ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (265-270), Month: April-June 2014, Available at: www.researchpublish.com

An Evaluation of Personalization Systems using Web Mining Techniques

K.Nirosha [1] V.Karthick[2] and Mrs.E.Jaya[3]

[1,2]Research Scholars, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai-63
[3]Professor and Head, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai-63

Abstract: **Web Personalization is viewed as an application of data mining and machine learning techniques to build models of user behavior that can be applied to the task of predicting user needs and adapting future interactions with the ultimate goal of improved user satisfaction. We start by providing a description of the personalization process and a classification of the current approaches to Web personalization. We discuss the various sources of data available to personalization systems, the modeling approaches employed and the current approaches to evaluating these systems. Recently a lot of research has been done on personalized applications, which are able to tailor the information presented to individual users. The main goal of these personalized systems is to learn the users needs without asking for it explicitly. In many approaches a web user profile is constructed that help to customize the information presented. Typically, the personal profiles are composed of the browsing data that was collected from the particular users previously. At the present time, such web user profiles already find application in various areas of information retrieval. They are employed to re-rank search results, modify user queries or assist during the retrieval process.**

Keywords: **Web Mining, Personalization, Collaborative, Content Mining, Structure Mining.**

1. INTRODUCTION

World Wide Web has become the biggest and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs (blogs). Every day, the web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line. Because of its rapid and chaotic growth, the resulting network of information lacks of organization and structure. Users often feel disoriented and get lost in that information overload that continues to expand. On the other hand, the e-business sector is rapidly evolving and the need for web market places that anticipate the needs of their customers is more than ever evident. Therefore, the ultimate need nowadays is that of predicting the user needs in order to improve the usability and user retention of a web site.

2. WEB MINING

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories.

2.1Web 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. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, 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.

Figure 1 represents web taxonomy is shown in figure 1.

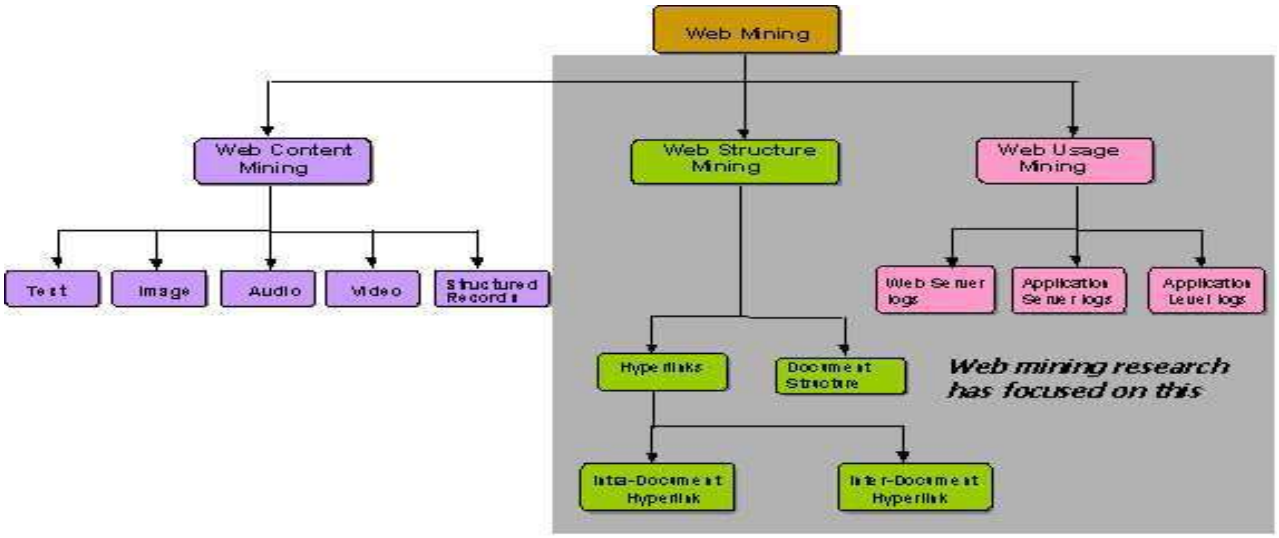

Figure 1. Web Mining Taxonomy

2.2Web 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.

