PNS: Personalized Multi-source News Delivery

Georgios Paliouras[1], Mouzakidis Alexandros[1], Christos Ntoutsis[2], Angelos Alexopoulos[3], and Christos Skourlas[2]

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
{paliourg, alexm}@iit.demokritos.gr
[2] Department of Informatics ,Technological Institute of Athens, Greece
[3] Department of Informatics and Telecommunications, University of Athens, Greece

Abstract. This paper presents a system that integrates news from multiple sources on the Web and delivers in a personalized fashion to the reader. The presented service integrates automatic information extraction from various news sources and presentation of information according to the user's interests. The system consists of source-specific information extraction programs (wrappers) that extract highlights of news items from the various sources, organize them according to pre-defined news categories and present them to the user through a personal Web-based interface. Dynamic personalization is used based on the user's reading history, as well as the preferences of other similar users. User models are maintained by statistical analysis and machine learning algorithms. Results of an initial user study have confirmed the value of the service and indicated ways in which it should be improved.

Keywords: Personalization, Information Extraction, Machine Learning.

1 Introduction

The rapid increase of interest for the Internet stems from the fact that it is very easy to access and publish information on the Web. At the end of 1993 the community of the World Wide Web numbered about 300,000 users, the majority of which were university researchers and large IT companies. Today it is estimated that there are over 1 billion users on the Internet, most of which cannot be considered computer experts. The increase in the number of users does not only cause an increase of information, but also increases the variety of information that is available (music, news, products, movies, services, etc.) so that more and more users find it useful to surf the Web and contribute to its expansion. This phenomenon is responsible for the "explosion" of the Web, which leads to the *"information overload"* of Web users. Moreover, the large number of Web users and the globalization of the economy led businesses to offer e-services such as e-commerce, e-learning and e-papers. The increased competition and the need for successful services obliged the businesses to add value to e-services in order to create loyal visitors – customers.

To realize this added value and to face the reality of information overload personalization techniques have been developed. Web personalization is simply defined as the process of making Web-based information systems adaptive to the

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part II, LNAI 4252, pp. 1152 – 1161, 2006.

needs and interests of individual users. Typically this concerns data collection about the users, analysis of these data, and retrieval of the suitable data for the specific user at the suitable time [1].

Users have their own preferences about news content and news sources. Databases of news sources change continuously, making it impossible for users to identify and follow every single news item that is published and could be of interest to them. Therefore, news delivery on the Web is one of the most typical services where personalization can add significant value. Nowadays all large news media (newspapers and news channels) offer in some grade personalization services.

However, the personalization of individual news sources is not a sufficient solution to the problem of information overload, as the users still have to visit many different sites, in order to keep up-to-date with current news. Therefore, an integrated service that aggregates information from various sources and presents it to the users according to their own preferences is very desirable. This is the kind of service by PNS, the system presented in this paper. PNS includes information extraction programs (wrappers), which retrieve continuously highlights of new items that appear at various news sources. This information is organized according to predefined news categories and presented to the users through a Web-based interface. Personalization is achieved with the use of a separate personalization server that provides a variety of services. PNS makes use of four types of adaptive personalization: (a) personal user statistics, (b) stereotype modeling, (c) community modeling, (d) news itemsets. Each of the four types requires the acquisition and maintenance of a different user model, which is achieved with the use of statistical analysis and machine learning methods.

The rest of this paper is structured as follows. Section 2, reviews the state-of-the-art systems for news personalization. Section 3 briefly describes the design and implementation of PNS. Section 4 presents the results of an initial user study, while in the last section conclusions and future directions are discussed.

2 Related Work

A wide variety of both research prototypes and commercial systems offer personalized news on the Web. The goals of these systems are [2]:

- *Personalized news presentation*: The system can provide personalized news, tailored to individual preferences.
- *Personalized advertisements*: The system publishes advertisements targeted to individual groups of people.
- *Effective search capability*: It is possible to search for news items related to a given topic by providing meaning to some keywords.

For successful Web personalization the following three steps are important [1,3]:

1. Gathering of useful information about the user and his interests. The collection of data is performed explicitly, through form-filling, and/or implicitly, through the logging of usage data, possibly combined with legacy data.
2. Creating user models. The collected data are processed and interesting patterns are discovered. Users are clustered and modeled according to their interests.

Non-adaptive user models are predefined and cannot change even if the user interests have changed. Machine learning methods are used to create adaptive user models that capture changes in the user's interests.

