Optimized web personalization -implementation using web mining techniques

R.LAVANYA
PG Student
Sri Manakula Vinayagar Engineering College
Pondicherry
lavanya27382@gmail.com

I.SASIKALA
PG Student
Sri Manakula Vinayagar Engineering College
Pondicherry
sskl529@gmail.com

M.GOWRI
Associate Professor
Sri Manakula Vinayagar Engineering College
Pondicherry
kani0138@gmail.com

ABSTRACT
In recent years there has been a growing interest in automatic and semi-automatic harvesting and mining of data from Web pages. Web mining is the application of data mining techniques to extract knowledge from Web. In Web mining, different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. The Web mining is a converging research area from several research communities, such as Databases, Information Retrieval and Artificial Intelligence. In this paper, we highlight the significance of studying the evolving nature of the Web personalization and its implementation using web mining techniques.

Keywords
Structure Mining, Usage Mining, Pattern Analysis, Content Mining.

1. INTRODUCTION

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the properties of the huge, diverse, dynamic and unstructured nature of Web data, Web data research has encountered a lot of challenges, such as scalability, multimedia and temporal issues etc. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the web. A user interacts with the Web, there is a wide diversity of user's navigational preference, which results in needing different contents and presentations of information. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for a Web developer or designer to know what the user really wants to do, predict which pages the user is potentially interested in, and present the customized Web pages to the user by learning user navigational pattern knowledge [1,2,3].

2. WEB MINING

Web mining 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.

3. WEB MINING TECHNIQUES

3.1 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).

3.2 Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Thus, Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level.

3.3 Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Some of the typical usage data collected at a Web site include IP addresses, page references, and access time of the users.

3.4 Text Mining

Due to the continuous growth of the volumes of text data, automatic extraction of implicit previously unknown and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining, therefore, corresponds to extension of the data mining approach to textual data

and its concerned with various tasks, such as extraction of information implicitly contained in collection of documents or similarity- based structuring. Text collection in general, lacks the imposed structure of a traditional database. The text expresses the vast range of information, but encodes the information in a form that is difficult to decipher automatically.

4. WEB DATA

Web data are those that can be collected and used in the context of Web personalization. These data are classified in four categories according to [6]:

Content data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.

Structure data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.

Usage data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.

User profile data provide information about the users of a Web site. A user profile contains demographic information for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

5. PERSONALIZATON ON THE WEB

Web personalization is a strategy, a marketing tool, and an art. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate what information you present to your users and how you present it [7,12]. Correctly executed, personalization of the visitor's experience makes his time on your site, or in your application, more productive and engaging. Personalization can also be valuable to you and your organization, because it drives desired business results such as increasing visitor response or promoting customer retention. Unfortunately, personalization for its own sake has the potential to increase the complexity of your site interface and drive inefficiency into your architecture. It might even compromise the effectiveness of your marketing message or, worse, impair the user's experience. Few businesses are willing to sacrifice their core message for the sake of a few trick web pages. Contrary to popular belief, personalization doesn't have to take the form of customized content portals, popularized in the mid-to-late 90s by snap.com and My Yahoo!. Nor does personalization require expensive applications or live-in consultants. Personalization can be as blatant or as understated as you want it to be. It's a tired old yarn, but if you hope to implement a web personalization strategy, the first and most important step is to develop and mature your business goals and requirements [4,5]. It is important to detail what it is you hope to do and, from that knowledge, develop an understanding of how you get from an idea to implementation. You might be surprised to discover that it won't require most of next year's budget to achieve worthwhile results. Web personalization can be seen as an inter-disciplinary field that includes several research domains from user modeling [7,12], social networks [14], web data mining [8,13], human-machine interactions to Web usage mining[13]; Web usage mining is an example of approach to extract log files containing information on user navigation in order to classify users. Other techniques of information retrieval are based on documents categories' selection [13]. Contextual information extraction on the user and/or materials (for adaptation systems) is a technique fairly used also includes, in addition to user contextual information, contextual information of real-time interactions with the Web. [8] proposed a multi-agent system based on three layers: a user layer containing users' profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. [3] propose reformulation query by adding implicit user information. This helps to remove any ambiguity that may exist in query: when a user asks for the term "conception", the query should be different if he is an architect or a computer science designer. Requests can also be enriched with predefined terms derived from user's profile [8] develop a similar approach based on user categories and profiles inference. User profiles can be also used to enrich queries and to sort results at the user interface level [11]. Other approaches also consider social-based filtering [12] and collaborative filtering. These techniques are based on relationships inferred from users' profile. Implicit filtering is a method that observes user's behavior and activities in order to categorize classes of profile. Other approaches consider information semantics. For example, user queries can be enriched by adding new properties from the available domain ontologies [12]. [14] Assume that reading, scanning and interacting with a document considered as relevant takes much time for the user. They consider that three sources of implicit feedback are the most relevant to approximate user's interest for a given web page: reading time, scrolling over the same page and interacting with the system. The Web information retrieval and mining usually consider web pages as the element to be analyzed, organized and presented to the user. However, the content of these Web pages is complex and inter-related. This has lead to an interest on integrating semantic knowledge; Personalization process has been enriched at the semantic level, based

on user modeling and on log files analysis. These approaches can be combined. User modeling by ontology can be coupled with dynamic update of user profile using results of information-filtering and Web usage mining techniques.

6. PERSONALIZATION STRATEGIES

Personalization falls into four basic categories, ordered from the simplest to the most advanced: Memorization, Customization and Task Performance Support.

6.1 Memorization

In this simplest and most widespread form of personalization, user information such as name and browsing history is stored (e.g. using *cookies*), to be later used to recognize and greet the returning user. This mode depends more on Web technology than on any kind of adaptive or intelligent learning. It can also jeopardize user privacy.