 Hyperlinks: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within 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 within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents.

2.3Web 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. Web usage mining itself can be classified further depending on the kind of usage data considered:

 Web Server Data: The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

 Application Server Data: Commercial application servers such as Weblogic , Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

 Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

3. USER PROFILES

The Web has taken user profiling to completely new levels. For example, in a 'brick and-mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user, providing a much more detailed insight into the decision making process. Adding such behavioral information to other kinds of information about users, e.g. demographic, psychographic, etc., allows a comprehensive user profile to be built, which can be used for many different applications. While most organizations build profiles of user behavior limited to visits to their own sites, there are successful examples of building 'Web-wide' behavioral profiles. These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the Web.

4. PERSONALIZATION SYSTEMS

A common feature of most personalization systems is the application of user models/profiles to customize the systems behaviour to individual users. User models represent the information about users that is essential to support the adaptation functionality of the particular system. In this chapter we want to give a general overview about the most common techniques used for web user personalization.

5. PERSONALIZATION APPROACHES

Basically personalization systems can be divided into three main categories: rule-based systems, content-filtering filtering systems, and collaborative-filtering systems.

5.1 Rules-Based Systems

These kinds of systems rely on decision rules that are used to recommend items to users. The rules are used to affect the content served to a user whose profile satisfies one or more conditions. User profiles are mainly obtained through explicit interactions with users. Because they are often static, the systems performance degrades over time.

5.2 Content-Based Filtering Systems

In content-based filtering systems, a user profile represents the content descriptions of items in which the user has previously expressed interest. Items are represented by a set of features or attributes. The recommendation task in such systems usually involves the comparison of extracted features from unseen items with content descriptions in the user profile. In most systems content descriptions are textual features extracted from web pages. User profiles and new items are normally both represented as weighted term vectors. The primary drawback of such systems is their tendency to overspecialize the item selection, since profiles are only based on the previous rated items.

5.3 Collaborative-Filtering Systems

Systems falling into this category generally involve matching the ratings of a current user for objects with those of similar users in order to produce recommendations for objects not yet rated or unseen. The primary technique used to accomplish this task is the standard memory-based k-Nearest-Neighbour classification approach. Profiles of the target user are compared with the historical profiles of other users in order to find the top k users who have similar taste or interest. For instance, the e-commerce application Amazon applies this technique to recommend products to users based on the items other users with same interests purchased before. One of the main drawbacks is that a huge amount of data from numerous users has to be collected before any useful recommendations can be made. Furthermore, the bottleneck of such systems is that the neighbourhood formation phase is performed as an online process. In our research we focus on the classical content-based filtering approach, which builds the profiles merely from features associated with items previously seen or rated by the active user.

6. EVALUATION OF PERSONALIZATION SYSTEMS

Evaluation of personalization systems remains a challenge due to the lack of understanding of what factors affect user satisfaction with a personalization system. It seems obvious that a system that accurately predicts user needs and fulfils these needs without the user needing to expend the same resources in achieving the task as he would have, in the absence of the system, would be considered successful. Hence personalization systems have most commonly been evaluated is terms of the accuracy of the algorithms they employ. Recent user studies have found that a number of issues can affect the perceived usefulness of personalization systems including, trust in the system, transparency of the recommendation logic, ability for a user to refine the system generated profile and diversity of recommendations. For a business deploying a personalization system, accuracy of the system will be little solace if it does not translate into an increase in quantitative business metrics such as profits or qualitative metrics such as customer loyalty. Hence the evaluation of personalization systems needs to be carried out along a number of different dimensions, some of which are better understood that others and have well established metrics available. The key dimensions along which personalization systems are evaluated includes :

– User Satisfaction
– Accuracy
– Coverage
– Utility
– Explainability
– Robustness
– Performance and Scalability

7. CLASSIFICATIONS OF APPROACHES TO PERSONALIZATION

In this section we discuss various dimensions along which personalization systems can be classified based on the data they utilize, the learning paradigm used, the location of the personalization and the process that the interaction takes with the user.

