Evolving Trends and its Application in Web Usage Mining: A Survey

Vivek Dogne, Anurag Jain, Susheel Jain

Abstract: With the abundance of information available on the World Wide Web (WWW), the issue of how to extract useful knowledge from the Web has gained significant attention among researchers in data mining and knowledge discovery areas. Web mining is applied to reflect the importance of Webpages and to predict the web domain visits ofvarious users. This article provides a survey of the available literature on Web usage mining and reviews the research and application issues in web usage mining

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

I. INTRODUCTION

Web mining refers to the effort 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 huge amount of information available from the web. It is often categorized into three major areas. [1, 2]

II. WEB CONTENT MINING

Web content mining is the procedure 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.

Manuscript Received on January 2015.

Vivek Dogne, Department of Computer Science, Radharaman Institute of Technology and Science, Bhopal (M.P.), India.

Anurag Jain, Department of Computer Science, Radharaman Institute of Technology and Science, Bhopal (M.P.), India.

Susheel Jain, Department of Computer Science, Radharaman Institute of Technology and Science, Bhopal (M.P.), India.

Hyperlinks A hyperlink is a structural element that connects a location in a web page to a different location, alsoinside 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.

Document 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 usersalong 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, pageposition and access time.

Application Server Data: Commercial application servers such since Weblogic, StoryServer have significantfeatures to allow E-commerce applications to be build on top of themby little effort. A key characteristic is the ability to path various kinds of businessevents 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 be renowned, however, that various end applications need a combination of one or more of the techniques functional in the above the 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 datais introduced, outlining thedetails related to web log data preprocessing. Techniques on the web usage mining are discussed in section

4. Finally section 5 concludes thepaper.

V. LITERATURE REVIEW

S. park 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.[3].

X. Zhang et al, describes a toolset thatexploits 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 beentrained 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 usergroup, and recommend a unique set of product browsing options appropriate to an individual user.[4]

Z. Li et al, present a novel ontology basedWeb usage miningframeworkthat leverages search engine queries to improve the accuracy of unemployment rate prediction. The proposedFramework is underpinned by a domain ontology which captures unemployment related concepts andtheir semantic relationships to facilitate the extraction of useful prediction features from relevant search enginequeries. In addition, state-of-the-art feature selection methods and data mining models such as neural networksand support vector regressions are exploited to enhance the effectiveness of unemployment rate prediction. [5]

M. Belk et al, focuses on modeling users'cognitive styles based on a put of Web usage mining techniqueson client navigation patterns and clickstream data. Main aim is to inspect whether exact clusteringtechniques can group user of particular cognitive style bymeasures obtained from psychometric testand content navigation behavior. [6]

M. Wu. et al, proposes an approach based on web mining to analyze product usability. This approach uses the massive online customer reviews onanalogous products and features as data source, which are easy to get from Web and can reflect the mostupdated customer opinions on product usability. Association rule mining techniques are adopted toextract customer opinions on the usability of product features. [7]

S.G.Matthewsetal, presentedgenetic algorithm (GA)-based solution isdescribed that uses the elastic nature of the 2-tuple linguistic illustration to discover rules that occurat the intersection of fuzzy set borders. The GA-based advance is enhanced from previous work byincluding a graph illustration and anbetter fitness function. [8]

Y. T. Wang et al, introduced the concept of throughout-surfing patterns (TSP) and then present an competent method for mining the patterns. Authors propose a compact graph structure,term a path traversal chart, to record information about the navigationpaths of website visitors. The graph contains the frequent surfing paths that are required for miningTSPs.[9]

X. Wang et al, propose a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. We made use ofthe cluster information generate by a self organizingdiagram for pattern analysis and a fuzzy inferencesystemto capture the chaotic movement to provide short-term (hourly) and long-term (daily) Web trafficmovementpredictions.[10]

G. Castellano et al, proposed NEWER(NEuro-fuzzy Web Recommendation), a usage-based Webadvice 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 users distribution similar interests and to determine a recommendation model as a set of fuzzy rules express the associations between user category andrelevancesof pages.[11]

C. C. Aggarwal et al, designed an algorithm which combine classical partition algorithms among probabilistic models in order to produce an effectiveclustering approach. Then show how to enlarge the approach to the categorization problem. [12]

VI. PREPROCESSING STEPS OF LOG DATA

One objective of web usage mining is to extract sequentialusage patterns from a large collection of web logs [13]. 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 [14]. Five preprocessing steps have been identified [15]:

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 necessary recorded in consecutive order rather they could be separated by records from other users.
3. Session Identification: To divide pages accessed byeach user into individual sessions. A session is a sequence of pages visisted 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.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

5. Formatting: Format the data to be readable by data mining systems.

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.[16, 17, 18].

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:

Predict the association and relationship among set of items "wherever the presence of one set of objects in a operation implies (with a certain degree of assurance) the presence of extraitems[20].