6.2 Customization

This form of personalization takes as input a user's preferences from registration forms in order to customize the content and structure of a web page. This process tends to be static and manual or at best semi-automatic. It is usually implemented on the Web server. Typical examples include personalized web portals such as My Yahoo and Google.

6.3 Task Performance Support

In these client-side personalization systems, a personal assistant executes actions on behalf of the user, in order to facilitate access to relevant information. This approach requires heavy involvement on the part of the user, including access, installation, and maintenance of the personal assistant software. It also has very limited scope in the sense that it cannot use information about other users with similar interests. This approach requires heavy involvement on the part of user, including access, installation, and maintenance.

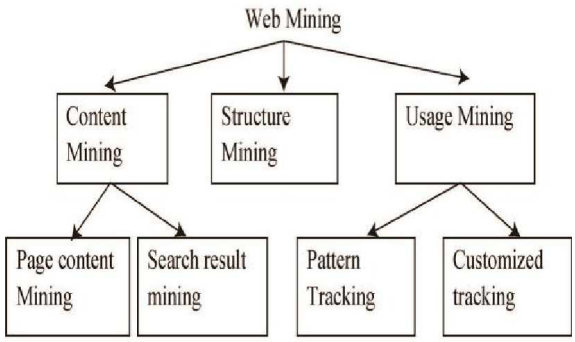

Fig 1: Web mining classification

7. REQUIREMENTS OF WEB USAGE MINING

It is necessary to examine what kind of features a Web usage mining system is expected to have in order to conduct effective and efficient Web usage mining, and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to: Gather useful usage data thoroughly, Filter out irrelevant usage data, Establish the actual usage data, Discover interesting navigation patterns, Display the navigation patterns clearly, Analyze and interpret the navigation patterns correctly, and Apply the mining results effectively.

8. CONCLUSION

In this paper, we have outlined three different modes of web mining, namely web content mining, web structure mining and web usage mining. Needless to say, these three approaches cannot be independent, and any efficient mining of the web would require a judicious combination of information from all the three sources. We have presented in this paper the significance of introducing the web mining techniques in the area of web personalization. Personalization requires analysis of your goals and the development of business requirements, use cases, and metrics. Once these are fully understood, you may find that your personalization strategy doesn't require substantial augmentation of your application environment. If you do find that the integration of a personalization tool is necessary, with this knowledge, you'll be able to better analyze and judge the offerings. In less than a decade, the World Wide Web has become one of the world's three major media, with the other two being print and television. Electronic commerce is one of the major forces that allow the Web to flourish, but the success of electronic commerce depends on how well the site owners understand users' behavior and needs. Web usage mining can be used to discover interesting user navigation patterns, which can then be applied to real-world problems such as Web site/page improvement, additional product/topic recommendations, user/customer behavior studies, etc. This paper has provided the requirements of Web usage mining and the introduction of web 2.0 technology. Improving quality and extension of our models will be the following steps in our project. The development and application of Web mining techniques in the context of Web content, usage, and structure data will lead to tangible improvements in many Web applications, from search engines and Web agents to Web analytics and personalization. Future efforts, investigating architectures and algorithms that can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web applications.

9. REFERENCES

[1]. Jose Aguilar 2009 Information science and applications, Issue 9, Volume 6, page 1523-1532.

[2]. S. Asur and B. A. Huberman, "Predicting the future with social media," HP Laboratories, Tech. Rep. HPL-2010-53, 2010.

[3]. Berners-Lee J, Hendler J, Lassila O (2001) The Semantic Web. Scientific American, vol. 184, pp34-43.

[4]. Cecconi A, Galanda M (2002) Adaptive Zooming in Web Cartography. In Proceedings of SVG Open 2002 (Zurich, Switzerland), pp787-799.

[5]. Chen L, Sycara K (1998) A Personal Agent for Browsing and Searching. In Proceedings of the 2nd International Conference on Autonomous Agents, Minneapolis/St. Paul, May 9-13, pp132-139.

[6]. Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. *ACM Transactions On Internet Technology (TOIT)*, 3(1), 1-27.

[7]. Ghani, R. and A. Fano. Building Recommender Systems Using a Knowledge Base of Product Semantics. in Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (AH2002). 2002, p. 11-19, Spain.

[8]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, January 2000/Vol. 1, Issue 2, pp. 12-23.

[9]. Mobasher, B., Web Usage Mining and Personalization, in Practical Handbook of Internet Computing, M.P. Singh, Editor. 2004, CRC Press. p. 15.1-37.

[10]. Meo R., Lanzi P., Matera M., Esposito R. (2004).Integrating Web Conceptual Modeling and Web Usage Mining. In *Proc. of "WebKDD- 2004 workshop on Web Mining and Web Usage Analysis"*, part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.

[11]. Mobasher, B., Cooley, R., and Srivastava, J. (2000). Automatic personalization based on web usage mining, *Commuunications of the. ACM*, 43(8) 142–151.

[12]. Mobasher B., Dai H., Luo T., and Nakagawa M. (2001). Effective personalizaton based on association rule discovery from Web usage data, *ACM Workshop on Web information and data management*, Atlanta, GA.

[13]. Pierrakos, D., et al. *Web* Community Directories: A New Approach to Web Personalization. in Proceeding of the 1st European Web Mining Forum (EWMF'03). 2003, p. 113-129, Cavtat-Dubrovnik, Croatia.

[14]. Schafer J.B., Konstan J., and Reidel J. (1999). Recommender Systems in E-Commerce, In *Proc. ACM Conf. E-commerce*, 158-166.