be related according to user view. In user clustering the goal is to group users which have same browsing patterns. Such understanding can be used in business to perform market segmentation and Web site personalization [7]. Created a model by applying clustering algorithm, and then the model is adjusted by statistical approach based on the change of behavior of users or data domain of website periodically [11, 12]. Proposed to integrate Markov model based sequential pattern mining with clustering. Classification is a method that maps a data item into one of several predefined classes. In Web usage mining the users is in different classes according to their profiles.

RELATED WORK

Pre-Processing Log data

In the pre-processing phase, sample server log file, was processed to transform the raw data into structured information. The purpose of data cleaning is to eliminate irrelevant items [13].

Log Data File

The raw data for mining purpose is collected from NASA website. The records of ten days are considered for further analysis. It contains approximately 4,00,000 records in Common log file format.

Data Cleaning

Log knowledge is keep in info for any process of information by means that of queries and program. Data file obtained

was terribly immense and it takes nearly eightieth of total time to mine the information. In knowledge cleansing method, the unwanted data is off from the log info.

The data cleaning takes the following steps:

Step 1: Removal of the entries having image files, graphic or multimedia files. The records which are accessing file with extension gif, jpg, jpeg etc. are to be removed. After performing this step around 1,23785 records left.

Modules

Web Usage Mining and Pattern Discovery

Web Usage Mining is that the application of information Mining techniques to find usage pattern from internet data. Internet usage mining consists of 3 phases particularly preprocessing, pattern discovery and pattern analysis. We tend to take the only session containing only 1 question is introduced, that distinguishes from the standard session. Meanwhile, the user session during this project is predicated on one session, though it is often extended to the entire session. The planned user session consists of each clicked and unclicked URLs and ends with the last URL that was clicked in a very single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user sessions. We Process the Analysis through given procedure:

•Individual System Web Log User Interests Extracting.

•Multiple Systems or Online Web Log User Interests Extracting.

Original Web Log Data or User Identification

This step focuses on separating the Web users from others. User Identification means identifying Unique users considering their IP address. Following heuristics are used to identify unique users:

(1) If there is a new IP address, then there is a new user.

(2) For more logs, if the IP address is the same, but the operating system or browsing software are different, a reasonable assumption is that each different agent type for an IP address represents a different user.

Existence of local caches, corporate firewalls and proxy servers greatly complicate user identification task. The WUM methods that rely on user cooperation are the easiest ways to deal with this problem. However, it is difficult because of security and privacy.

Session Identification

Visited pages during a user's navigation browsing should be divided into individual sessions. A session suggests that a collection of web content viewed by a specific user for a specific purpose. At present, the ways to spot user session embody timeout mechanism and supreme forward reference in the main. The following rules are used to identify a session:

(1) For any new IP address in Web log file, a new user and also a new session will be created.

(2) In one user session, if the refer page in an entry of Web log file is null, a new session will be considered.

(3) If the time between page requests is more than 25.5 or 30 minutes, it is assumed that the user is starting a new session.

Pattern Discovery and Classification

Pattern discovery is a phase which extracts the user behavioral patterns from the formatted data. For this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of this phase. In pattern discovery phase, several

data mining techniques are applied to obtain hidden patterns reflecting the typical behavior of users. Some important techniques for this phase are: path analysis, standard statistical analysis, clustering algorithms, association rules, classification algorithms, and sequential patterns. In the following, some of these techniques are described.

Classification

Classification is to make mechanically a model that may classify a group of pages. It is the task of mapping a page into one amongst many predefined categories. Within the internet domain, classification techniques permit one developing a profile of users that are happiness to explicit class or class and access particular server files. This needs extraction and choice of options that supported demographic data offered on these users, or supported their access patterns. This system has 2 steps. The primary step is predicated on the gathering of coaching data set and a model is made to explain the options of a group of information categories. During this step, information categories are predefined thus it is called supervised learning. Within the second step, the made model is employed to predict the categories of future information. For example, classification on server access logs may lead to the discovery of interesting patterns such as the following:

(1) Users from state or government agencies who visit the site tend to be interested in the page /company/lic.html.

