

World Wide Web Personalization

Olfa Nasraoui

University of Louisville, USA

INTRODUCTION

The Web information age has brought a dramatic increase in the sheer amount of information (Web content), in the access to this information (Web usage), and in the intricate complexities governing the relationships within this information (Web structure). Hence, not surprisingly, information overload when searching and browsing the World Wide Web (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 Web site, 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 (e.g., grocery stores or libraries) to virtual stores of products or information (e.g., 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 knows the products and the 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 1990s,

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, p.). 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, 2002).

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 used later to recognize and greet the returning user. It usually is implemented on the Web server. This mode depends more on Web technology than on any kind of adaptive or intelligent learning. It also can 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 usually is 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 Web site (Mobasher

Table 1. Possible goals of Web personalization

- | |
|--|
| <ul style="list-style-type: none"> • 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 Web site • Making results of information retrieval/search more aware of the context and user interests |
|--|

et al., 2000; Nasraoui et al., 2002; Schafer et al., 1999). It usually is implemented on the Web server and relies on data that reflect 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—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 (Mobasher et al., 2000; Schafer et al., 1999):

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 (e.g., e-commerce transactions) also can be used. In some cases, Web content, structure, and application data can be added as additional sources of data in order to shed more light on the next stages.
2. **Preprocessing of Web Data:** Data is frequently preprocessed 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 (e.g., 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 (Nasraoui et al., 1999; Spiliopoulou & Faulstich, 1999; Srivastava et al., 2000), this step applies machine learning or data-mining techniques in order 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 off-line so that it does not add a burden to 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 & Vazirgiannis, 2003):

1. **Content Data:** Text, images, and so forth 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 Web site, including time of visit, IP address, and so forth. This data typically is recorded in Web server logs, but it also can 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 either can be static or dynamic. They also can be individualized (one per user) or aggregate (summarize several similar users in a given group).

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/world-wide-web-personalization/10787

Related Content

Group Pattern Discovery Systems for Multiple Data Sources

Shichao Zhang and Chengqi Zhang (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 546-549).
www.irma-international.org/chapter/group-pattern-discovery-systems-multiple/10657

Differential Association Rules: Understanding Annotations in Protein Interaction Networks

Christopher Besemann, Anne Denton, Ajay Yekkirala, Ron Hutchison and Anderson Marc (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1747-1758).
www.irma-international.org/chapter/differential-association-rules/7729

Factor Analysis in Data Mining

Zu-Hsu Lee, Richard L. Peterson, Chen-Fu Chien and Ruben Xing (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 498-502).
www.irma-international.org/chapter/factor-analysis-data-mining/10648

Formal Concept Analysis Based Clustering

Jamil M. Saquer (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 514-518).
www.irma-international.org/chapter/formal-concept-analysis-based-clustering/10651

Data Mining in Diabetes Diagnosis and Detection

Indranil Bose (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1817-1824).
www.irma-international.org/chapter/data-mining-diabetes-diagnosis-detection/7734