Research on a Hybrid Approach for Web Usage Mining

Asmita Patil[1] Jagruti Kadam[2] Pratiksha Bharmal[3] Jagruti Babaria[4]
[1,2,3,4]Student
[1,2,3,4]Department of Information Technology
[1,2,3,4]Atharva College of Engineering, Mumbai, Maharashtra

Abstract— With the large number of companies using the Internet to distribute and collect information, knowledge discovery on the web or web mining has become an important research area. Basically data mining techniques are used in web mining. Web mining is extended version of data mining. Data mining is work upon Off Line whereas Web mining is work upon On-Line. In data mining data is stored in (database) data warehouse and in web mining data is stored in server database & web log. The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. Web mining can be divided into three areas, namely web content mining, web structure mining and web usage mining (also called web log mining).Web content mining focuses on discovery of information stored on the Internet, i.e., the various search engines. 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 association patterns, clustering of web documents and classification 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. Web structure mining can be used when improving the structural design of a website.
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.

Key words: distribute and collect information, Knowledge Discovery,

I. INTRODUCTION

Web content mining is the process of extracting knowledge from web documents such as text and multimedia. Knowledge extraction from the structure of web and hyperlink references is called web structure mining. Web usage mining is the process of knowledge exploitation from the secondary data [1].

Web usage mining is a type of web mining, which exploits data mining techniques to discover valuable information from navigations of Web users. Web usage mining tries to make sense of the data generated by the Web surfer's sessions or behaviors. While the Web content and structure mining utilize the real or primary data on the Web, Web usage mining mines the secondary data derived from the interactions of the users while interacting with the Web. The Web usage includes the data from Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions, transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of the interactions.

Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship between different Web sites. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs) [4].

	Web mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	Db View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	-Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Serves Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing -User Modeling

Fig. 1: Comparison for Web Mining

II. LITERATURE SURVEY

Web Mining – It is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

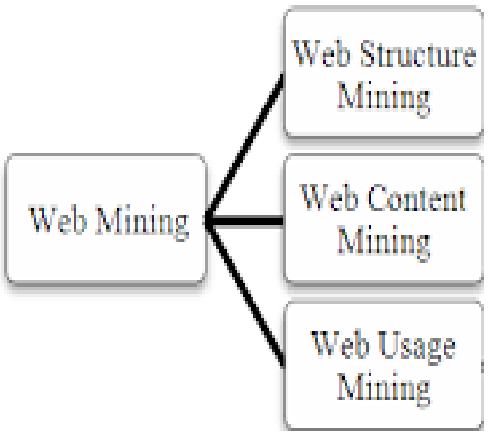

Fig. 2: Web Mining

Web Content Mining (Analyse the content of web pages as well as results of web Searching) Web content mining is a process of extracting up information from texts, images and other contents. The technologies that are mainly used in web content mining are NLP (Natural language processing) and IR (Information retrieval). Web Structure Mining (Hyperlink Structure) Web structure mining is a process of extracting up information from linkages of web pages. Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site. This graph structure can provide information about ranking and enhance search results of a page through filtering. Web Usage mining (analysing user web navigation) Web usage mining is a process of extracting information from user how to navigate web sites. Web usage mining also known as web log mining, aims to discover interesting and frequent user access patterns from web browsing data that are stored in web server logs, proxy server logs or browser logs.[14]

A. Pre Processing

Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery [5].

As said in [12], pre-processing "consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery". This step can break into at least four sub steps: Data Cleaning, User Identification, Session Identification and Formatting. Unneeded data will be deleted from raw data in web log files in the data cleaning step. When a user requests a page, the request is added to the Log File; but if this page contains images, java scripts, flash animations, video, etc., they are added to the Log file as well. Most of the time, these are not needed for pattern discovery and should be omitted from log files [11].

B. Pattern Discovery

Pattern discovery is the discovery of frequently occurring ordered events or subsequences as patterns.

An example of it is "Customers who buy a Canon digital camera are likely to buy an HP color printer within a month."There are several methods and techniques for pattern discovery. They are:

1) Clustering: A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.
2) Classification: Classification is the technique to map a data item into one of several predefined classes [2]. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines etc [5].

C. Pattern Analysis

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the Interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website[6].