7.1 Individual Vs Collaborative

The term personalization impresses upon the individuality of users and the need for systems to adapt their interfaces to the needs of the user. This requires data collected on interactions of users with the system to be modeled in a user-centric fashion. Typically, data is collected by the business with which the user is interacting and hence the business has access to data associated with all its customers. A personalization system may choose to build an individual model of user likes and dislikes and use this profile to predict/tailor future interactions with that user. This approach commonly requires content descriptions of items to be available and are often referred to as content-based filtering systems.

7.2 Reactive Vs Proactive

Reactive approaches view personalization as a conversational process that requires explicit interactions with the user either in the form of queries or feedback that is incorporated into the recommendation process, refining the search for the item of interest to the user. Most reactive systems for personalization have their origins in case-based reasoning research. Reactive systems can be further classified based on the types of feedback they expect from the user. Common feedback mechanisms used by these systems include value elicitation, critiquing/tweaking, rating and preference feedback. Value elicitation and tweaking/critiquing are feature based approaches to feedback. While in value elicitation the user must provide a rating for each feature of each recommendation object presented to the user, based on its suitability to the users needs, in tweaking/critiquing the user only provides directional feedback (for example, “too high”, “too low”) on feature values for the recommended object. Rating and preference are feedback approaches at the object level. In rating based feedback, the user must rate all the recommendations presented to him, based on their ‘fit’ with his requirements. In preference feedback the user is provided with a list of recommendations and is required to choose one of the recommendations that best suits his requirement. The system then uses this feedback to present the user with other, similar objects. The iterations continue until the user finds an object of interest or abandons the search.

7.3 User Vs Item Information

Personalization systems vary in the information they use to generate recommendations. Typically, the information utilized by these systems include:

– *Item Related Information:* This includes content descriptions of the items being recommended and a product/ domain ontology

– *User Related Information:* This includes past preference ratings and behaviour of the user, and user demographics

Systems that use item related information generally deal with unstructured data related to the items. Once this data has been processed, into relational form such as a bag-of-words representation commonly used for textual data, a user profile is generated.

8. WEB USAGE MINING AND PERSONALIZATION

Web usage mining is the process of identifying representative trends and browsing patterns describing the activity in the web site, by analyzing the users' behaviour. Web site administrators can then use this information to redesign or customize the web site according to the interests and behavior of its visitors, or improve the performance of their systems. Moreover, the managers of e-commerce sites can acquire valuable business intelligence, creating consumer profiles and achieving market segmentation. There exist various methods for analyzing the web log data. Some research studies use well known data mining techniques such as association rules discovery, sequential pattern analysis, clustering, probabilistic models, or a combination of them. Since web usage mining analysis was initially strongly correlated to data warehousing, there also exist some research studies based on OLAP cube models. Finally some proposed web usage mining approaches that require registered user profiles, or combine the usage data with semantic meta-tags incorporated in the web site's content. Figure 2 represents the web personalization process.

