A new approach of web directory creation and personalization

Asst. Prof. Madhavi S. Darokar
Department of Computer Enginnering
(JSPM's Imperial College of Engineering & Research, Pune.)

Abstract: Now a day's huge amount of data is available on the internet. Every day the size of this data is increasing. So the accessing this big data is a challenging task in today's era. It causes information overload on the internet. A novel Web directory creation is the solution to this problem. Here we are applying personalization to the web directory and creating our own web directory by clustering process then applying probabilistic theorem. This will help to reduce the narrowing the user search and display of unwanted data contents on the web.

Keywords: *Web directory, Internet service Provider, clustering, Open Directory Project, Clustering, Machine Learning, Personalization, Web Usage Mining.*

I. INTRODUCTION

The web has tremendous amount of information. The number of user accessing the World Wide Web is increasing every day. So to retrieve perfect required information online on the web is a cumbersome task. Personalization of web is the solution to this problem, which uses web mining technique to overcome the problem. So here we are creating web directory. As per user interest the web pages they have searched under specific theme is organized hierarchically generally known as web directory. Here user can get the interested information by searching inside the directory from broad category and gradually narrowing down until they get the thematic contents .So the user has to check deep inside the directory until they get satisfied information on the web. The clustering and probabilistic approaches are used for the pattern analysis to build user community specific personalized web directory [5]. As the web data has high rate of thematic diversity (increased dimensionality and semantic incoherence), we are creating a knowledge discovery framework for construction of community web directory by applying personalization to web directory which will become automatic machine learning process [3].

The explosive amount of information available on the web, it become the important platform to add, update and retrieve the information and mine the web for important knowledge But the web data is large, non-structured, scattered and dynamic so there is need of research to overcome these challenges[1]. So it is mandatory for web developers to solve the issue of overloading of information. This can be achieved by knowing what the web user really want, what is their interest, what is the web searching pattern and customize his interest by pattern learning and mining it to get knowledge .

To make web based information available as per the need and interest of each user or multiple user we can use the novel method of web personalization and to achieve the same we have to create an correct practical model to tackle the above deficiencies of web that is "information overload"[2]. We are creating here usable web directories of same interest of user communities. The web contents are organized into thematic hierarchies called as Web Directories which are constructed by human manually.

The web personalization is done by the web usages mining which automatically stores frequently access of web patterns from history stored in the web log files. Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files [2].

The Web has not achieved its goal of providing easy access to online information. As its size is increasing, the abundance of available information on the Web causes the frustrating phenomenon of information overload to Web users. Organization of the Web content into

ISSN: 2278 – 7798

thematic hierarchies is an attempt to alleviate the problem. These hierarchies are known as Web Directories and correspond to listings of topics which are organized and overseen by humans. A Web directory, such as Yahoo (www.yahoo.com) and the Open Directory Project (ODP) like (dmoz.org), allows users to find Web sites related to the topic they are interested in, by starting with broad categories and gradually narrowing down, choosing the category most related to their interests [3].

II . PAST WORK

The many research work has been carried out on the Web Personalization which is method of making Web-based information systems created as per the needs and interests of individual users, or number of users, to tackle information overload. However, to achieve the web personalization, we need to construct the accurate and practical user models, but construction and maintaining of these models are difficult [6][9].

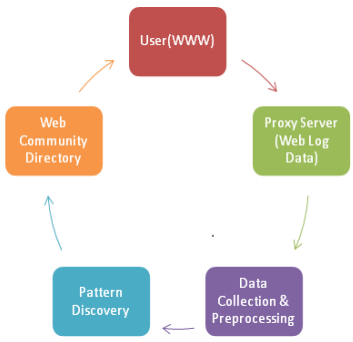
III. PRESENT WORK

A Personalized Web directory is constructed according to the needs and interests of particular group of user communities. Web directory is created using clustering algorithm on the web log data found on the proxy server generally called web usages data. The web directory is personalized by the level of user interest analyzed in usages data. Furthermore, it presents the complete novel method for the construction of such hierarchical web directories by using web usage data. . We tested this methodology to the ODP directory using access logs from the proxy servers of an Internet Service Provider(ISP) and provided results indicates the usability of the Personalized Web directories. Here we approximated the gain and losses of the end user. The web directory is also evaluated according to the user preferences.

