Dynamic User Context Web Personalization in Meta-Portals

Christos Bouras, Prof. and Vassilis Poulopoulos, MsC
Research Academic Computer Technology Institute and
Computer Engineer and Informatics Department,
University of Patras
Rion, Greece, GR26504
bouras@cti.gr, poulop@cti.gr

***Abstract*—We present the dynamic web personalization and document grouping infrastructure for meta-portals and the evaluation of the mechanism on peRSSonal, a system that collects articles from news portals and blogs worldwide. A meta-portal is an informational node where articles from different sources are collected and presented in a categorized and personalized manner. The web personalization mechanism is based on dynamic creation and update of user profile according to the user's preferences when browsing and on user grouping details. Assuming that required information, such as article tagging, keywords to categories matching and articles to categories relation is already part of the meta-portal we present a novel mechanism that can build and maintain a user profile which is formed without disturbing the user. Furthermore, we describe the real-time user centered document grouping mechanism that is implemented to support the web personalization system and present the experimental evaluation of the whole system.**

***Keywords*-web personalization; on-line document grouping; user profiling; meta-portal;**

I. INTRODUCTION

The last decade can be inevitably referenced as the decade of dramatic changes to almost every aspect of our everyday life. The advances of technology are huge and the evolution of World Wide Web can be recognized as enormous. This weird freedom that the WWW offers, attracts more and more people. More attractive is the fact that people are free to produce WWW content in an extremelly easy way making thus the production of web content a trend. Online text production has never been so ordinary and internet users were converted from internet readers to internet authors. The WWW is a vast place of article production and it can be referenced without any doubt as a large newsletter.

The problem that arises from the fact that the Internet becomes a place where the sources (media) are more than the consumers (readers) is that the customers are usually unable to locate useful information. By useful information we define the information that a user would like to be presented. Searching across the internet through the wide variety of search engines can be considered as a possible solution to this problem, but the outrageous number of results is uninviting. Moreover, when searching for articles which contain breaking news it is impossible for a search engine to locate them. The search tools that exist within article's sources and the communication channels provided can be presented as a solution or even the ultimate solution; however, the user must "invent" these places before (s)he starts using these services. Creating customized and personalized sections within web pages is another viable solution but some recent examples seem to become misleading for the plethora of different types of users that exist on the web. Meta-portals is the trend and their usage is not recent. The point is that most of them let the users add their own communication channels assuming that the communication channel of the user includes important information about the user. User personalization and user profiling seems to be the panacea of the current chaotic web status.

Many efforts were presented in the latest years in order to solve the problem of user profiling within web sites or even across the internet. There is a slight but enormous difference between user profiling (or personalization) and customization of web sites. Customization is the capability that is provided to the user to alter the layout of the website that is watching; which is the color, the font, the position of the elements, the order of the information and others. In the context of the Web, personalization implies the delivery of dynamic content, such as textual elements, links, advertisement, product recommendations, etc., that are tailored to needs or interests of a particular user or a segment of users [4]. Personalization techniques [10] are an alternative, user-centric, approach to addressing the problem of information overload. The ultimate goal of any user-adaptive system is to provide users with what they need without them asking for it explicitly [5].

Some important efforts towards personalization can be found in [6] and [8] where it is obvious that for more than one decade the research community is trying to apply web personalization through data mining activities and generally heuristics while [1] present some of the first more "advanced" techniques of web personalization for the web2.0 that was born back in 2005. The approaches described in [7] and [9] are of high importance in the research literature on the issue as the first one introduces a cube model for knowledge extraction about the user's behavior and the second deals with usage patterns from web extracted

978-1-4244-7755-5/10/$26.00 ©2010 IEEE

data. Kim and Chan in [11] present a robust context for personalization based on UIH which is the user's interest hierarchy that is constructed with the usage of a tree model of the user profile. Other approaches like the ones presented in [12] and [13] that are applying personalized features either on portals or on search procedures by utilizing semantic information of the user are also interesting as they gather information from meta-data and not only direct information from the user. Evaluation of the user models learned from the data involves the estimation of the accuracy of the models for predicting content that may be interesting to a user as well as other aspects such as explainability of the recommendations, diversity of the recommendation set, serendipity of the recommendations, and user satisfaction [2]. Finally, it is important to have a reference on the ongoing discussion that is focused on the part of privacy and web personalization. It is a fact that some of the constructed mechanisms are utilizing private information which is obtained without the user's consent. Extended information about the ease of use of privacy and web personalization can be found in [14] where the formula for reconciling both is presented and analysed.

