INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A NOVEL APPROACHES IN WEB MINING TECHNIQUES IN CASE OF WEB PERSONALIZATION

Y.Raju[1], Dr. D. Suresh Babu[2]
[1] Geethanjali College of Engineering and Technology, IT Department, Hyderabad, India.
Email: raju.yeligeti@gmail.com
[2] Departments of Computer Science, Head, Kakatiya Government College, Kakatiya University.
Email: sureshd123@gmail.com

Abstract: Web mining is the application of data mining techniques to extract knowledge from Web. Web mining has-been explored to a vast degree and different techniques have been proposed for a variety of applications that includes Web Search, Classification and Personalization etc. Most research on Web mining has been from a 'data-centric' point of view. In this paper, we highlight the significance of studying the evolving nature of the Web personalization. Web usage mining is used to discover interesting user navigation Patterns and can be applied to many real-world problems, such as improving Web sites/pages, making Additional topic or product recommendations, user/customer behavior studies, etc. A Web usage mining system performs five major tasks: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. Each task is explained in detail and its related technologies are introduced. The Web mining research is a converging research area from several research communities, such as Databases, Information Retrieval and Artificial Intelligence. In this paper we implement how Web mining techniques can be apply for the Customization i.e. Web personalization.

Keywords: Usage Mining, Navigation Patterns, Pattern Analysis, Content Mining, Structure 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. Typically, the following problems are often mentioned in Web related research and applications:

(1) Finding relevant information: To find specific information on the web, users often either browse Web documents directly or use a search engine as a search assistant. When a user utilizes a search engine to locate information, he or she often enters one or several keywords as a query, then the search engine returns a list of ranked

pages based on the relevance to the query. However, there are usually two major concerns associated with the query-based Web search [1]. The first problem is low precision, which is caused by a

lot of irrelevant pages returned by the search engine. The second problem is low recall, which is due to the lack of capability of indexing all Web pages available on the Internet. This causes the difficulty in locating the unhindered information that is actually relevant.

(2) Finding needed information: Most search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine don't exactly match what a user really needs due to the fact of the existence of the homology. For example, when one user with an information technology background wishes to search information with respect to "Python" programming language, he/she might be presented with information on the creatural python, one kind of snake rather than the programming language, given entering only one "python" word as query. In other words, the semantics of Web data [3] is rarely taken into account in the context of Web search.

(3) learning useful knowledge: With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them (data mining oriented). More interestingly, more studies [4-6] have been conducted on how to utilize the Web as a knowledge base for decision making or knowledge discovery recently.

(4) Recommendation/personalization of information: While 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

2. WEB MINING TECHNIQUES

Web mining is a rapid growing research area. It consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks. Web content mining aims to extract/mine useful information or knowledge from web page contents.

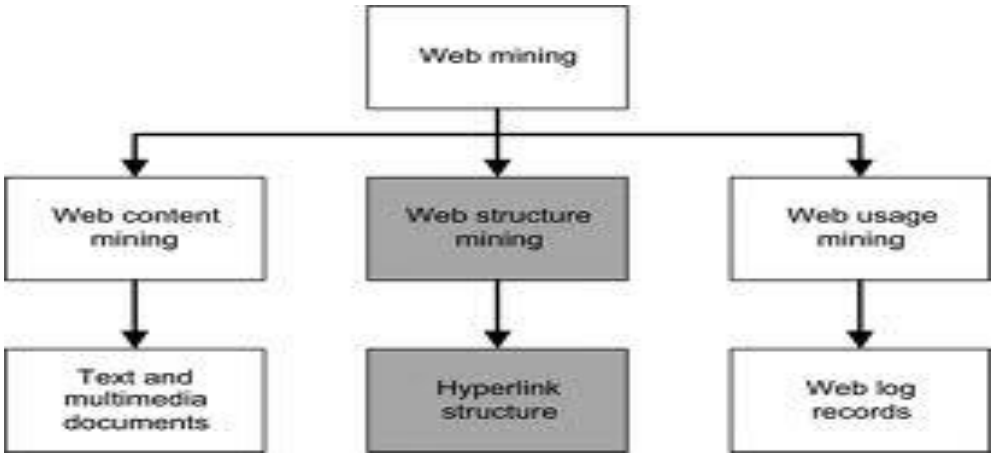

Figure 1. Web Mining Techniques

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 and 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).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.

