Web Page Personalization using Content Based Retrieval in Data Mining

Author

Ajay Eknathji Meshram

Department of Computer Science, Patel Institute of Technology
Rajiv Gandhi Proudyogiki Vishwavidyalaya (RGPV), Bhopal

Email-kusumsut@gmail.com

Abstract

Today’s World Wide Web market is becoming more competitive, so it’s more important than ever, to provide customers with an interactive, personal Web experience. The ease and speed with which business transaction can be carried out over the web, has been a key driving force in the rapid growth of e-commerce. The ability to track users browsing behavior down to individual mouse clicks, has brought the vendor and end customer closer than ever before. Sites must plan now to respond to this expectation or risk being left behind, as the Web changes to a personal medium. It is now possible for vendor to personalize the product message, for individual customers at a massive scale, which is referred as Web Page Personalization. There are various techniques to serve the purpose of the personalization task. Web Personalization allows you to have a Web site, that tailors Web content to a Web user’s preferences, and other profile information. In addition, a personalization system logs every Web page displayed to every user so you can develop a ”clickstream” view of what they saw, when they saw it, and for how long. To engage visitors to a web site at a very early stage (i.e. before registration or authentication), Personalization Tools must rely primarily on click stream data, captured in web server logs. The lack of explicit user ratings, as well as the sparse nature and the large volume of data, such as setting, lead serious challenges. The problem with the www is the users behavior, being dynamic in nature, which changes depending on the users interest and many constraints. The size of the data to be handled is in terms of terabytes. To deal with such problem the data mining techniques is one solution, which is applied on the WWW database, and called as Web Mining. The personalized Web site provides benefits for your customers, and your company, at the same time, the Web
users, will enjoy receiving information, that fits their unique needs. With Personalization, you are able to form loyal and long-lasting relationships with each individual.

I. INTRODUCTION

The web contains the rich and dynamic collection of hyperlink information, and web page access and usage information, providing the rich sources of data mining. So one of the popular applications of data mining is World Wide Web Mining. The web poses great challenges for effective recourse and knowledge discovery. The challenges are due the characteristics of the web sites which are mentioned below:- We can structure web application development into four different dimensions

1. Evolution related characteristics (Evolution of Product)

2. Development related characteristics (Development of Product)

3. Application related characteristics (Product itself)

4. Usage related characteristics (Use of Product)

A. Evolution Related Characteristics

Web applications are subjected to frequent changes and permanent evolution. Unlike, conventional application software, that evolves over a series of planned, chronologically spaced releases, web applications evolve continuously. In the course of evolution, negotiability of quality often sacrifices maintainability. The development of web applications requires a methodologically, sound engineering approach called Web Engineering. Web Engineering aims to address, and resolve the multifaceted problems of Web based system development. Web Engineering is the application of systematic, disciplined and quantifiable approaches to the cost-effective development and evolution of high quality applications, in the World Wide Web. While Web Engineering adopts and encompasses many software engineering principles, it incorporates many new approaches, methodologies, tools, techniques and guidelines to meet the unique requirements of Web-based systems. Developing Web-based systems is significantly different from traditional software development, and poses many additional challenges. These challenges promoted research into efficient and effective discovery and use of resources on the internet. To increase business on the internet, it is required now to modify the web access, or the contents of the web page, to better fit the desires of the user. This may involve actually creating Web pages, that are unique per user, to determine what Web documents to retrieve. The Personalization can be achieved with the help of Web mining. The web mining is comprised of the following three parts-
Web Content Mining

Web Usage Mining - Web Structure Mining

The Web Mining System should do the following function

Authoritative web page Mining - Mining technique should be applied to identify the authoritative web pages for the particular topic. The algorithm like HITS (Hyperlink Induced Topic Search) can be used to find out the Authoritative Web links associated with particular topic.

Web Document Classification: The web document should be classified according to their category, based on some training set data. Key based document classification or Content Classification is used to classify the web document.

1) Multidimensional and Multilayer web information base:

The web can be layered to provide the multidimensional hierarchical view of the web. The Content Based Classification or Web Document Classification serves the purpose. On the web, the web logs are maintained, to keep the record of web page request. These web logs can be mined to find out the performance usage of the web pages on the internet.

2. Development Related Characteristics

Web application developers always need to deal with conditions, risks and uncertainty.

1) Process: Web application development processes are characterized by frequent changes and adjustments which are necessary due to rapid technological developments, fast changing trends, volatile requirements and rigid schedules.

2) Development environment: The technical infrastructure is characterized by a high degree of volatility and heterogeneity. Web application development relies on a broad spectrum of different components, such as Web Server, Application Server, Database Server. Because of increased time-to-market pressure, the components are always immature and fall short in stability, reliability and desired functionality.

3) Development team: Development team has to take multidisciplinary effort for the web development. Involvement of open source communities is also required for the web development. Younger team members may not have sufficient experience for the development should be guided time to time.

4) Legacy integration: Web applications often need to integrate legacy systems. The external services provided by these systems are rarely documented, and often change without notice, thus negatively affecting the quality.
3. Application Related Characteristics

While developing web applications, one has to consider not only functionality, but also content, hypertext and presentation aspects.