We present a novel mechanism for user profile construction and maintenance in meta-portals. Some worldwide known meta-portals are Yahoo [1] and Google news [2]. We enhance the operation of our meta-portal peRSSonal by providing dynamically changing user profiling features fully adapted on the user's needs and without need of any user input.

The rest of the paper is structured as follows. The next section presents the architecture of the system while in the third section a first algorithmic analysis is presented. The next section includes sets of experiments and the paper is finalized with future work and concluding remarks on the implemented system.

II. Architecture

The architecture of the system relies on distributed components which form the dynamic web user profiling system. We are putting the focus on the personalized profiling subsystem, though brief analysis of the other modules is presented in order to cross-connect the features of our complete system, peRSSonal[3].

The architectural schema consists of a series of subsystems, as depicted in figure 1. The collaboration between the distributed parts is based on the open standards (for input data and output data) and on the communication with a centralized database. The general procedure is as follows: at first, web pages are captured and only the useful text is extracted from them. Then, the extracted text is parsed followed by summarization and categorization. Finally we have the presentation of the personalized results to the end user.

[1]http://news.yahoo.com/ - News from Yahoo
[2]http://news.google.com/ - News from Google
[3]http://perssonal.cti.gr/perssonal/ - peRSSonal Meta-Portal

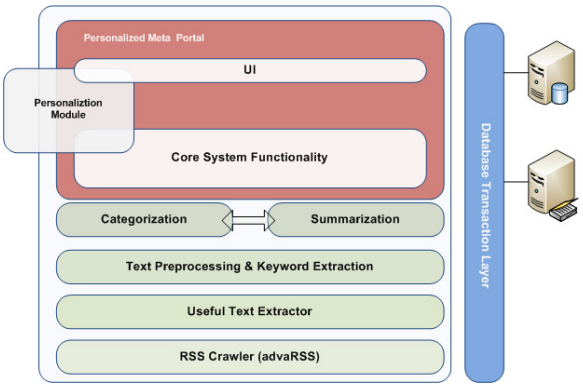

Figure 1. System Architecture

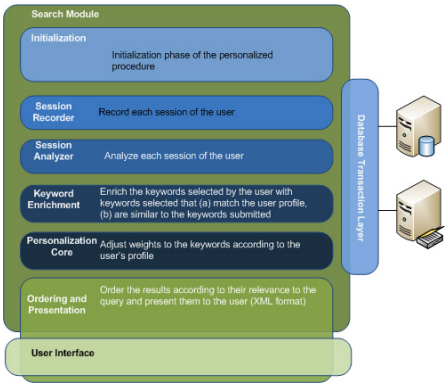

Figure 2. User Profiling Flow of Information

A. Flow of Information

In figure 2, we can see the general schema and flow of the advanced and personalized profiling system.

The personalization procedure of the portal that is supported as a medium of communication between all the procedures and the users can be used in order to personalize the summarization on each user. According to the algorithmic procedures of the personalized portal, the system creates a vector that represents the user's profile. To be more precise, each user has two vectors for his profile: a "positive" vector and a "negative" one. The positive vector represents semantically the interests of the user on the article content and the negative represents what is out of user's interest. The vectors are constructed from tables with keyword/value pairs. The value represents how much is the user interested or not interested in the specific keyword. According to the user's behavior when browsing the meta-portal the vectors are dynamically altered. The main factors that affect the user's behavior are: a. selecting or not selecting an article that is presented to the user, b. position in the page of article selected or not selected, c. time spending reading the

article in combination to the article's length, d. selection of similar articles, e. selecting to browse the original web page of an article, f. adding the article to the "starred" articles, g. utilizing the document-clustering service on an article (locate all the articles that are identical to the one that the user is reading), h. use of the tag the article service (add the article's tags into the user's profile), i. user the tracking service (inform the user about every identical article that will occur) and finally j. remove the article from the article's list.

B. Document Grouping

The system that we are presenting is utilizing the user's behavior in order to achieve enhanced document grouping. The document grouping procedure of the system leads to creating sets of articles that are identical. By identical, we define the articles that refer to exactly the same fact but have different sources. The document grouping procedure is a never ending procedure because articles occur every five minutes (execution time of our crawler) and the groups should be constructed and maintained simultaneously.