3. News filtering/ranking. News articles to be presented are chosen, together with the order of presentation. When the filtering is based on the content of the articles then it is called content-based filtering. Due to the high cost of data preprocessing and analysis, an alternative (or complementary) personalization technique that can be used is collaborative filtering, where the system groups the users into communities according to common characteristics and reading interests.

Table 1 summarizes some well-known personalized news systems with a small description of the algorithms that are used for personalization.

Table 1. Summary of Web news personalization systems, according to their features

System	*Data Collection*	*User modeling*	*Filtering*
Personal Wall Street Journal, San Francisco Chronicle, Fishwrap [4]	Explicit input, legacy data	Non-adaptive	Content-based
Krakatoa, Anatagonomy [5]	Explicit and implicit input	Adaptive	Content-based, Collaborative
SmartPush [6]	Explicit input	Non-adaptive	Content-based
Newsweeder [7]	Explicit and implicit input	Adaptive	Content-based, Collaborative
Aggrawal and Yu [8]	Explicit and implicit input	Adaptive	Collaborative
WebMate [9]	Explicit and implicit input	Adaptive	Content-based
NewsDude [10]	Explicit input	Adaptive	Content-based
Findory (http://www.findory.com)	Implicit input	Adaptive	Content-based
Google (http://news.google.com), Yahoo (http://dailynews.yahoo.com)	Explicit input	Non-adaptive	Content-based
Newsjunkie [11]	Explicit and implicit input	Adaptive	Content-based

3 Personalized News Service Description

The Personalized News Service (PNS) provides users with personalized access to news items from multiple Web sources. For user modeling, the system makes use of a generic Personalization Server (PServer)[1]. In comparison to the existing systems this service has the following characteristics:

1. *Data collection:* The system collects both explicit (optional) user data, through a registration form, and implicit data through usage logging.

[1] PServer has been developed in the Institute of Informatics and Telecommunications of NCSR "Demokritos" and will soon be made available under a BSD-like license.

2. *User modeling:* The system is highly adaptive with the use of PServer.
3. *Filtering:* The system uses content-based and collaborative filtering.

The key feature of the proposed system that differs from many existing ones is that it aggregates news highlights from multiple Web sources. News integration is achieved through a combination of source-specific information extraction (wrappers) and RSS input. To our knowledge, PNS is the first system that provides highly adaptive personalization together with multi-source news integration.

3.1 Description of the PNS Content Server

The PNS Content Server is the main server of the system that scans news sources, at regular time intervals. The output of the server is a personalized electronic newspaper consisting of recent article titles that match the interests of the user.

Figure 1 shows the overall architecture of the system, which is made up of three main units: a) Content Scanner, b) Content Selector, c) Content Presenter, as well as the Content Index Database where highlights about the news items and the wrappers are stored. Respecting the copyright of the sources, the server does not store the content of the articles, but simply indexes it, according to its own categorization.

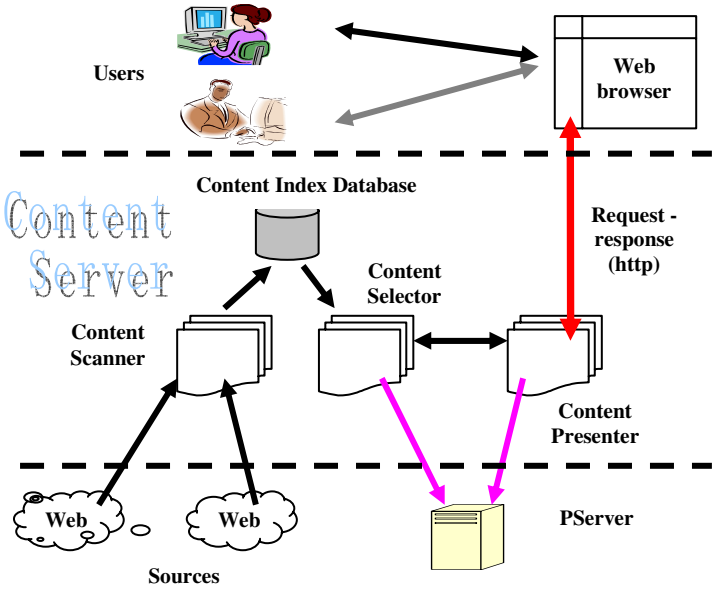

Fig. 1. Overall architecture of the personalized news delivery service

The system collects information about users in two ways:

1. A username and password are specified for logging and memorization purposes. During the registration, the user can also provide personal information, such as age, gender, occupation, etc. for improved personalization.
2. The browsing activity of users is tracked and stored for personalization purposes.