IV. PROPOSED METHODOLOGY

The main theme of the paper is creating the Web directories according to the preferences and interest of user communities. A Web directory, such as Yahoo and the Open Directory Project (ODP) (dmoz.org), users can find any topic that they are interested by searching the websites, starting with large categories and gradually moving down and choosing the category related to their interests. Searching for a particular interested topic, user has to check deep inside the directory. Hence, the Web directory accessing complexity is increased due to the size and information overload problem will occur. To overcome this problem of web directories, we research on the creation of community web directories. The usages data is collected from ISP and personalized web directory is created. User communities are created using data collected from the log files of Internet Service Provider. The goal is creating personalized web directories based on web usage data. The users of a community can use this personalized directory as a entry point for accessing the Web, based on the topics that they are interested in, instead of accessing available vast Web directories. There are three methods of web usages mining to achieve the goal.

1. Collection of Web data such as web browsing history recorded in Web server logs.
2. Preprocessing of Web data such as filtering user's requests, requests to images and multimedia, and identifying unique sessions, Web data analysis also known as Web Usage Mining [3], to discover user's interesting in specific patterns or profiles, and Web usage mining can use various data mining or machine learning techniques.
3. Discovery of community web directory from user's navigational pattern.
4. Evaluation of the discovered web directory for personalization. The complete process is shown in the below diagram.



ISSN: 2278 – 7798

V. ALGORITHMS

We combine the clustering algorithm and probabilistic approaches presented in previous work and also present a new algorithm that combines these two approaches[24]. The resulting community model is the Personalized Web Directory. The proposed personalization methodology is tested and evaluated on a ODP and a general-purpose Web directory, indicating it's importance to the Web user. The processing steps for discovery of the web directory are as follows.

A. Collection of Web Usage Data:

To identify user's interest in the web the user communities are formed using data collected from Web proxies as users search the Web and construct community Web directories based on those patterns[15]. Web Usage Data Preparation consists of the collection and cleaning of the usage raw data, as well as the identification of user sessions, removing graphics and multimedia files. Then to assemble these data into a consistent, integrated, and concise view. A user session is consider as a sequence of log entries and the access to Web pages by the same IP address, where the time interval between two subsequent entries should not exceed a certain time period.

B. Initialization of Web Directory:

It provides the categorization of the Web pages included in the usage data, according to the categories of a Web directory. We compare two different approaches for the characterization of the Web pages. The first approach organizes Web pages into an artificial Web directory using K-means clustering [16]. The second approach classifies them onto an existing Web directory, like ODP.

C. Discovery of Community Web Directory:

It is the main process of discovering the user practical models from web log data, using machine learning techniques [15]. We are using here unsupervised learning to discover the interested pattern of thematic user session by coordinating the association between user session ,categories of web directory and searched web pages and exploiting these novel models to build the Personalized Web directories. The creation of Personalized Web directories is a fully automated process, resulting into operational personalization knowledge, in the form of user new models. A new algorithm K-means clustering and probabilistic model PLSA is combined here to get the result [4][23] The Algorithmic mathematical steps for constructing web directory

VI. MATHEMATICAL MODEL

Input: Web log data provided from the Internet Service Provider server. Then performing data cleaning.
Step1: Loading clean ISP log file removing images and multimedia content etc.
Step2: Extract information from the log file like session from IP address. While extracting information from log file, we extract sessions using below formula [10]:
Let P = (p1, p2 . . . , pu) is the sequence of Web pages accessed from a certain IP between t1 and tu. Then, a user session v (t1, tf), tf δ tu, is defined as : v(t1,tf)=(p1,p2, . . . ,pf) :4 =tj tj-1 δ ,1 < j δ f) \(f = u[tf+1 δ tf > δ), where δ is a predefined time threshold.
Step3: Extract websites from log file. Then extract web pages from websites using below formula[4]:

$$Sim(pi, vi) = \sum_{k=1}^{k=m} WikQik \Big/ \sqrt{\sum_{k=1}^{k=m} W^2 ik}\ \sqrt{\sum_{k=1}^{k=m} W^2 ik}$$

Step4: Extract directories of the website .
Step5: Cluster community specific directories based on training the data set into various clusters. If cluster is not found then cluster is denoted as "unknown".