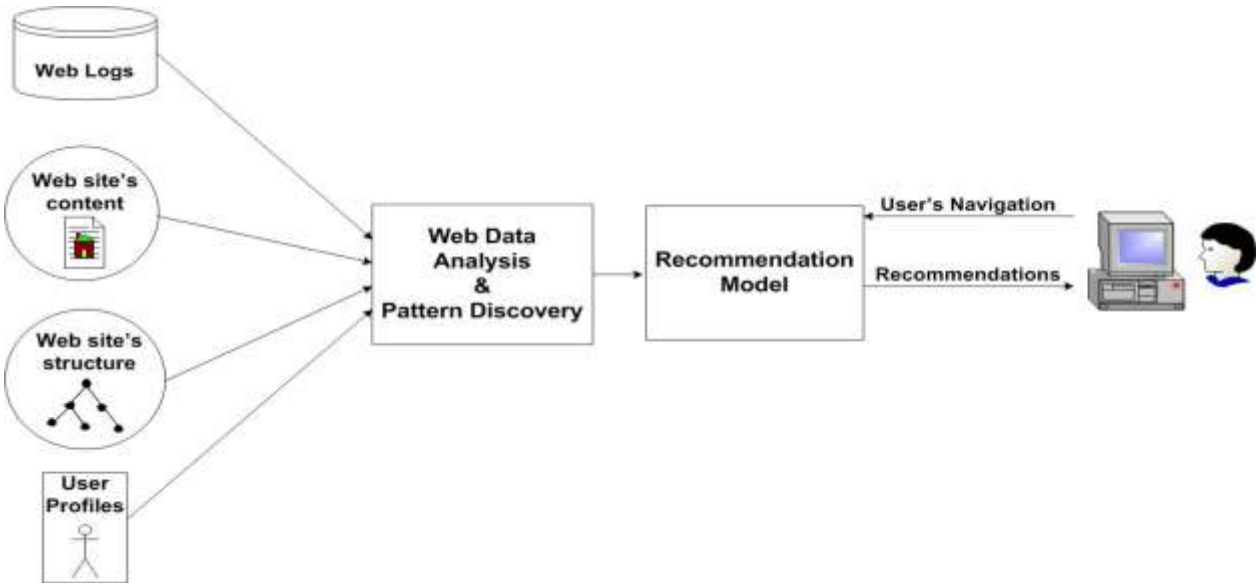

Fig.2 Web Personalization Process

Furthermore, this knowledge can be used to automatically or semi-automatically adjust the content of the site to the needs of specific groups of users, i.e. to personalize the site. As already mentioned, web personalization may include the provision of recommendations to the users, the creation of new index pages, or the generation of targeted advertisements or product promotions. The usage-based personalization systems use association rules and sequential pattern discovery, clustering, Markov model, machine learning algorithms, or are based on collaborative filtering in order to generate recommendations. Some research studies also combine two or more of the aforementioned techniques.

9. CONCLUSION AND FUTURE WORK

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key computer science contributions made by the field, a number of prominent applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion. The ultimate goal of personalization is a lift in user satisfaction. However, most research into personalization has focussed evaluation on the accuracy of predicted ratings and little agreement has emerged as to what factors, other than prediction accuracy affect user satisfaction. Even less agreement exists with regard to how the effect of personalization on these factors should be measured. A lot more user studies need to be carried out to gain a better understanding of these issues. The development of more personalization exemplars with the necessary infrastructure to conduct large scale user testing is required.

REFERENCES

[1] Alessandro Micarelli, Filippo Sciarrone, Mauro Marinilli; Department of Computer Science and Automation, Roma Tre University: Web Document Modeling (2007)

[2] C. Anderson, P. Domingos, D. S. Weld, Relational Markov Models and their Application to Adaptive Web Navigation, in Proc. of the 8th ACM SIGKDD Conference, Canada (2002)

[3] M. Albanese, A. Picariello, C. Sansone, L. Sansone, A Web Personalization System based on Web Usage Mining Techniques, in Proc. of WWW2004, New York (2004)

[4] B. Berendt, Using site semantics to analyze, visualize and support navigation, in Data Mining and Knowledge Discovery Journal, 6: 37-59 (2002)

[5] J. Borges, M. Levene, A Dynamic Clustering-Based Markov Model for Web Usage Mining, Technical Report, available at http://xxx.arxiv.org/abs/cs.IR/0406032 (2004)

[6] R. Cooley, B. Mobasher, J. Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web, in Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '97)

[7] E. Colet. Using Data Mining to Detect Fraud in Auctions, 2002.

[8] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, 2000.

[9] JiaWei Han, Micheline Kamber: Data Mining – Concepts and Techniques: Chapter 7 - Cluster Analysis: China Machine Press Second Edition (2006)

[10] S. Holland, M. Ester, W. Kiebling, Preference Mining: A Novel Approach on Mining User Preferences for Personalized Applications, in Proc. of the 7th PKDD Conference (2003)

[11] Mirza, B.J.: Jumping connections: A graph theoretic model for recommender systems. MSc Thesis, Virginia Tech (2001)

[12] Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Using sequential and non-sequential patterns for predictive web usage mining tasks. In: Proceedings of the IEEE International Conference on Data Mining. (2002)