The user's input and choices are forwarded to the PServer from the Presenter. Additionally, the Presenter supplies the registration and identification information to the PServer. This information is used for maintaining the user models. The three component modules of the system are described in more detail below.

3.2 Content Scanner

Content Scanner is an autonomous module that retrieves information about news items from a range of content sources at specified time intervals, using information extraction techniques. This is a typical web *information integration* system that extracts and combines data from multiple web sources. Content Scanner follows the local-as-view approach where, for every information source a specific wrapper is used to extract the desired information. Following this approach, it is simple to add or delete sources and it is also easier to describe constraints on the contents of the sources. Furthermore, the Scanner obtains news through RSS feeds. The extracted information is stored in the Content Index Database. The input for the Content Scanner is a set of news sources associated with a set of wrappers that are used to identify and extract the relevant information.

Figure 2 shows the architecture of Content Scanner. Because of the hierarchical structure of the content sources, the extraction of the relevant news is performed in two levels, thus using two levels of wrappers. The wrapper of the first level (source wrapper) extracts from the main page of a news source (e.g. yahoo), the URL address of the most recent articles. This URL is used by the content wrapper, which extracts the news highlights and stores them in the database. The wrappers are source-specific HTML patterns. For example, in order to extract the titles of articles in a specific source the TITLE tag was used (start pattern: <TITLE>, end pattern: </TITLE>). The wrappers are stored in the Content Index Database and can thus easily be maintained.

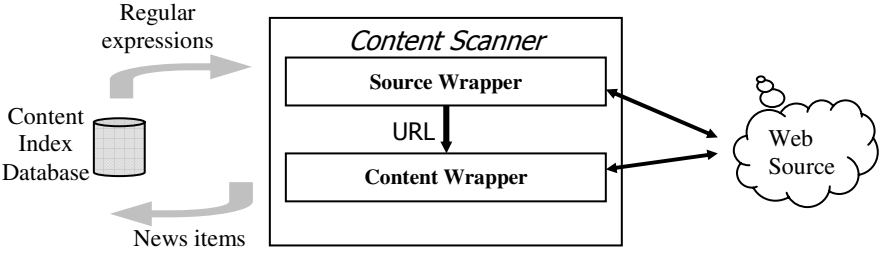

Fig. 2. Architecture of Content Scanner

The Content Scanner can easily be extended so to extract news highlights from additional Web sources. This can be done by defining new wrappers for the additional sources and inserting the wrappers into the Content Index Database.

3.3 Content Selector

The Content Selector chooses the recent content from the Content Index Database to be used for the composition of the personalized newspaper for a particular user. For

the selection both content–based filtering and collaborative filtering are used. In the first case, news is classified along two orthogonal dimensions, the category (e.g. Sports) and the source (e.g. yahoo). For example, a user may prefer to read financial and sports news, while another might be interested specifically in the world news of yahoo. PNS also supports various types of collaborative filtering: (a) according to personal characteristics, optionally provided by them, the users are assigned to a stereotype, the model of which is dynamically maintained based on usage data, (b) users are clustered into communities according to their common preferences alone, (c) news are clustered into news itemsets according to the usage data.

3.4 Content Presenter

The Content Presenter module is the user interface of PNS. All services are available through this unit, which is responsible for the following tasks:

- Registration of new users.
- Identification of registered users.
- Presentation of the daily news that exist in the Content Index Database according to the user's preferences (personal e-paper).
- Presentation of the daily news according to the preferences of users that belong to the same stereotype.
- Presentation of the daily news according to the preferences of users that belong to the same user community.
- Presentation of the daily news that belong to the same news itemset as the currently viewed item.
- Personalized presentation of the news of previous days.
- Text -search and presentation of news titles using keywords.
- Ability to retrieve news highlight in specific dates from the database.