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 pattern discovery:

It is applied to web access server transaction logs. The purpose is to determine sequential patterns that specify user visit patterns over aassured 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 [22].

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 [23].

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 [24].

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 todesign algorithm, which can help to better understand the mined knowledge.

REFERENCES

[1] R. Kosala, H. Blockeel, Web mining research: a survey, ACM SIGKDD Explorations Newsletter 2 (1) (2000)pp, 1–15.

[2] F.M. Facca, P.L. Lanzi, Mining interesting knowledge from weblogs: a survey, Data and Knowledge Engineering 53 (3) (2005)pp, 225–241.

[3] 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 Engineering 65.3 (2008)pp, 512-543.

[4] Zhang, Xuejun, John Edwards, and Jenny Harding. "Personalised online sales using web usage data mining." Computers in Industry 58.8 (2007)pp, 772-782.

[5] Li, Ziang, et al. "An ontology-based Web mining method for unemployment rate prediction." Decision Support Systems 66 (2014) pp,114-122.

[6] Belk, Marios, et al. "Modeling users on the World Wide Web based on cognitive factors, navigation behavior and clustering techniques." Journal of Systems and Software 86.12 (2013) pp, 2995-3012.

[7] Wu, Mingxing, et al. "An approach of product usability evaluation based on Web mining in feature fatigue analysis." Computers & Industrial Engineering 75 (2014) pp, 230-238.

[8] Matthews, Stephen G. et al. "Web usage mining with evolutionary extraction of temporal fuzzy association rules." Knowledge-BasedSystems 54 (2013) pp, 66-72.

[9] Wang, Yao-Te, and Anthony JT Lee. "Mining Web navigation patterns with a pathtraversal graph." Expert Systems with Applications 38.6 (2011) pp,7112-7122.

[10] Wang, Xiaozhe, Ajith Abraham, and Kate A. Smith."Intelligent web traffic mining and analysis."Journal of Network and ComputerApplications28.2 (2005) pp, 147-165.

[11] Castellano, Giovanna, Anna Maria Fanelli, and Maria AlessandraTorsello. "NEWER: A system for NEuro-fuzzy WEb Recommendation." Applied Soft Computing 11.1(2011) pp,793-806.

[12] Aggarwal, C., Yuchen Zhao, and P. Yu. "On the use of Side Information for Mining Text Data." (2012) pp, 1-1.

[13] Cooley, R., P.-N. Tan, and J. Srivastava, "Discovery of intersting usage patterns from Web data," presented at WEBKDD, (1999) pp,5-32.

[14] Kohavi, R., "Mining e-commerce data: The good, the bad, and the ugly," presented at 7th ACM SIGKDD International Conference on Knowledge Discovery, San Francisco, California,(2001) pp,8-13.

[15] Cooley, R., B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, vol. 1, (1999)pp, 5-32.
[16] Fu, X., J. Budzik, and K. J. Hammond, "Mining navigation history for recommendation," In Proceedings of the 2000 Conference on Intelligent User Interfaces, (2000) pp,106-112.
[17] Pazzani, M., J. Muramatsu, and D. Billsus.Syskill&Webert, "Identifying interesting Web sites," In Proceedings of the Thirteenth National Conference on Artificial Intelligence, (1996), pp, 54-61.
[18] Mobasher , B., R. Cooley, and J. Srivastava,"Creating adaptive web sites through usage-based clustering of URLs," In Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), (1999) pp, 19-25.
[19] Cooley, R.; Mobasher, B.; Srivastava, J.; "Web mining: information and pattern discovery on the World Wide Web".In Proceedings ofNinth IEEEInternational Conference., 3-8 Nov. (1997)pp, 558 – 567.
[20] Peng, Huiping. "Discovery of interesting association rules based on web usage mining." Multimedia Communications (Mediacom), 2010 International Conferenceon.IEEE, (2010) pp, 272-275.
[21] Masseglia, Florent, DoruTanasa, and Brigitte Trousse. "Web usage mining: Sequential pattern extraction with a very low support." Advanced Web Technologies andApplications.Springer Berlin Heidelberg, (2004) pp,513-522.
[22] Varghese, NayanaMariya, and Jomina John. "Cluster optimization for enhanced web usage mining using fuzzy logic." Information andCommunication Technologies (WICT), 2012World Congress on.IEEE, (2012) pp,948-952.
[23] Raghavendra, Prakash S., Shreya Roy Chowdhury, and SrilekhaVedulaKameswari. "Comparative study of neural networks and k-means classification in web usage mining." Internet Technology and Secured Transactions (ICITST), 2010 International Conference for.IEEE, (2010) pp,1-7.
[24] Li, Yan, BoqinFeng, and Qinjiao Mao. "Research on path completion technique in web usage mining." Computer Science and Computational Technology, 2008.ISCSCT'08.International Symposium*on*.Vol.1.IEEE,(2008) pp,554-559.