(2) 60% of users, who placed an online order in /company/products /Music, were in the range of 18-25 years old and lived in Chandigarh.

Clustering

Clustering is another mining technique similar to classification, however, unlike classification there are no predefined classes, therefore, this technique is an unsupervised learning process. This technique is used to group together users or data items that have similar characteristics, so that members within the same cluster must be similar to some extent, also they should be dissimilar to those members in other clusters.

In the WUM domain, clustering techniques are mainly used to discover two kinds of interesting clusters: user clusters and page clusters. Clustering of users is to cluster users with similar preference, habits and behavioral patterns. Such knowledge is especially used for automated return mail to users falling within a certain cluster, or dynamically changing a particular site for a user, on a return visit, based on past classification of that user (provide personalized Web content to the users). On the other hand, clusters of Web pages contain pages that seem to be conceptually related according to the users' perception. The knowledge that is obtained from clustering in WUM is useful for performing market segmentation in ecommerce, designing adaptive Websites and designing recommender systems.

Frequent Patterns through FP Growth Algorithm

Many of vital page accesses are lost within the journal file as a result of the existence of native cache and proxy server. The task of path completion is to fill in these missing page references and makes bound wherever the request came from and what all pages are concerned within the path from the beginning until the top. We tend to are proposing FP-Growth algorithmic rule for net usage mining since no real time server out there.

CONCLUSION

Pre-processing the online log information may be an important and necessity innovate net mining. It removes irrelevant things and identifies users and sessions

together with the browsing data. The output of this part ends up in the creation of a user session file. The various patterns will be then discovered patterns by applying the mining techniques. The discovered patterns will then be used for numerous net usage applications like user identification, usage categorization, website improvement, business intelligence and proposals.

REFERENCES

1. Yan Wang. Web mining and knowledge discovery of usage patterns. CS 748T Project (Part I); 2000.
2. Sumathi, Padmaja Valli, Santhanam. An overview of preprocessing of web log files for web usage mining. *Journal of Theoretical and Applied Information Technology*. 2011; 34(2).
3. Qiang Yang. Building association-rule based sequential classifiers for web-document prediction. *Data Mining and Knowledge Discovery*. 2004; 8: 253–273p.
4. María J. Martín-Bautista, María-Amparo Vila, Víctor H. Escobar-Jeria. Obtaining user profiles via web usage mining. Iadis European Conference Data Mining; 2008.
5. K. R. Suneetha, Dr. R. Krishnamoorthi. Identifying user behavior by analyzing web server access log file. *IJCSNS International Journal of Computer Science and Network Security*. 2009; 9(4).
6. Sang T.T. Nguyen. Efficient web usage mining process for sequential patterns. Proceedings of Iiwas; 2009.
7. Saeed R. Aghabozorgi, Teh Ying Wah. Recommender systems: Incremental clustering on weblog data. *ICIS*. 2009.
8. Karuna P Joshi, Anupam Joshi, Yelena Yesha, Raghu Krishnapuram. Warehousing and mining web logs. *Workshop on Web Information and Data Management*; 1999.
9. J. Vellingiri, S. Chenthur Pandian. A novel technique for web log mining with better data cleaning and transaction identification. *Journal of Computer Science*. 2011; 7(5): 683–689p.
10. V.V.R. Maheswara Rao, Dr. V. Valli Kumari. An enhanced pre- processing research framework for web log data using a learning algorithm. *Netcom*. 2010; 01–15p.
11. Robert Walker Cooley. Web usage mining: Discovery and application of interesting patterns from web data; 2000.
12. A. Anitha. A new web usage mining approach for next page access prediction. *International Journal of Computer Applications*. 2010 8(11).
13. Liping Sun, Xiuzhen Zhang. Efficient frequent pattern mining on web log data; 2011.