Web Structure Mining: Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, and linking the information through reference links to bring forth the specific page containing the desired information.

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Web Usage Mining: Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. 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 the Web server. Typical data includes IP address, page reference and access time.
- Application Server Data: Commercial application servers 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 thus 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 categories above. We describe the web usage mining activities of an on-going project, called Click World that aims at extracting models of the navigational behavior of web site users. The models are inferred from the access logs of a web server by means of data and web mining techniques. The extracted knowledge is deployed to the purpose of f offering a personalized and proactive view of the web services to users.

4. PERSONALIZATON ON THE WEB

Personalization technology enables the dynamic insertion, customization or suggestion of content in any format that is relevant to the individual user, based on the user's implicit behavior and preferences, and explicitly given details.

Web pages: Web pages are personalized based on the characteristics (interests, social category, context,) of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed. The term customization is used instead when the site only uses explicit data such as ratings or preferences.

There are three categories of personalization:

1. Profile / Group based
2. Behavior based (also known as Wisdom of the Crowds)
3. Collaboration based

Web personalization models include rules-based filtering, based on "if this, then that" rules processing, and collaborative filtering, which serves relevant material to customers by combining their own personal preferences with the preferences of like-minded others. Collaborative filtering works well for books, music, video, etc. However, it does not work well for a number of categories such as apparel, jewelry, cosmetics, etc. Recently, another method, "Prediction Based on Benefit", has been proposed for products with complex attributes such as apparel.

There are three broad methods of personalization:

1. Implicit
2. Explicit
3. Hybrid

With implicit personalization the personalization is performed by the web page (or information system) based on the different categories mentioned above. With explicit personalization, the web page (or information system) is changed by the user using the features provided by the system. Hybrid personalization combines the above two approaches to leverage the best of both worlds.

5. PERSONALIZATION STRATEGIES

1. Embrace the Process:

Don't make the mistake of measuring all your marketing efforts in decimals and dollar signs. Big Data can allow us to break free of this limiting yardstick and start measuring in three dimensions. Many visitor interactions are valuable and successful for other reasons. Visitors might read the great blog post you shared, download a free report, chat with a customer service rep, subscribe to an e-newsletter, or compare products without giving you a dime. But all these touch points have likely made them more familiar, trusting, and loyal to your brand. And they've allowed you to gather more behavioral data, which you can now use to enhance personalization over time.

On the path to conversion, every step is important. Identify each step, determine what makes it valuable, and optimize accordingly. You can then measure success by whether or not one touch point led to the next.

2. Cultivate Existing Customers:

It may seem counterintuitive to use Big Data to hone in on smaller targets, but this strategy has proven results. For one, customer acquisition can be costly and time intensive. Studies show "it's more cost-effective to cultivate existing customers than to find new ones." Big Data can yield tremendous insight into the loyal folks who've already clicked through, subscribed, and made repeat orders. Integrate these individuals' demographic, behavioral, and purchase-history data and continue populating emails and WebPages with personalized messaging that anticipates their questions and needs. You'll make it easy (i.e., simple) for them to stay loyal.

3. Generate Insights, Not Reports:

A report tells you the percentage of visitors who clicked through, or the number of would-be customers who abandoned their cart. Insights tell you why. With Big Data, it's possible to answer questions that we used to only guess at, such as "Who are the 15% who signed up and why?"

4. Pay Attention to Deviations:

Big Data can liberate marketers from limiting categorizations of customers. When an individual does something unexpected or deviates from the trend, we can take note and respond. What may have been perceived as a failure of analytics, can now be an opportunity for engagement. If a customer buys something out of the ordinary, or fails to renew a standing order, a company can immediately reach out with discounts on a new line of products, a simple survey, or a freebie. This reminds people that you value them and can be flexible enough to grow with them.