The basic idea of the document grouping procedure is the on-line creation of the document groups. When the user selects an article to read the system checks if a cluster exists for the specific article. In case the cluster exists then all the documents of the cluster are fetched and presented to the user as a single fact with different instances/sources. From the behavior of the article publishing procedure we assume that articles published with time difference greater than 16 hours cannot be identical. Still, the system can recognize such articles as relevant. This means that if the oldest article in a cluster is more than 16 hours old then the cluster is considered to be "closed".

III. Algorithm Analysis

The first procedure of the system is fetching articles from the internet in a periodical manner. The next procedure includes the steps of analyzing the HTML files, extracting the useful text from them and preprocessing the useful text in order to extract the article's keywords. The useful text extraction is based on the fact that HTML pages can be depicted as a tree with every HTML tag holding a node on the tree, while every leaf includes pure text. The preprocessing techniques and the keyword extraction techniques are described in [10]. Following we apply summarization and categorization procedures on the document. At this stage we assume that we have all the prerequisites in order to construct and maintain a user's profile. The user's profile is created in a single step and it is maintained while the user is using peRSSonal portal and utilizing the information and services presented to her/him.

A. Dynamic User Profile Creation and Maintenance

Some first information is obtained by the system when the user is registered in order to create an initial profile. This procedure is done through the web registraton procedure. During this procedure the user is asked to enter his preferences against the seven major categories of the portal (business, entertainment, health, politics, technology, education, science). The preference varies from -5 to +5 indicating total reject of the category to total accept respectively. For each of the categories a vector exists in the database constructed from pairs of keywords/value; each pair indicating the representative keywords of the categories and the quantity of relevance to the category. The relevance derives from the tf-idf weight of the keyword into the set of documents of the specific category. If for example the user has selected to see articles labeled as "business" when registering to the web environment then equation 1 is utilized in order to construct a simple vector for the user with keywords and relevance.

$$\beta(x) = \sum_{\kappa=1}^{n} \beta\kappa_x(\kappa) * \epsilon(\kappa) \quad (1)$$

where β will be the relevance for keyword x, $\beta\kappa_(x)$ is the relevance of keyword x in category κ and $\epsilon(\kappa)$ is the user's selection against a category. ϵ is defined as:

$$\epsilon(\kappa) = \Delta * X_{\kappa}^{2} \quad (2)$$

where Δ can range from 1 to 5 according to how much we want to force or affect the profile towards the choice of the user. A price of 1 is mild while 5 is considered to be too high especially if the user selects negative values for many categories. In general, if the user selects high values for many categories we set Δ equal to 1 and when the user selects low values for many categories then we set Δ equal to 5. In general and in most of the cases we set Δ equal to 3.

This is done for the first $\epsilon(\kappa)$ keywords of the category (κ) as they are considered to be the most representative in order to create an initial profile for the user. The profile is an initial vector that is used to present the first list of articles that are similar to the user's profile and the values can be positive or negative. The positive values are used in order to obtain the articles that are relevant to the user and the negative are used in order to reject articles from the ones that are selected for the user. The algorithm that is used in order to measure the relevance of an article (document) to the user (terms - query) is a variant of the Okapi BM25 set of algorithms [3] which is out of the scope of the current work.

For the maintenance of the user profile a set of algorithms is utilized which derive from the usage of services of the meta-portal. The scope of each of the algorithms is to add new pairs into the user's vector or update the existing ones, and the information on how to do this derive from the user's activities in the meta-portal. Experimental evaluation that was done into news articles has shown that the first 10

keywords can represent fully the article's semantic quality. The information that is collected in order to update the profile of the user is the articles that a user has read, the articles that are rejected while all the other information are collected while the user is reading an article. When the user starts a session, a session recorder is responsible for recording the user's input. A weight is assigned to every keyword/relevance existing in any article that a user reads or rejects. When the user is reading an article an initial state of the weight is given according to equation 3.

$$weight(keywords) = \frac{min(timereading(x), time2read(length(x)))}{time2read(length(x))} \cdot (1 + \frac{articleposition(x)^2}{\sqrt{articleposition(1)^2 + articleposition(2)^+ articleposition(n)^2}}) \quad (3)$$