1) Presentation: In conventional software systems, the look feel is often determined by standardized user interface, elements and style guides. Presentation is a central quality factor for Web applications, not least to the high competitive pressure on the web, where visual appearance is subject to trends and new technical features. Web applications need to be self explanatory, requiring particular attention to visual design and the consistency of the interaction style behaviour.

2) Content: The origin of web is its role as information medium. Web applications are heavily content driven. Content comprises not only structured data residing in database systems, but also unstructured and semi-structured data, such as textual descriptions or multimedia information. Complexity arises especially from the fact, that the content is often highly dynamic and continuously updated. Also users typically demand for high content quality, in terms of topicality, accurateness, consistency and reliability.

3) Hypertext: For structuring information, web applications advocate the hypertext paradigm. The basic elements of the Web's notion of hypertext are nodes, links and anchors. Examples of accessing hypertextual information, include browsing (like in online store catalogues), querying (in e-learning applications). The essential features of the hypertext are, its nonlinearity requiring from both authors and users to address the potential issues of disorientation, and cognitive overload. This can be achieved through specific navigation design (site maps, keyword searches traversed paths etc) is essential to preserve quality of access.

D. Usage Related Characteristics

The users of web applications often vary in numbers and cultural background, use heterogeneous devices, and can freely choose the time and location of accessing the web application. Developers can not always predict all these potential settings.

1) Unpredictable Technical Infrastructure: Available end-user devices vary in hardware and software capabilities such as display, computational power or browser version. Network connections differ with respect to bandwidth, reliability, stability and availability. Complexity is even more increased because the actual representation of web on the client device is to a large extent outside the control of the developers. For example, Users configure their browsers individually and may even disallow certain features.

2) Natural Context: This includes aspects of the location and the time of access, offering new kinds of context based service. The possibility of immediate and permanent availability of web applications,
3) requires special quality considerations such as 24/7 availability.

4) Diversity and magnitude of User-Base: Web application users differ in age, social and cultural background, goals and skills. This heterogeneity has to be considered by developers, since the Web entails no obligations, and Web applications will only be used, if they bring immediate advantage. The way users interact with web can be hardly predicted, and may leave the application any time. Also the number of users may vary, making Scalability another crucial factor.

II. MOTIVATION

First, the user may want to personalize, to enable access to information content. This would be the case, when creating a bookmark, or when indicating the type of information, that should be displayed on the starting page of a portal. This aspect of motivation is particularly important in the World Wide Web, where the amount of information exceeds our processing capacities. With Web Page Personalization, advertisements to be sent to a potential customer, are chosen based on specific knowledge, concerning that customer. The second work-related reason to personalize would be to accommodate work goals. For example, some user found that stopping something on web page, that is annoying, were triggers for customizing software. Third, personalization may be to accommodate individual differences. An instance of this, would be providing access for people with visual disabilities. More precisely, Dominant individuals seem to prefer a computer with a dominant dialogue style, whereas submissive users prefer a submissive computer.

Unlike targeting, Personalization may be performed on the target Web page. The goal here is to entice a current customer to purchase something, he or she may not have thought about purchasing. Personalization is almost the opposite of Target-ing. With Targeting, businesses display their advertisements at other sites, visited by the users. With Personalization, when a particular person visits a Web site, the advertising can be designed specifically for that person. Personalization include such techniques as use of cookies, use of databases and more complex Data Mining and machine learning strategies. Web Usage Mining focuses on techniques, that could predict user behavior, while the user interacts with the web. One of the most successful and widely used technologies for building the Personalization System is record collaborative filtering (CF). Given a target users record of activity, CF based technique, such as K-Nearest Neighbor (KNN) approach, compare the record with the historical record of the other users, in order to find top K Neighbor, who have similar test or interest. But these techniques suffer through some of its basic limitation, which reduces the efficiency, scalability etc. Using Data Min-ing Technique, is one of the solutions to it. Numerous internet based business models have been developed recently, of which Electronic-Commerce is playing a key role.
The goal of the Personalization System is to provide the user, most likely links of the web page, depending upon users current navigated links. The system will use the Data Mining Techniques, like association rule and sequence pattern matching, on the Web Data and Transaction Data, to find out the likely accessed pattern. Until now what ever is the model given for the Web Personalization System is based only on the server log data, and the current user-browsing pattern. The system can be designed, which have combination of market-basket analysis and server log rule extraction.

III. STATEMENT OF THE PROBLEM

The ability to track users browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. Due to Web Page Personalization, it is now possible for vendor to personalize the product message for individual customers at a massive scale. Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the users navigational behavior (usage data) in correlation with other information collected in the Web context, namely structure, content and user profile data. There are various techniques to serve the purpose of the personalization task.