3.5 Personalization Server

The Personalization Server (Pserver) is a general purpose personalization server, using a feature-based representation of user models. Pserver constructs and maintains models for individual users, stereotypes, user communities and feature groups. For PNS the features of users are the sources and categories of news articles. User models are used by the Content Selector to personalize the content that is presented to the users. Each personal user model may contain the following information: (a) personal information about the users, as provided during the registration and (b) sources and categories of news articles with a weight parameter which is based on the frequency at which the user chooses the particular source or category. Stereotypes are similar to personal user models, but they accumulate frequency statistics for all users with the same personal characteristics. User communities are also aggregate models, but they do not contain personal information about the users. Finally feature groups are orthogonal to the user communities in that they aggregate statistics about features, rather than users.

User communities and feature groups are not predefined, but are constructed with the use of machine learning algorithms. Pserver's architecture supports the usage of

alternative machine learning algorithms for that purpose. One such example is Cluster Mining [12], which discovers patterns of common behavior by looking for all fully connected sub-graphs (cliques) of a graph that represents the user's characteristic attributes. It starts by constructing a weighted graph $G(A,E,W_A,W_E)$. In order to construct feature groups, the set of vertices A corresponds to the features used in the user models, and the set of edges E corresponds to feature co-occurrence as observed in the models. For instance, in our application that we examine, if the user reads economy news from BBC and ANT1 an edge is added between the relevant vertices. The weights on the vertices W_A and the edges W_E are computed as aggregate usage statistics. An example graph is shown in Figure 3.

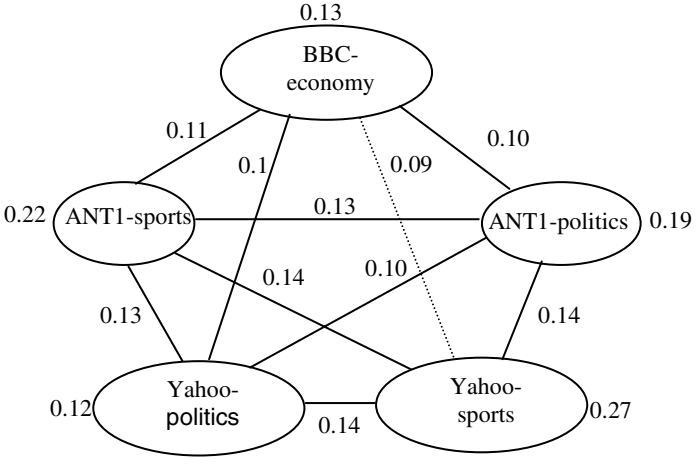

Fig. 3. Feature graph for cluster mining

The connectivity of the graph is usually very high. For this reason we make use of a *connectivity threshold* aiming to reduce the edges of the graph. In our example in Figure 2, if the threshold equals 0.1 the edge ("BBC-economy", "yahoo-sports") is dropped. Cliques are then found on the reduced graph. In addition to the construction of feature groups, this algorithm can be used to construct user communities, by placing users at the vertices of the graph. However, PServer supports the use of any other clustering algorithm for that purpose, too. The administrator of the system should specify the frequency at which communities and feature groups will be updated. The communication between PServer and the Content Server is done by simple HTTP requests and replies.

4 User Study

In order to evaluate PNS, users of different background were asked to test the system for a short period of time. On a daily basis, the system collected the most recent news, which were then presented to the users. The users were asked to fill an electronic

questionnaire with their observations. The role of the user study was to gather information in several different areas, with the general aims of:

- Validating the personalization services.
- Evaluating the functionality of the system.
- Providing input into the design of the system.

Table 2 presents the most important subjects that were evaluated, followed by the results according to the answers supplied by the users. The results obtained from this initial user study confirm the added value provided by the service and point out a number of interesting improvements. The most important subject is the enhancement of the system with more new sources and categories, which will make the system more interesting for the users and the added value of personalization more clear.

Table 2. Summary of the results of the user study

Subject tested	Positive	Partly	Negative
Satisfaction with the order that news highlights are presented	55%	35%	10%
Satisfaction with the news presentation	70%	25%	5%
Satisfaction with the number of news categories	20%	45%	35%
Satisfaction with the number of news sources	10%	50%	40%
Feeling of ease in searching for news	70%	20%	10%
Satisfaction from the interface	70%	30%	0%

5 Conclusions and Future Work

News personalization is an emerging technology that serves both users and businesses. Despite the wide-adoption of personalization by various news sources on the Web, there is still a need for an integrated service that will aggregate information from multiple sources and present the results to the user in a personalized manner. Our service (PNS) uses source-specific information extraction programs to retrieve highlights of news articles and organize the extracted information according to predefined news categories. Using a Personalization Server (PServer), the system provides a personalized view of the collected news items through a Web interface. Dynamic personalization techniques are used for that purpose, analyzing both the user's own interaction with the system, as well as the preferences of similar users. The results of an initial user study have confirmed the added value provided by the service and have pointed to a number of interesting extensions.

