World Wide Web Personalization

Olfa Nasraoui
Department of Computer Engineering and Computer Science
Speed School of Engineering
University of Louisville
Louisville, KY 40292
USA

email: o0nasr01@louisville.edu

INTRODUCTION

The Web information age has brought a dramatic increase in the sheer amount of information (*Web content*), the access to this information (*Web usage*), as well as the intricate complexities governing the relationships within this information (*Web structure*). Hence, not surprisingly, information overload, when searching and browsing the WWW, has become the "plague du jour". One of the most promising and potent remedies against this plague comes in the form of personalization. *Personalization* aims to customize the interactions on a website depending on the user's explicit and/or implicit interests and desires.

BACKGROUND

The Birth of Personalization: No Longer an Option, But a Necessity

The move from *traditional* physical stores of products or information (such as grocery stores or libraries) to virtual stores of products or information (such as *e-commerce sites* and *digital libraries*) has practically eliminated physical constraints traditionally limiting the number and variety of products in a typical inventory. Unfortunately, the move from the physical to the virtual space has drastically limited the traditional three dimensional layout of products for which access is further facilitated thanks to the sales representative or librarian who *know* their *products and* their *customers*, to a dismal *planar* interface *without* the sales representative or librarian. As a result, the customers are drowned by the huge number of options, most of which they may never even get to know. In the late 90's, Jeff Bezos, CEO of Amazon™ once said, "*If I have 3* *million customers on the Web, I should have 3 million stores on the Web*" (Schafer et al., 1999). Hence, in both the e-commerce sector and digital libraries, Web personalization has become more of a necessity than an option. Personalization can be used to achieve several goals, ranging from increasing customer loyalty on e-commerce sites (Schafer et al., 1999) to enabling better search (Joachims T., 2002).

Table 1: Possible Goals of Web Personalization

- Converting browsers into buyers
- Improving web site design and usability
- Improving customer retention and loyalty
- Increasing cross-sell by recommending items related to the ones being considered
- Helping visitors to quickly find relevant information on a website
- Making results of information retrieval/search more aware of the context and user interests

Modes of Personalization

Personalization falls into four basic categories, ordered from the simplest to the most advanced:

(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. It is usually implemented on the Web server. This mode depends more on Web technology than on any kind of adaptive or intelligent learning. It can also jeopardize user privacy.

(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!™.

(3) *Guidance or Recommender Systems* – A guidance based system tries to *automatically* recommend hyperlinks that are deemed to be relevant to the user's interests, in order to facilitate access to the needed information on a large website (Schafer et al., 1999; Mobasher et al., 2000; Nasraoui et al., 2002). It is usually implemented on the Web server, and relies on data that reflects the user's interest *implicitly* (browsing history as recorded in Web server logs) or *explicitly* (*user profile* as entered through a registration form or questionnaire). This approach will form the focus of our overview of Web personalization.

(4) *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.

In the following, we concentrate on the third mode of personalization, namely, automatic Web personalization based on *recommender systems*, because they necessitate a minimum or no explicit input from the user. Also, since they are implemented on the server side, they benefit from a global view of all users' activities and interests in order to provide an *intelligent* (learns user profiles automatically), and yet *transparent* (requiring very little or no explicit input from the user) Web personalization experience.

MAIN THRUST

Phases of Automatic Web Personalization

The Web personalization process can be divided into four distinct phases (Schafer et al., 1999; Mobasher et al., 2000):

(1) *Collection of Web data – Implicit* data includes past activities/clickstreams 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 clickthrough 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* (Spiliopoulou and Faulstich, 1999; Nasraoui et al., 1999; Srivastava et al., 2000), 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.

Categories of Data used in Web Personalization

The Web personalization process relies on one or more of the following data sources (Eirinaki and Vazirgiannis, 2003):

(1) *Content Data* – Text, images, etc, in HTML pages, as well as information in databases.

(2) *Structure Data* –Hyperlinks connecting the pages to one another.

(3) *Usage Data* – Records of the visits to each web page on a website, including time of visit, IP address, etc. This data is typically recorded in Web server logs, but it can also be collected using cookies or other session tracking tools.

(4) *User Profile* – Information about the user including demographic attributes (age, income, etc), and preferences that are gathered either explicitly (through registration forms) or implicitly (through Web server logs). Profiles can be either static or dynamic. They can also be individualized (one per user) or aggregate (summarize several similar users in a given group).