After opening an article to read it the weight can still be changed. The weight can be changed according to the actions of the user which are described in section Flow of Information. Each of the action is assigned a weight according to its importance which derives from questionnaires provided to internet users and their opinion on the importance of some factors. This implies that selecting to track an article is more important than reading similar articles on the issue while rejecting an article must affect the weight negative and more specifically affect with a high negative weight. Each action is recorded and when the article page is unloaded the sum of the weight percentage to be added to the pairs is added directly to the weight deriving from eq. 3. At the end of the session, each keyword that was located in one of the documents read or in one of the document rejected is followed by a weight either positive or negative. If the keyword exists then the system updates the value of the keyword into the user's profile by directly adding the weight to the already existing one without exceeding the triple value. If the keyword does not exist into the user's vector then a new entry is added with a value equal to the weight that was recorded but not larger than the double of the maximum existing value. The limits exist in order not to overload the user's profile quickly with keywords but form a user's profile gradually. In case of negative weights we apply the same procedure with the same limits.

B. On-line document grouping

The on-line document grouping is a procedure that takes place while a user is reading an article and its scope is to collect and interconnect every document that is identical to any other document but they derive from different sources. This is done in order to omit any duplicate instances of the articles and create linkages between articles. Moreover, this procedure speeds up the presentation of sets of articles to the end user. The algorithm that is utilized in order to locate all the identical articles is based on the cosine similarity measure and is done real-time. In order to present how the on-line document grouping is done we are making two basic assumptions: a. the system has never done document clustering before, b. a document can be related to any other document if they have 16 hours difference at most and c. the oldest article in a cluster and the newest article in the cluster should have 16 hours differrence at most. When the user is selecting an article to read, a function fetches asynchronously the group of documents that are directly related to the one that the user is reading. If the document cluster already exists then all the articles within the cluster are directly presented to the user. In parallel, even if the cluster exists, and because of the fact that articles are added every five minutes, if the newest article in the cluster is not older than 16 hours and if the difference between the oldest and newest article is the cluster is no more than 16 hours the system keeps checking for articles that may belong to the current cluster. If an article that the user is reading does not belong to a cluster then the cluster is created while the user is reading. The user is invloved in the procedure of cluster creation in order to assure that the cluster consists of identical articles only. Assuming that a user is presented an article this indicates that the user is interested into reading the article, which furthermore means that the vector of the user's profile is close to the vector of the article. What we expect is the identical articles to be close to the user's vector. The limit that we have with the addition of the user's interaction is another limit that has to be passed excempt from the similarity of the documents. If the article that the user is reading has similarity A with the current user then the rest of the articles that are meant to form the cluster should have $\pm \beta *$A where β varies from 0.07 to 0.1 according to our experimental evaluation and it is directly dependand on A. If A is relatively small (less than 30%) then it seems that the limit of β should be 0.1 while when A reaches values of 80% or more then β could be 0.07. It seems that the use of the median (0.085) is sufficient taking into account that most of the articles that are presented to a user have usually more than 50% relation to the user profile's vector.

IV. Experimental Evaluation

The experimental evaluation of the system consists of experiments conducted in order to present the creation and maintenance of the user profile and to provide information about the statistics of the document grouping procedure. We utilize peRSSonal meta-portal which we enhance with the dynamic user profiling mechanism and the document clustering sub-system, and we are executing our experiments on both real and virtual users that are registered officially in peRSSonal.

A. Experiments on the dynamic user profile

In order to conduct experiments on the dynamic user profile we are first experimenting with the reason of existance of a user profile within a meta-portal. Figures 3 and 4 present how many articles interest a user without

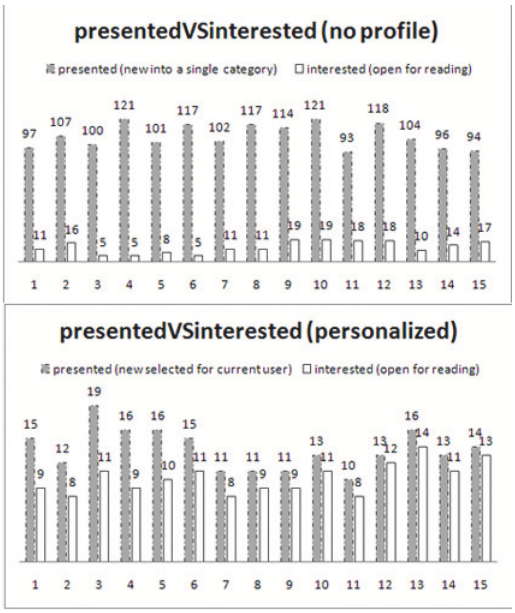

Figure 3. Presented VS Interested without profile — Presented VS Interested with use of dynamic profile