The most highly demanded extension is the increase of the coverage of the system with new sources. The extension of the news retrieval module (Content Scanner) is also very important. Frequent changes of the structure of the news sources require the manual update of the wrappers, which is a time-consuming process. Wrapper

induction and verification techniques [13, 14] can be used to automatically update the wrappers.

Another important issue is to create a module that detects sudden changes in news trends and create new topics and categories. For example, during the war in Iraq, CNN dedicated a special page with news from the war. The proposed system could not extract any information about the war, since "war" isn't a predefined topic for extraction.

Concluding, news aggregation and dynamic personalization from various sources is a promising, highly requested application, which also leads to a variety of interesting research challenges. The service presented in this paper will serve as a base for the development of further innovative applications and services.

Acknowledgements

The presented work is part of a long-term project of the Software and Knowledge Engineering Laboratory at the Institute of Informatics and Telecommunications of NCSR "Demokritos". Part of this work was done in collaboration with the Department of Informatics of the Technological Institute of Athens, in the context of the research project PA_CO_CLIR (Parallel, COntent Based Cross Language Information Retrieval) that is co-funded by the European Social Fund and National Resources (EPEAEK-II)-ARXIMHDHS.

References

[1] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, 2003, Web Usage Mining as a Tool for Personalization: A Survey, *User Modeling and User-Adapted Interaction*, v. 13, n. 4, pp. 311-372.

[2] L. Ardissono, L. Console, and I. Torre 2000, On the application of personalization techniques to news servers on the Web, *Lecture Notes in Computer Science,* Torino, Italy, pp. 1-12.

[3] A. Kobsa, J. Koenemann, and W. Pohl. 2001, Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *The Knowledge Engineering Review* 16 (2), pp. 111-155.

[4] P.R. Chesnais, M.J. Muckle, and J.A. Sheena. 1995, The fishwrap personalized news system. In *Proceedings IEEE 2nd Intl Workshop on Community Networking Integrating Multimedia Services to the Home*, Princeton, New Jersey, USA.

[5] T. Kamba, K. Bharat and M.C. Albers. 1995, The Krakatoa Chronicle - an interactive personalized newspaper on the Web In *Proceedings 4th Intl WWW* Conference, p. 159-170.

[6] T. Kurki, S. Jokela, R. Sulonen and M. Turpeinen, 1999, Agents in delivering personalized content based on semantic metadata In *Proceedings 1999 AAAI Spring Symposium Workshop on Inteligent Agents in Cyberspace,* p. 84-93.

[7] K. Lang. 1994, Newsweeder: An adaptive multi-user text filter. *Technical Report.* School of Computer Science, Carnegie Mellon University.

[8] C.C. Aggarwal and P.S. Yu, 2002, An Automated System for Web Portal Personalization, *Technical Report*, *IBM T. J. Watson Research Center Yorktown*, USA.

[9] L. Chen and K. Sycara, 1998, WebMate: A personal agent for browsing and searching. *In Proceedings of the Second International Conference on Autonomous Agents, Minneapolis*, p. 132-139.

[10] D. Billsus, and M. J. Pazzani, 1999, A Hybrid User Model for News Classification. In *Kay J. (ed.), UM99 User Modeling - Proceedings of the Seventh International Conference*, pp. 99-108. Springer-Verlag, Wien, New York, USA.

[11] E. Gabrilovich, S. Dumais, and E. Horvitz, 2004, Newsjunkie:Providing Personalized Newsfeeds via Analysis of Information Novelty, *InProceedings of the 13th international conference on World Wide Web,* New York,USA, pp.482-490

[12] G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos, 2000, Clustering the Users of Large Web Sites into Communities, *Proceedings of the International Conference on Machine Learning (ICML),* pp. 719-726, Stanford, California.

[13] N. Kushmerick, 2000, Wrapper Verification, *World Wide Web J.* **3**(2), pp.79-94, Special issue on Web Data Management.

[14] N. Kushmerick, 1997, Wrapper induction for information extraction, *PhD Thesis*, University of Washington.