Different Ways to Compute Recommendations

Automatic Web personalization can analyze the data to compute recommendations in different ways, including:

(1) *Content-based or Item-based filtering* – This system recommends items deemed to be similar to the items that the user liked in the past. Item similarity is typically based on domain specific *item attributes* (such as author and subject for book items, artist and genre for music items). This approach has worked well for Amazon™

(Linden et al., 2003), and has the advantage of easily including brand new items in the recommendation process, since there is no need for any previous implicit or explicit user rating or purchase data to make recommendations. (2) *Collaborative filtering* – Based on the assumption that users with similar past behaviors (rating, browsing, or purchase history) have similar interests, this system recommends items that are liked by other users with similar interests (Schafer et al., 1999). This approach relies on a historic record of all user interests such as can be inferred from their ratings of the items on a website (products or web pages). Rating can be *explicit* (explicit ratings, previous purchases, customer satisfaction questionnaires) or *implicit* (browsing activity on a website). Computing recommendations can be based on *lazy* or *eager* learning phase to model the user interests. In *lazy* learning all previous user activities are simply stored, until recommendation time, when a new user is compared against all previous users to identify those who are similar, and in turn generate recommended items that are part of these similar users' interests. Lazy models are fast in training/learning, but they take up huge amounts of memory to store all user activities, and can be *slow at recommendation time* because of all the required comparisons. On the other hand, *eager* learning relies on data mining techniques to learn a summarized model of user interests (a decision tree, clusters/profiles, etc) that typically requires only a *small* fraction of the memory needed in lazy approaches. While eager *learning* can be slow, and is thus performed *offline*, using a learned model *at recommendation time* is generally much *faster* than lazy approaches.
(3) *Rule-based filtering* – In this approach, used frequently to customize products on e-commerce sites such as Dell on Line™, the user answers several questions, until receiving a customized result such as a list of products. This approach is mostly based on heavy planning and manual concoctions of a judicious set of questions, possible answer combinations, and customizations by an expert. It suffers from a *lack* in *intelligence* (no automatic learning), and tends to be *static*.

Recommender Systems

One of the most successful examples of personalization comes in the form of *recommender systems*. Several approaches to automatically generate Web recommendations based on user's Web navigation patterns or ratings exist. Some involve learning a usage model from Web access data or from user ratings. For example, lazy modeling is used in collaborative filtering which simply stores all users' information and then relies on *K-Nearest-Neighbors* (*KNN*) to provide recommendations from the previous history of similar users (Schafer et al., 1999). *Frequent itemsets* (Mobasher et al., 2001), a partitioning of user sessions into *groups* of *similar* sessions, called session *clusters* (Nasraoui et al., 1999; Mobasher et al., 2000), or user *profiles* (Nasraoui et al., 1999; Mobasher et al., 2000) can also form a user model obtained using data mining. Association rules can be discovered offline, and then used to provide recommendations based on web navigation patterns.

Among the most popular methods, the ones based on collaborative filtering and the ones based on fixed support association rule discovery may be the most difficult and expensive to use. This is because, for the case of *high-dimensional* (too many web pages or items) and extremely *sparse* (most items/web pages tend to be unrated/unvisited) Web data, it is difficult to set suitable support and confidence thresholds to yield reliable and complete web usage patterns. Similarly, collaborative models may struggle with sparse data, and do not scale well to a very large number of users (Schafer et al., 1999).

Challenges in WWW Personalization

WWW personalization faces several tough challenges that distinguish it from the main stream of data mining:

(1) *Scalability* – In order to deal with large websites that have huge activity, personalization systems need to be *scalable*, i.e. efficient in their time and memory requirements. To this end, some researchers (Nasraoui et al., 2003) have started considering web usage data as a special case of *noisy data streams* (data that arrives continuously in an environment constrained by stringent memory and computational resources. Hence the data can only be processed and analyzed *sequentially*, and cannot be stored).

(2) *Accuracy* – WWW personalization poses an enormous risk of upsetting users or e-commerce customers in case the recommendations are inaccurate. One promising approach (Nasraoui and Pavuluri, 2004) in this direction is to add *an additional data mining phase* that is separate from the one used to discover user profiles by clustering previous user sessions, and whose main purpose is *to learn an accurate recommendation model.*

This approach differs from existing methods that do not include adaptive learning in a separate second phase, and instead base the recommendations on simplistic assumptions (e.g. nearest profile recommendations, or deployment of pre-discovered association rules). Based on this new approach a new method was developed for generating *simultaneously accurate and complete* recommendations, called *Context Ultra-Sensitive Approach based on two-step Recommender systems (CUSA-2-step-Rec)* (Nasraoui and Pavuluri, 2004).