ISSN: 2278 – 7798

VII. RESULTS

The K-means clustering algorithm for calculating edge instances is shown in following diagrams.

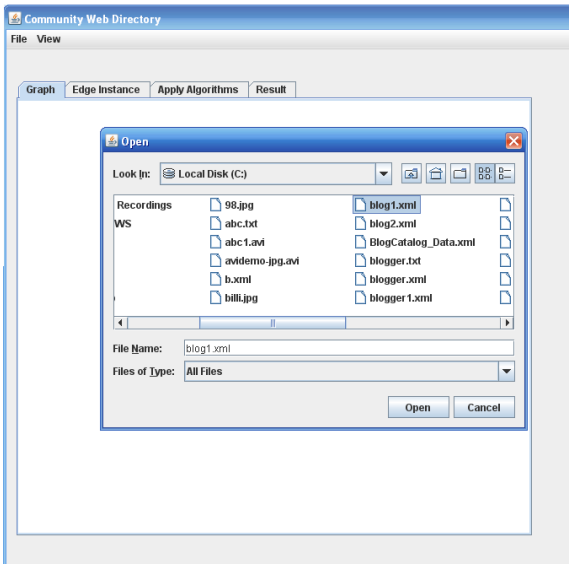


Fig 2: Input XML file of user interest while visiting blog.

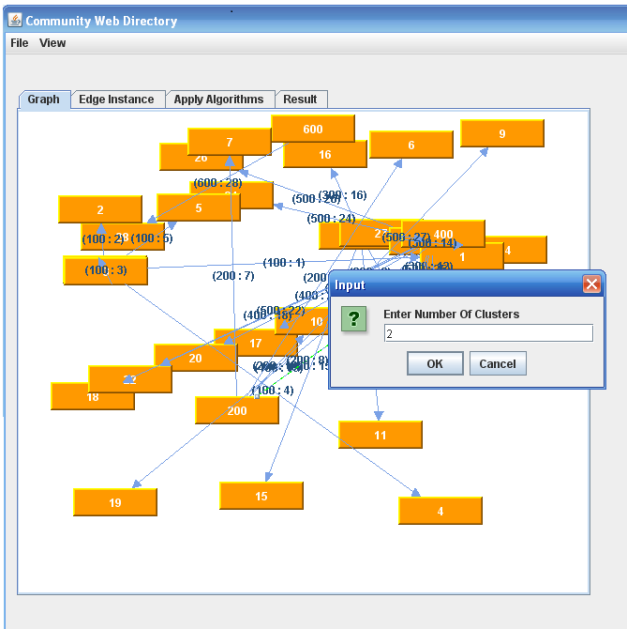


Fig 3: Creation of number of Cluster as per user interest .

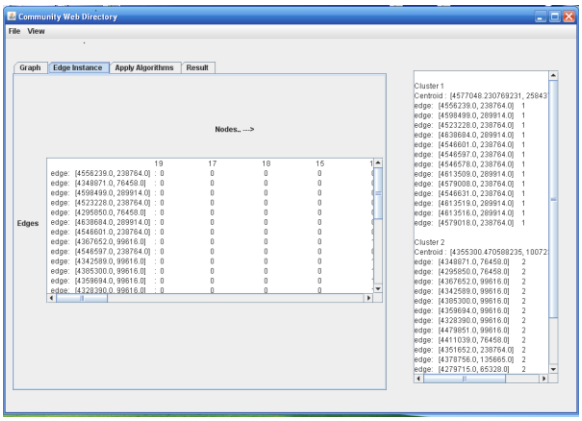


Fig 4: Calculating Edges and Nodes and created Cluster of user interest shown in right panel.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This paper elaborates the concept of a Personalized Web directory, as a Web directory that satisfies the needs and interests of particular user communities. Furthermore, it presents the complete methodology for the creation of such web directories with the aid of novel machine learning methods. User community models are considered as the form of thematic hierarchies and are constructed by combining clustering and probabilistic learning approaches. We applied our methodology to the ODP directory. The proposed work reduces the information overload problem by reducing the dimensionality by the categorization of web pages into the correct category of web directory. We have not only showed the result of user session gain but the losses encountered during browsing the web.