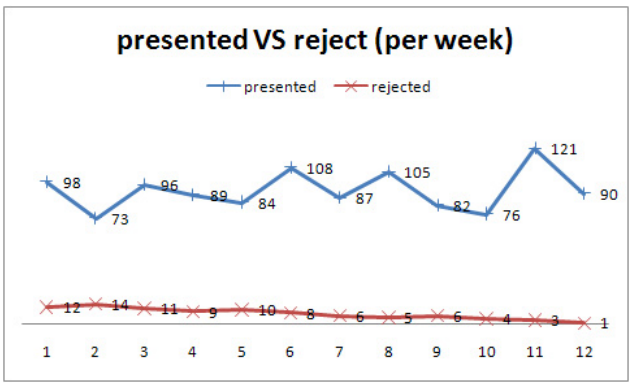

Figure 4. presented VS rejected (per week)

profile from the ones that are presented to her/him and what is the corresponding fraction when user profiling and personalization is used.

It is obvious that when a user enters and browses the meta-portal without creating a profile the articles that are presented to her/him concerning one category of the portal are usually more than 100. From them, the user normally selects 7-15 articles to read. It seems that only a 10% of the articles presented are of the user's interest.

On the other side, when a user creates a profile the articles presented to her/him concerning one of the categories that the user has selected are no more than 15. From the articles selected more than 75% are selected by the user to be read. The difference is huge and implies that the personalization is essential for a meta-portal that presents huge amounts of information. Another set of experiments is conducted with the help of an add-on to the meta-portal in order to obtain information about how much time is required for a user in order to create a sufficient profile and present only information that are of high interest to the user. The adaptability of the dynamic profile mechanism can be measured by asking the testers of the system to record how many of the weekly presented articles they reject.

As it is obvious after 4 weeks of the user browsing the meta-portal more than 90% of the articles presented to the user are of the user's interest.

B. Experiments on the document clustering

In order to test the document grouping mechanism we conducted a basic experiment to test its efficiency. We searched through the major portals that the system checks to find manually the same article published in all of them. After ensuring that our meta-portal has obtained all these articles we open one of them to see if the system is able to construct a linkage between all of them. For this reason we are checking the RSS feeds of politics news of seven major portals of Europe and the USA. When we locate an article that is published to all of them we check our meta-portal and open one instance of the articles and check if all the other 6 instances (one shown and 6 more articles from the seven portals) are present. Figure 5 presents the efficiency of the mechanism against 7336 articles (7x1048 articles). We expect the system to construct 1048 distinct document clusters.

We furthermore analyze the 54 clusters that fail to include the 7 instances of the articles in order to see how many instancs they were able to include. Figure 5 presents the results. It is clear that the vast majority of the inclomplete clusters include at least 5 of the 7 published articles (more than 60%).

V. Conclusion and Future Work

In this paper we presented a mechanism that is able to complete a procedure of collecting news from news portals and blogs and present them personalized back to the end-users by applying furthermore document clustering algorithms. This mechanism is helpful for internet users who are spending a considerable amount of time trying to locate news of their interest through major or minor news portals or even through RSS feeds (RSS readers). Despite the fact that the personalization micro-sites that exist, even within some portals, resolve part of the problem, still the refinement of the results and the personalization on the specific device of the user and the specific needs of the user is a huge problem. The procedure of accessing all the news portals in order to collect useful information is part of our everyday life, though, the information that is shown to the screen of the end user includes almost 80% of not needed information or even trash information. The mechanism that we are proposing is able to collect the articles from news portals