The problem with the www is the users dynamic behavior, which changes depending on the users interest and many constraints. The size of the data to be handled is in terms of terabytes. To deal with such problem the data mining techniques is one solution, which is applied on the WWW database. Use of data mining technique on the web data to serve the personalization task is one of the best choices to solve the problem. The main task in Web Personalization System is to provide the most likely visiting link in the future for particular user depending on various users current navigated browsing pattern. The job is to design the Web Personalization System for an e-commerce site. In our project, the site is of the superstore shop, which sells everything from food to electronic goods. The sellers want the site to be modified:- To increase the cross sells. To provide the different advertising policy to different user. To provide the ease of surfing the site to the user. To get the group of user of particular interest. This will help the seller to get potential customer and new market policy in future. Here the shop is maintaining the data for the purchased items and the customers of the shop. It is also maintaining server log data. The task is to mine both these data and get the association rule from these and provide the recommendation to most likely visiting link on the site.

A. Detailed Statement of the Problem

Considering all the requirements of the system, the complete system is divided in to following parts:

1) Designing Web Site : This part of the system will deal with the site design for e-commerce retailing, and creating the web meta data from the site in indexed based format.

2) Implementation of Personalization Using Registration Data: This part will do the personalization of the
homepage of each user, by using the information provided by user, during registration to the web site. The
data will help in taking decision for personalization, depending on the users likes and dislikes. ID3
Classification algorithm is used to take decision about users future choice.

3) Database Design Data Transfer: The web data is huge in size and have different dimensions associated
with it. So maintaining the data in warehouse is suitable choice for the system. The data will be collected
from the different input sources, it will go through the generalize steps of the database design.

4) Implementation of Data Mining Algorithms: After building the data base, the data mining algorithm is
operated on the data. Here the task is to get the Association Rule from the data base. The data mining
algorithm will take the data from the Database, and perform the mining and produce the rules, which will be
stored in a database again, along with their support. Apriori Association Algorithm is chosen for getting the
rules. In this engine, the administrator will specify the support for the rules. A set of association rule mined
from the database will be mainly related to market analysis, and navigational pattern. Different advertising
policy will be displayed, depending on particular user interest, or based on some constraint such as festival
season or some other parameter.

5) Recommendation Engine: This part of the system will take the input from the user, and mined database.
The Recommendation Engine will automatically modify the original site graphical user interface by making
the comparison of user window with the associations rule, and will give the most likely navigated links to be
used in the site. Also the links based on previous browsing patterns of users which can be derived from the
access log files of Web Server, will be recommended to user will be shown dynamically.

IV. CONCLUSION

Application of Data Mining for achieving Personalization has revolutionized the entire Electronic
Commerce (e-Commerce). In addition to being able to store and retrieve information in an individual’s
profile “on the fly,” you can also create “wizards” - software that can analyze information in the profile
database and make recommendations or comments specific to the individual. Data from the profile database
and clickstream log files can be used, by content creators to understand, which material is used more than
others, how long material is viewed, and in what order material is accessed. The previous work done for the
Personalization motivates us to implement the project. The system analyzes the user browsing pattern on
specific window, compares it with association rule and provide most likely visiting link pattern.
Personalization will serve to provide the user, easy navigation of the web sites. It will also play a vital role in
deciding the business policy and will also result in increase in cross sale of the organization. Producers of
television programs use similar audience data, based on samples (diaries, interviews, set monitors, etc.) to
understand the viewing habits of the audience. With personalization, you can gather detailed information
about every member of your audience, and tailor the presentation to each individual.
V. ACKNOWLEDGEMENT

I would like to express my special thanks to the Principal, Patel Institute of Technology, Ratibad, Bhopal, who have been a tremendous mentor for me. I would like to thank you, for encouraging my work, and for allowing me to be guided by the Head of the Department, Department of Computer Science, Patel Institute of Technology (PIT), Bhopal. Your advice on this work have been priceless. I would also like to thank all the Professors, Teaching and non-teaching staff, for their co-operation. I would also like to thank all of my family members and friends who supported me in writing, and incented me to strive towards my goal. Once again, Thanks and Regards to all!

VI. REFERENCES

1. Filip Coenen, Gilbert Swinnen, Koen Vanhoof, Geert Wets, A Framework for Self Adaptive Websites: Tactical versus Strategic Changes, in proceedings of WEBKDD2000 Web Mining for Ecommerce Challenges and Opportunities, 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2000, Boston, MA

2. Zhexue Huang, Joe Ng, David W. Cheung, Michael K. Ng, Wai-Ki Ching, A Cube Model for Web Access Sessions and Cluster Analysis, in Proceedings of the Mining Log Data Across All Customer TouchPoints Workshop (WEBKDD01), August 2001

3. Bettina Berendt, Understanding Web usage at different levels of abstraction: coarsening and visualizing sequences, In Proceedings of the Workshop ”WEBKDD 2001 - Mining Log Data Across All Customer TouchPoints”, 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, August 2001, San Francisco, CA


11. Bettina Berendt, Web usage mining, site semantics, and the support of navigation, in Proceedings of the Workshop WEBKDD2000 - Web Mining for E-Commerce - Challenges and Opportunities, 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, August 2000, Boston, MA 29


13. Alex Buchner, Maurice D. Mulvenna, Discovering internet marketing intelligence through on line analytical Web usage mining, SIGMOD Record, 27(4), pp. 54-61, 1998


16. Jiawei Han, Micheline Kamber, Data Mining, Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers