

Data Mining based User Modeling Systems for Web Personalization applied to people with disabilities

Julio Abascal

Dept. of Computer Architecture and Technology
University of the Basque Country
Donostia
julio.abascal@ehu.es

Olatz Arbelaiz

Dept. of Computer Architecture and Technology
University of the Basque Country
Donostia
olatz.arbelaiz@ehu.es

Javier Muguerza

Dept. of Computer Architecture and Technology
University of the Basque Country
Donostia
j.muguerza@ehu.es

Iñigo Perona

Dept. of Computer Architecture and Technology
University of the Basque Country
Donostia
inigo.perona@ehu.es

Abstract

This position paper tackles the problem of automatic web personalization using machine learning techniques to model the users' behavior. The target population is people with physical, sensory or cognitive restrictions. In this paper we present our plans to study the possibility of creating user models using the information extracted from web navigation logs by means of data mining methods. We discuss the expected advantages of adopting data mining to generate information about the user, in comparison with traditional methods.

1 Introduction

The main objective of Web personalization is to adapt the browsing or navigation, the presentation and the contained information of the web pages to the needs of the user without him or her doing an explicit demand. The goal is to enhance the speed and the achievement using the web and to decrease the physical and cognitive effort to perform it. Besides, this adaptation becomes especially critical when the users have special needs. In this context, in order to make the web automatically adaptive it is compulsory to build a model of the user that aggregates its main characteristics, and to use this model to perform actions that make easier its navigation in the Web. These user models are usually built based on the analysis of their browsing history and this implies the need to work with log files containing large amount of data. Therefore the use of data mining techniques is very adequate in this context.

Most of the works done in the Web personalization context mainly focus on the improvement or simplification of three aspects: navigation, presentation or layout of the shown data (typography, colors, where to locate the information, compact link lists...); and content of the shown web pages.

In order to the Web being adaptable in the three mentioned contexts, it is compulsory to first obtain user models that permit the characterization of their different needs. These models could be directly designed by experts in the area (rule based approach), they can be built based on previous information from that user such as the logs of previous navigations (content based approach) or they can be induced from information about groups of users with similar characteristics (collaborative approach). The first approach is somehow static and requires previous knowledge of the users and redesigning when new behaviors appear in the users. However, the last two approaches focus on automatic techniques for user characterization. Our aim is to focus on these

two approaches and to combine them to find a trade-off between the high specialization of the content-based approaches and the computationally too expensive collaborative approach.

Section 2 describes the basis of the adaptive web and Section 3 is devoted to describe the use of Data Mining techniques in this context.

2 The Adaptive Web

Web personalization is based on modeling user features such as interests, navigational behavior, preferences, physical sensory or cognitive restrictions, etc. The information stored in the model is used to make assumptions about the current user-system interaction that allows adapting the system to the actual user needs or preferences. User adaptation methods have been frequently adopted by intelligent interface designers to adjust the interface to the user (contrarily to the usual situation where the user adapts him or herself to the interface).

Adaptive systems are usually composed of a modeling component where the information about the user is stored and processed. The model is composed of relevant and observable parameters that allow creating profiles and stereotypes that are used to make assumptions about the current state of the user. In this way the adaptive system can dynamically adapt the interaction to the evolving conditions of the user. Currently the user component is usually built by means of ontologies that allow to store, manipulate and extract assumptions from data about the user, its context, tasks, etc.

Web personalization seeks for the adaptation to the user in three areas: [10]:

1. Navigation. Browsing pages is the most common task when using the web. This task can be slowed, for instance, if the page contains a high number of links that are not interesting for the user. Knowing the objectives and interests of the user the browser can make easier the navigation task giving more emphasis to the most interesting or probable links for a specific user. The inclusion of a specific navigation menu with selected links is another navigation facilitating possibility.
2. Presentation: The presentation of a web page can be adapted to the specific needs of each user applying cascading style sheets (CSS). The most convenient style sheet(s) may be stored in the (static) model of the user.
3. Content. Even if it is not convenient to automatically change the content of a web page, some inclusions may enhance its readability. For instance, the automatic inclusion of text captions in simple language used to explain longer and more difficult texts are useful for people with reading difficulties.

In the case of users with disabilities, adaptation becomes crucial and may very much speed up the navigation,

In order to be able to model the user, the modeling component must collect information about a number of observable parameters such as interest, characteristics, etc. This information can be requested to the user in a previous session, but this is annoying, disruptive and can produce false assumptions. Another option is to collect this information while the user is accessing the web. In this way the system can learn its interests, likes, etc.

Data mining for web personalization has many advantages. It is not disruptive, is based in statistical data obtained by real navigation exercises (decreasing the possibility of false assumptions) and is itself adaptive (when the characteristics of the user change, collected data allows the automatic change of the interaction schema). When the user is a person with physical, sensory or cognitive restrictions, data mining is the easiest (and frequently almost the only) way to obtain information about the uses of the person.

Data mining in this context has also some drawbacks. The most important one is its impact over privacy, due to the need of storing large quantities of data about the users. Diverse laws in different countries protect user rights for privacy. Even if it is difficult to reach a balance among privacy and personalization, some appealing proposals have been recently published.

As previously mentioned, learning from the own interaction allows maintaining a dynamic profile of the user, avoiding the application of all assumptions when the interest, characteristics or circumstances of the user change.

3 Data Mining

The use of data mining techniques can be understood as the data processing to extract knowledge from it. Nowadays, the amount of data stored magnetically makes these kinds of techniques essential. In this context, the research group Aldapa, has great experience solving different real world problems such as: automatic character recognition, fraud detection, customer fidelity, intrusion detection, etc. To solve these problems we used adequate machine learning techniques adapted to different restrictions such as: large data volumes, combinatorial explosion, real time, uncertainty in the data, etc.

The aim in this work is to use our data mining experience in the context of Web personalization for people with special requirements or people with disabilities. We propose the use of data mining to model the browsing/needs of the users automatically and not manually as it is done in a customization process. We will base this modeling work on a large amount stored information such as log files... that are stored in servers, proxies, desktops, etc., as a consequence of previous browsing tasks performed by the users. Depending on the used techniques, we could use the models achieved by data mining techniques to dynamically enrich the ontologies used to build user components or directly as user models by themselves.

The treatment of the mentioned information requires a complete data mining process including all its steps: data collection, data preprocessing (retrieving non useful information, generating aggregate features, derived features, etc.), selecting the most adequate machine learning technique and the application of the selected techniques to finally obtain knowledge. Furthermore, there is a need of validation of the achieved results and a feedback to repeat the process with more knowledge in the case it is needed.

We plan to confront the machine learning problem in two phases. The first step will consist on generating the models of the users. We propose to use unsupervised learning techniques or clustering techniques in this first step that will provide as output sets of users with similar characteristics or needs. The second phase would be the exploitation phase where a user navigating in the web needs to be matched with one of the previously generated profiles based on a supervised classification technique. This way, the navigation of the user will be personalized based on wider information related to the set of users belonging to the cluster (collaborative approach).

In both phases, the unsupervised part and the supervised one we will need to define or select a distance metric that is suitable for the characteristics of the data. We will need to analyze different kinds of distances depending on the nature of the data. If the problem is instantiated in a determined dimensional space we can use distances such as: Manhattan, Euclidean distance, cosine similarity measure [7], etc. In contrast, if the data are represented using sequences (click streams inside a web, visited web pages, etc.) we will need to use other kinds of distances such as: Edit distance [2], Normalized Compression Distance [4], etc.

Based on the nature of the data (vector like, sequence like) and the selected distances for the first phase we will analyze different clustering techniques we used previously in other contexts such as: SAHN [8], Fixed-width or Leader algorithm [9], k-means [3], etc.

On the other hand, we propose different strategies to establish the profiles related to each cluster or group of similar users:

- the use of paradigms such as association rules [6] or frequent episodes [5] so that we can predict the most probable transitions between links (and similarly for Item-based approaches).
- the use of meta-learning techniques based for example in classification trees to profile the built clusters.

We propose to complement the results achieved in the collaborative approach using the same kind of techniques mentioned in the preceding paragraphs from a content-based approach point of view.

Finally, the validation of the obtained results is a key issue. It is necessary to check if the user models built in this way are effective to model the user and allow a valid adaptation. To this end it is necessary to design formal experiments involving real users navigating the web. These experiments have to compare the adaptation by classical methods with the results obtained by means of data mining, using static interaction as a control. In order to study its temporal behavior, these experiments may require observation and logging of information of web use for large periods (with the explicit consent of the user).

References

- [1] Brusilowsky, P., Kobsa, A. & Nejd, W. (2007) *The Adaptive Web. Methods and Strategies of Web Personalization*. Berlin-Heildeberg: Springer-Verlag.
- [2] Gusfield D. (1997) *Algorithms on strings, trees, and sequences*. Cambridge University Press.
- [3] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.
- [4] Li M., Chen X., Li X., Ma B., Vitanyi P.M.B. (2004) The similarity metric. *IEEE Transactions on Information Theory* 50, 3250–3264
- [5] Mannila H., Toivonen H., Verkamo A. I. (1997) *Discovery of Frequent Episodes in Event Sequences*. *Data Mining and Knowledge Discovery* 1, 259–289. Kluwer Academic Publishers
- [6] Piatetsky-Shapiro, G. (1991) *Discovery, analysis, and presentation of strong rules*, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA.
- [7] Salton G. (1989) *Automatic Text Processing*. Addison-Wesley.
- [8] Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*. Books in biology. W. H. Freeman and Company.
- [9] Spath H. (1980) *Cluster analysis algorithms*. Ellis Horwood, Chichester, UK.
- [10] Brusilowsky P. Adaptive Navigation Support (2007). in *The Adaptive Web. Methods and Strategies of Web Personalization*. Berlin-Heildeberg: Springer-Verlag. 263-290