CUSA-2-step-Rec relies on a committee of profile-specific URL-predictor neural networks. This approach provides recommendations that are accurate and fast to train because only the URLs relevant to a specific profile are used to define the architecture of each network. Similar to the task of completing the missing pieces of a puzzle, each neural network is trained to predict the missing URLs of several complete ground-truth sessions from a given profile, given as input several incomplete subsessions. This is the first approach that, in a sense, *personalizes* the recommendation *modeling* process itself depending on the user profile.

(3) *Evolving User Interests* – Dealing with rapidly evolving user interests and highly dynamic websites requires a migration of the complete web usage mining phases from an offline framework to one that is completely online. This can only be accomplished with scalable single-pass evolving stream mining techniques (Nasraoui et al., 2003). Other researchers have also studied web usage from the perspective of evolving graphs (Desikan and Srivastava, 2004).

(4) *Data Collection and Preprocessing* – Preprocessing Web usage data is still imperfect, mainly due to the difficulty to identify users accurately in the absence of registration forms and cookies, and due to log requests that are missing because of caching. Some researchers (Berendt et al., 2001) have proposed clickstream path completion techniques that can correct problems of accesses that do not get recorded due to client caching.

(5) *Integrating Multiple Sources of Data* – Taking semantics into account can also enrich the Web personalization process in all its phases. A focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications (Li J. and Zaiane O., 2004). In particular, there has recently been an increasing interest in integrating web mining with ideas from the semantic web, leading to what is known as *semantic web mining* (Berendt et al., 2002).

(6) *Conceptual Modeling for Web usage Mining* – Conceptual modeling of the web mining and personalization process is also receiving more attention, as web mining becomes more mature, and also more complicated. Recent efforts in this direction include (Meo et al., 2004; Maier, 2004).

(7) *Privacy Concerns* – Finally privacy adds a whole new dimension to WWW personalization. In realty, many users dislike giving away personal information. Some may also be suspicious of websites that rely on cookies, and may even block cookies. In fact, even if a web user agrees to giving up personal information or accepting cookies, there is no guarantee that websites will not exchange this information without the user's consent.

Recently, the W3C (World Wide Web Consortium) has proposed recommendations for a standard, called *Platform for Privacy Preferences (P3P)*, that enables Websites to express their privacy practices in a format that can be retrieved and interpreted by client browsers. However, legal efforts are still needed to ensure that websites truly comply with their published privacy practices. For this reason, several research efforts (Agrawal and Srikant, 2000; Kargupta et al., 2003) have attempted to protect privacy by masking the user data using several methods such as randomization, that will modify the input data, yet without significantly altering the results of data mining. The use of these techniques within the context of Web mining is still open for future research.

FUTURE TRENDS

The Web is an incubator for a large spectrum of applications involving user interaction. User preferences and expectations, together with usage patterns, form the basis for personalization. Enabling technologies include data mining, preprocessing, sequence discovery, real time processing, scalable warehousing, document classification, user modeling and quality evaluation models. As websites become larger, more competitive, and more dynamic, and as users become more numerous, and more demanding, and as their interests evolve, there is a crucial need for research that targets the above enabling technologies, and leads them toward the path of scalable, real-time, online, accurate, and truly adaptive performance.

From another perspective, the inherent and increasing heterogeneity of the Web has required Web-based applications to integrate a variety of types of data from a variety of channels and sources. 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. Table 2 summarizes most of the active areas of future efforts that target the challenges that have been discussed in the previous section.

Table 2: Projected Future Focus Efforts in Web Personalization

- Scalability in the face of huge access volumes
- Accuracy of Recommendations
- Dealing with rapidly changing usage access patterns
- Reliable data collection and preprocssing
- Taking Semantics into account

- Systematic conceptual modeling of the web usage mining and personalization process
- Adhering to privacy standards

CONCLUSION

Because of the explosive proliferation of the Web, Web personalization has recently gained a big share of attention, and significant strides have already been accomplished to achieve WWW personalization while facing tough challenges. However, even in this slowly maturing area, some newly identified challenges beg for increased efforts in developing scalable and accurate web mining and personalization models that can stand up to huge, possibly noisy, and highly dynamic web activity data. Along with some crucial challenges, we have also pointed to some possible future direction in the area of WWW personalization.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support of the National Science Foundation CAREER Award IIS-0133948

REFERENCES

Agrawal R. and Srikant R. (2000). Privacy-preserving data mining, In *Proc. of the ACM SIGMOD Conference on Management of Data*, Dallas, Texas, 439-450.