Even negative visitor experiences can be turned into positive opportunities. If you're tapping into Big Data, you can listen to customers across multiple channels, registering their complaint on Twitter, and sending them a personalized message to remedy the situation.

The Web personalization process can be divided into four distinct phases:

(1) Collection of Web data – Implicit data includes past activities/click streams as recorded in Web server logs and/or via cookies or session tracking modules. Explicit data usually comes from registration forms and rating questionnaires. Additional data such as demographic and application data (for example, e-commerce transactions) can also be used. In some cases, Web content, structure, and application data can be added as additional sources of data, to shed more light on the next stages.

(2) Preprocessing of Web data – Data is frequently pre-processed to put it into a format that is compatible with the analysis technique to be used in the next step. Preprocessing may include cleaning data of inconsistencies, filtering out irrelevant information according to the goal of analysis (example: automatically generated requests to embedded graphics will be recorded in web server logs, even though they add little information about user interests), and completing the missing links (due to caching) in incomplete click through paths. Most importantly, unique sessions need to be identified from the different requests, based on a heuristic, such as requests originating from an identical IP address within a given time period.

(3) Analysis of Web data – Also known as Web Usage Mining, this step applies machine learning or Data Mining techniques to discover interesting usage patterns and statistical correlations between web pages and user groups. This step frequently results in automatic user profiling, and is typically applied offline, so that it does not add a burden on the web server.

(4) Decision making/Final Recommendation Phase – The last phase in personalization makes use of the results of the previous analysis step to deliver recommendations to the user. The recommendation process typically involves

generating dynamic Web content on the fly, such as adding hyperlinks to the last web page requested by the user. This can be accomplished using a variety of Web technology options such as CGI programming.

8. CONCLUSION

In this article, 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.

REFERENCES

[1] Han, J. and M. Kamber, *Data Mining: Concepts and Techniques*. 2007: Morgan Kaufmann.
[2] Berners-Lee J, Hendler J, Lassila O (2001) The Semantic Web. Scientific American, vol. 184, pp34-43.
[3] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). Measuring the accuracy of sessionizers for web usage analysis, In Workshop on Web Mining, at the First SIAM International Conference on Data Mining, 7-14.
[4] Berendt B., Hotho A., and Stumme G. (2002). Towards semantic web mining. In Proc. International Semantic Web Conference (ISWC02).
[5] Cecconi A, Galanda M (2002) Adaptive Zooming in Web Cartography. In Proceedings of SVG Open 2002 (Zurich, Switzerland), pp787-799
[6] 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.
[7] Desikan P. and Srivastava J. (2004), Mining Temporally Evolving Graphs. In Proceedings of "WebKDD- 2004 workshop on Web Mining and Web Usage Analysis", B. Mobasher, B. Liu, B. Masand, O. Nasraoui, Eds. part of the ACM KDD: Knowledge Discovery and Data Mining Conference, Seattle, WA.
[8] Eirinaki M., Vazirgiannis M. (2003). Web mining for web personalization. ACM Transactions On Internet Technology (TOIT),
[9] 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 the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (AH2002). 2002, p. 11-19, Malaga, Spain.
[10] 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
[11] Kargupta H., Datta S., Wang Q., and Sivakumar K. (2003). On the Privacy Preserving Properties of Random Data Perturbation Techniques, In Proc. of the 3rd ICDM IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL.

[12] Linden G., Smith B., and York J. (2003). Amazon.com Recommendations Item-to-item collaborative filtering, IEEE Internet.

Author Biography

Y. Raju Working as an Associate professor in IT Dept at GCE, Hyderabad. He received M.Tech from JNTUH. Presently Pursuing Ph.D. from JNTUH. He has published papers in international journal and conferences. His main research area includes Data Mining, Information retrieval System and Artificial Intelligence.

Dr. D.Suresh Babu is currently working Head, Department of Computer Science, Kakatiya Government College, Kakatiya University, Warangal India. He has received his Ph.D. Degree in Computer science & Engineering from Acharya Nagarjuna University, Guntur, A.P., INDIA. His main research interest includes Data Mining, neural networks, Information retrieval System and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published more than15 papers in International journals and conferences.