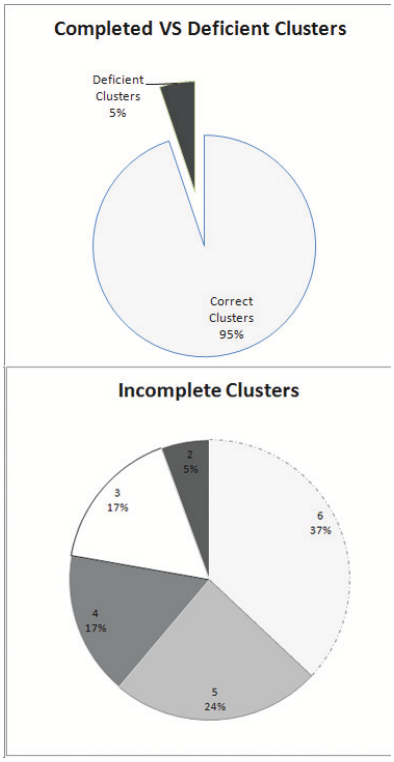

Figure 5. Completed VS Deficient Clusters — Analysis of Deficient Document Clusters

(through their RSS feeds), categorize the articles, summarize them and finally present them to the end-users according to their preferences and according to their device capabilities by grouping them in document clusters containing only identical articles. As a future work for our mechanism we are thinking of a news tracker system which will be able to track the changes that are done on news articles and update accordingly the document clusters. As more and more articles about a specific theme are published on several news portals or even on the same news portal we should be able to collect all the similar news and present them as one to the end user, providing also with the several links that the articles derive from and let the user make the best choice on which link to follow. Additionally, the automated procedure of maintenance of a user profile can be enhanced with user grouping procedure that will let users with similar interests exchange information on news articles. Finally, as the system is able to work at a very high speed, creating dynamically RSS for the user in real time, we are thinking of creating an add-on for every news portal that will enable the real-time creation of personalized RSS feeds for the end-user directly through the news portals.

REFERENCES

[1] S. S. Anand and B. Mombasher "Intelligent techniques for Web personalization". In Intelligent Techniques for Web Personalization Eds. Lecture Noes in Artificial Intelligence, vol. 3169. Springer-Verlag, Berlin, Germany, 137 (2005)

[2] J. L. Herlocker, J. A. Konstan, L. G. Terveen and J. T. Riedl. "Evaluating collaborative filtering recommender systems". ACM Trans. Information Systems 22, 1, 553, (2004)

[3] K. S. Jones, S. Walker and S. E. Robertson "A Probabilistic Model of Information Retrieval: Development and Comparative Experiments" (parts 1 and 2). Information Processing and Management, 36(6):779-840. (2000).

[4] R. Baraglia and F. Silvestri "Dynamic personalization of web sites without user intervention". Communications of the ACM, Volume 50, Issue 2, pp. 63 - 67, (2007)

[5] M. Mulvenna, S. S. Anand and A. G. Buchner "Personalization on the net using web mining". Communication of ACM 43, pp. 122 - 125, (2000)

[6] O. R. Zaiane, M. Xin, and J. Han. "Discovering web access patterns and trends by applying olap and data mining technology on web logs". In Proceedings of Advances in Digital Libraries Conference (ADL98), Santa Barbara, CA, (1998)

[7] Z. Huang. A cube model for web access sessions and cluster analysis. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2001).

[8] B. Mobasher. Data Mining for Web Personalization, volume 4321. The Adaptive Web : Methods and Strategies of Web Personalization, Berlin Heidelberg New York, Springer-Verlag edition, 2007. Lecture Notes in Computer Science.

[9] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan. Web usage mining: Discovery and application of usage patterns from web data. ACM SIGKDD Explorations Newsletter, 1:12-23, January 2000.

[10] C. Bouras, V. Poulopoulos, V. Tsogkas. PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries/ Data and Knowledge Engineering Journal, Elsevier Science, 2008, Vol. 64, Issue , 2008, pp. 330 - 345

[11] H.-R. Kim and P. Chan, "Learning implicit user interest hierarchy for context in personalization," Applied Intelligence, vol. 28, no. 2, pp. 153-166, April 2008.

[12] A. Sieg, B. Mobasher and R. Burke, "Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search", IEEE Intelligent Informatics Bulletin, Vol. 8, No. 1, November 2007.

[13] J. Garofalakis, Th. Giannakoudi and A. Vopi "Personalized Web Search by Constructing Semantic Clusters of User Profiles", in Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Zagreb, Croatia, pp. 238 - 247, 2008.

[14] Y. Wang and A. Kobsa "Respecting Users' Individual Privacy Constraints in Web Personalization". User Modeling 2007, Lecture Notes in Computer Science, Springer-Verlang, Vol. 4511 pp. 157 - 166, 2007.