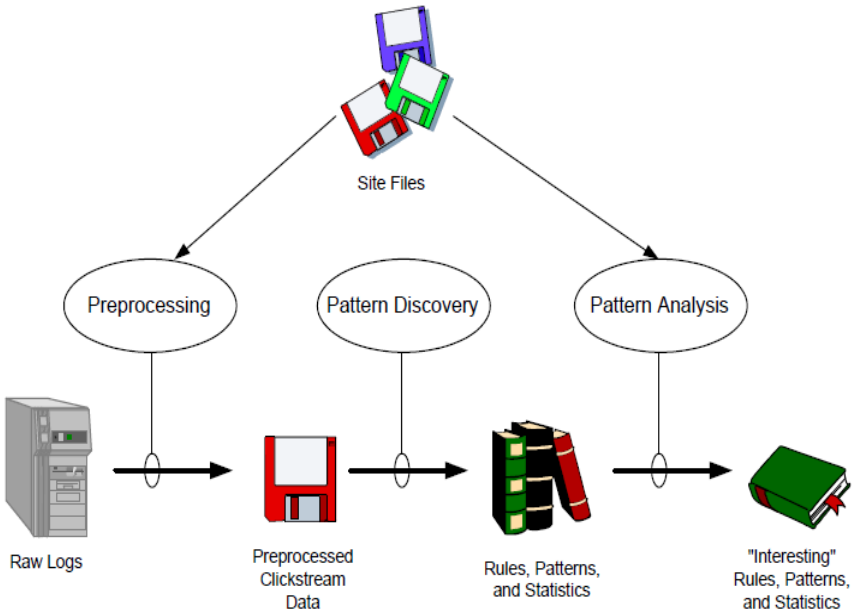

Fig. 2: Depicts three main tasks for Web Usage Mining process

D. Ant Colony Optimization

The complex social behaviors of ants have been much studied by science, and computer scientists are now finding that these behavior patterns can provide models for solving difficult combinatorial optimization problems. The attempt to develop algorithms inspired by one aspect of ant behavior, the ability to find what computer scientists would call shortest paths, has become the field of ant colony optimization (ACO), the most successful and widely recognized algorithmic technique based on ant behavior[3].

E. Ant Based Clustering

Deneubourg et al. in [7] proposed ant-based clustering and sorting. In the case of ant-based clustering and sorting, two related types of natural ant behaviors are modeled. When clustering, ants gather items to form heaps. And when sorting, ants discriminate between different kinds of items and spatially arrange them according to their properties [8]. Lumer and Faieta[9]. in proposed ant-based data clustering algorithm, which resembles the ant behavior described in [7]. The agents (ants) and data are randomly initialized on a toroidal grid. By moving agents, data is sorted according to its neighbors. The picking and dropping probabilities, given a grid position and a particular data item i, are computed using the density functions. Handl & Meyer in [10] proposed an extension of this algorithm where the parameter _ is adaptively updated during the execution of the algorithm. We applied Handl & Meyer's Ant-based clustering algorithm for detecting user's patterns [11].

III. CONCLUSION

In this paper we survey the area on the web mining. From this survey we can say that A hybrid approach of ant based clustering algorithm with lumer faieta can be used to remove the traffic from the weblogs.

REFERENCES

[1] R. Cooley, Web Usage Mining: Discovery and Application of Interesting patterns from Web Data, Ph. D. Thesis, University of Minnesota, Department of Computer Science, 2000.
[2] Kobra Etminani, Mohammad-R. Akbarzadeh-T. 2, Noorali Raeeji Yanehsari, Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method, IFSA-EUSFLAT, 2009.
[3] Marco Dorigo and Thomas Stützle, in "Ant colony Optimization," June 2009.
[4] Raymond Kosala, Hendrik Blockeel. Web Mining Research: A Survey. Volume 2, Issue 1-page 4, ACM SIGKDD, July 2000.
[5] Jaideep Srivastava , Robert Cooley, Mukund Deshpande, PangNing Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Volume 1, Issue 2-page 5, ACM SIGKDD, Jan 2000.
[6] Rajni Pamnani, Pramila Chawan, Web Usage Mining: A Research Area in Web Mining.
[7] J. Deneubourg -L., S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chrétien, The dynamics of collective sorting: robot-like ants and ant-like robots. Proceeding of the first international conference on simulation of adaptive behavior, pp. 356–365, MIT Press, 1991.
[8] J. Handl, B. Meyer, Ant-based and Swarm-based clustering, Swarm Intelligence, 1, pp. 95–113, 2007.
[9] E. Lumer, B. Faieta, Diversity and adaptation in populations of clustering ants. Proceeding of the third international conference on simulation of adaptive behavior, pp. 501–508, MIT Press, 1994.
[10] J. Handl, B. Meyer, Improved ant-based clustering and sorting in document retrieval interface. Proceeding of the Seventh International Conference on Parallel Problem Solving from Nature, Vol. 2439 of Lecture Notes in Computer Science, pp. 913-923, Germany: Springer-Verlag, 2002.
[11] Kobra Etminani , Mohammad-R. Akbarzadeh-T. 2, Noorali Raeeji Yanehsari, Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method, ISBN: 978-989-95079-6-8, IFSA-EUSFLAT 2009.
[12] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, 1(2), pp. 12-23, 2000.
[13] Christian Blum, Ant colony optimization: Introduction and recent trends, Physics of Life Reviews 2 (2005) 353–373.
[14] Aparna N. Gupta , Prof. Arti Karndik ar, "A Review: Study of Various Clustering Techniques in Web Usage Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014.