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.

Berendt B., Hotho A., and Stumme G. (2002). Towards semantic web mining. In *Proc. International Semantic Web Conference (ISWC02).*

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.

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

Joachims T. (2002). Optimizing search engines using clickthrough data. In *Proc. of the 8th ACM SIGKDD Conference*, 133-142.

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.

Li J. and Zaiane O. (2004), Using Distinctive Information Channels for a Mission-based Web-Recommender System. 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.

Linden G., Smith B., and York J. (2003). *Amazon.com* Recommendations Item-to-item collaborative filtering, *IEEE Internet Computing*, 7(1), 76-80.

Maier T. (2004). A Formal Model of the ETL Process for OLAP-Based Web Usage Analysis. 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.

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.

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

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.

Nasraoui O., Krishnapuram R., and Joshi A. (1999). Mining Web Access Logs Using a Relational Clustering Algorithm Based on a Robust Estimator, *8th International World Wide Web Conference*, Toronto, 40-41.

Nasraoui O., Krishnapuram R., Joshi A., and Kamdar T. (2002). Automatic Web User Profiling and Personalization using Robust Fuzzy Relational Clustering, in "*E-Commerce and Intelligent Methods" in the series "Studies in Fuzziness and Soft Computing*", J. Segovia, P. Szczepaniak, and M. Niedzwiedzinski, Ed, Springer-Verlag.

Nasraoui O., Cardona C., Rojas C., and Gonzalez F. (2003). Mining Evolving User Profiles in Noisy Web Clickstream Data with a Scalable Immune System Clustering Algorithm, in *Proc. of WebKDD 2003 – KDD Workshop on Web mining as a Premise to Effective and Intelligent Web Applications*, Washington DC, 71-81.

Nasraoui O. and Pavuluri M. (2004). Complete this Puzzle: A Connectionist Approach to Accurate Web Recommendations based on a Committee of Predictors. 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.

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

Spiliopoulou M. and Faulstich L. C. (1999). WUM: A Web utilization Miner, in *Proc. of EDBT workshop WebDB98,* Valencia, Spain.

Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000). Web usage mining: Discovery and applications of usage patterns from web data, *SIGKDD Explorations*, 1(2), 12-23.

TERMS AND DEFINITIONS

Clickstream: Virtual trail left by a user's computer as she surfs the Internet. The Clickstream is a record of every web site visited by a user, how long they spend on each page and in what order the pages are viewed in. It is frequently recorded in Web server logs.

Collaborative Filtering: Collaborative filtering is a method for making automatic predictions (filtering) about the interests of a user by collecting ratings and interest information from many users (collaborating).

Cookie: A message generated and sent by a web server to a web browser after a page has been requested from the server. The browser stores this cookie in a text file, and this cookie is then sent back to the server each time a web page is requested from the server.

Frequent itemset: A set of items, e.g. {A, B, C} that simultaneously co-occur with high frequency in a set of transactions. This is a pre-requisite to finding *association rules* of the form, e.g. {A, B} ➔ C. When items are URLs or products (books, movies, etc) sold or provided on a Website, frequent itemsets can correspond to implicit collaborative user profiles.

IP address: (Internet Protocol address). A unique number consisting of 4 parts separated by dots, such as 145.223.105.5. Every machine on the Internet has a unique IP address.

Recommender system: A system that recommends certain information or suggests strategies users might follow to achieve certain goals

Web client: A software program (browser) that is used to contact and obtain data from a Server software program on another computer (the server).

Web server: A computer running special server software (e.g. Apache), assigned an IP address, and connected to the Internet so that it can provide documents via the World Wide Web.

Web server log: Each time a user looks at a page on a Website, a request is sent from their client computer to the server. These requests are for files (HTML pages, graphic elements or scripts). The log file is a record of these requests.

CGI program: (Common Gateway Interface) A small program that handles input and output from a Web Server. Often used for handling forms input or database queries, IT cam also be used to generate dynamic Web content. Other options include JSP (Java Server Pages), and ASP (Active Server Pages), scripting languages allowing the insertion of server executable scripts in HTML pages, and PHP, a scripting language used to create dynamic web pages.