This research requires the evaluation of community Web directories in user practical studies which are in our next plans for future work. The proposed methodology provides a promising research direction, where new concepts of web mining are in demand.

IX. REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, and P.T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.

ISSN: 2278 – 7798

[2] D.Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," User Modeling and User-Adapted Interaction, vol. 13, no.4, pp. 311-372, 2003.

[3] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, and M. Dikaiakos, "WebCommunity Directories: A New Approach to Web Personalization," Web Mining: From Web toSemantic Web, B. Berendt et al., eds., pp. 113-129, Springer, 2004.

[4] D. Pierrakos and G. Paliouras, "Exploiting Probabilistic Latent Information for the Constructionof Community Web Directories," Proc. 10th Int'l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 89-98, 2005.

[5] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.

[6] The Adaptive Web, Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Neijdl, eds. Springer, 2007.

[7] T. Hofmann, "Learning What People (Don't) Want," Proc. 12th European Conf. in Machine Learning, pp. 214-225, 2001.

[8] G. Xu, Y. Zhang, and Y. Xun, "Modeling User Behaviour for Web Recommendation Using ldaModel," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence and Intelligent Agent Technology,pp. 529-532, 2008.

[9] W. Chu and S.-T.P. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," Proc. 18th Int'l Conf. World Wide Web (WWW), pp. 691-700, 2009.

[10] X. Jin, Y. Zhou, and B. Mobasher, "Task-Oriented Web User Modeling for Recommendation,"Proc. 10th10th Int'l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp.109118,2005.Advanced Computing: An International Journal (ACIJ), Vol.3, No.2, March 201248

[11] D. Chen, D. Wang, and F. Yu, "A PLSA-Based Approach for Building User Profile and Implementing Personalized Recommendation,"-Proc. Joint Ninth Asia-Pacific Web Conf. (APWeb '07) and Eighth Int'l Conf. Web-Age Information Management (WAIM '07), pp. 606-613, 2007.

[12] B. Mehta and N. Wolfgang, "Unsupervised Strategies for Shilling Detection and Robust Collaborative Filtering," User Modeling and User-Adapted Interaction, vol. 19, nos. 1/2, pp. 65-97, 2009.

[13] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using odp Metadata to PersonalizeSearch," Proc. 28th Ann. Int'lACM SIGIR Conf. Research and Development in InformationRetrieval, pp. 178-185, 2005.

[14] A. Sieg, B. Mobasher, and R. Burke, "Ontological User Profiles for Representing Context inWeb Search," Proc. IEEE/WIC/ACM Int'lConf. Web Intelligence and Intelligent AgentTechnology Workshops, pp. 91-94, 2007.

[15] Z. Ma, G. Pant, and O.R.L. Sheng, "Interest-Based Personalized Search," ACM Trans.Information Systems, vol. 25, no. 1, article no. 5, Feb. 2007.

[16] C.R. Anderson and E. Horvitz, "Web Montage: A Dynamic Personalized Start Page," Proc. 11thInt'l Conf. World Wide Web, pp. 704-712, May 2002.

[17] Chen, H. and S. T. Dumais. Bringing order to the web: automatically categorizing search results.In Proceedings of CHI'00, Human Factors in Computing Systems, 145-152, 2000

[18] Kamdar, T., and A. Joshi. On Creating Adaptive Web Sites using WebLog Mining. Technical Report TR-CS-00-05. Department of Computer Science and Electrical Engineering University of Maryland, Baltimore County, 2000

[19] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI '98), pp 43-52, 1998

[20] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, 1981.

[21] J. Hartigan, Clustering Algorithms. John Wiley & Sons, 1975.

[22] C. Bron and J. Kerbosch, "Algorithm 457-Finding All Cliques of an Undirected Graph," Comm. ACM, vol. 16, no. 9, pp. 575-577, 1973.

[23] T. Hofmann, "Probabilistic Latent Semantic Analysis," Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI' 99), pp. 289-296, 1999.

[24]D. Pierrakos, Georgios Paliours, "*Personalizing web directories with the Aid of Web Usages Data"* ,IEEETransactions on Knowledge and Data Engineering, vol.22,no.9,Sep 2010.

ISSN: 2278 – 7798