Creating hierarchical user profiles using Wikipedia

Krishnan Ramanathan, Julien Giraudi, Ajay Gupta

HP Laboratories
HPL-2008-127

Keyword(s):
personalization, user profiles, Wikipedia

Abstract:

Personalized information retrieval and search promises to improve the Internet experience. An important requirement for building personalized web applications is to build user profiles that represent the users' interests. There are two representations commonly used for user profiles. One is using frequently occurring words in user documents. This creates large profiles where profile terms have low precision and have insufficient context to determine the user interests. The other is using a pre-existing ontology such as DMOZ. While this approach alleviates the ontology creation and maintenance problem, it requires constructing classifiers for each DMOZ node. Besides, of all the topics in the DMOZ ontology, most people will have only a small fraction of the topics as their interests and hence most of the ontology is redundant for capturing the interests of a specific user. This paper presents an alternative method to construct a hierarchical user profile using Wikipedia as the vocabulary for describing the user interests. The profiles created in this manner are more compact and have high precision compared to profiles that use words. We also discuss a method to tag concepts in these profiles as being of recreational or transactional interest.

External Posting Date: October 6, 2008 [Fulltext] Approved for External Publication
Internal Posting Date: October 6, 2008 [Fulltext]

Creating hierarchical user profiles using Wikipedia

Krishnan Ramanathan, Julien Giraudi[1] and Ajay Gupta
HP Labs, Bangalore, India
Email: krishnan_ramanathan@hp.com, julien.giraudi@gmail.com, ajay.gupta@hp.com

Abstract

Personalized information retrieval and search promises to improve the Internet experience. An important requirement for building personalized web applications is to build user profiles that represent the users' interests. There are two representations commonly used for user profiles. One is using frequently occurring words in user documents. This creates large profiles where profile terms have low precision and have insufficient context to determine the user interests. The other is using a pre-existing ontology such as DMOZ. While this approach alleviates the ontology creation and maintenance problem, it requires constructing classifiers for each DMOZ node. Besides, of all the topics in the DMOZ ontology, most people will have only a small fraction of the topics as their interests and hence most of the ontology is redundant for capturing the interests of a specific user.

This paper presents an alternative method to construct a hierarchical user profile using Wikipedia as the vocabulary for describing the user interests. The profiles created in this manner are more compact and have high precision compared to profiles that use words. We also discuss a method to tag concepts in these profiles as being of recreational or transactional interest.

1 Introduction

Huge amount of information gets added to the Web everyday. Publicly visible text creation is of the order of 10 GB per day and private text creation (including user email, IM messages, tags, reviews etc) is of the order of 3 terabytes per day [15]. This rapidly increasing scale of the web is in many ways limiting the utility of the web. There is a high level of noise beginning from spam and ending with a lot of uninteresting, irrelevant and duplicated content. Search engines and other forms of ranking are unable to keep up with this. Recently, search engines have started showing Wikipedia links as the top search result because ranking has become very hard.

Personalization [16] is a playing an increasingly important role in creating better Internet experiences. Recent applications of personalization have focused on improving the search experience [9]. An important aspect of personalization is creation of a user profile. The user profile [13] could be created on the client PC or on an Internet server. Both these methods have different advantages. Client side profiles offer better privacy, a more complete view of the user data. Server side profiles enable collaborative filtering and profile portability.

[1] Participated in this work on an internship from EPFL, Lausanne, Switzerland

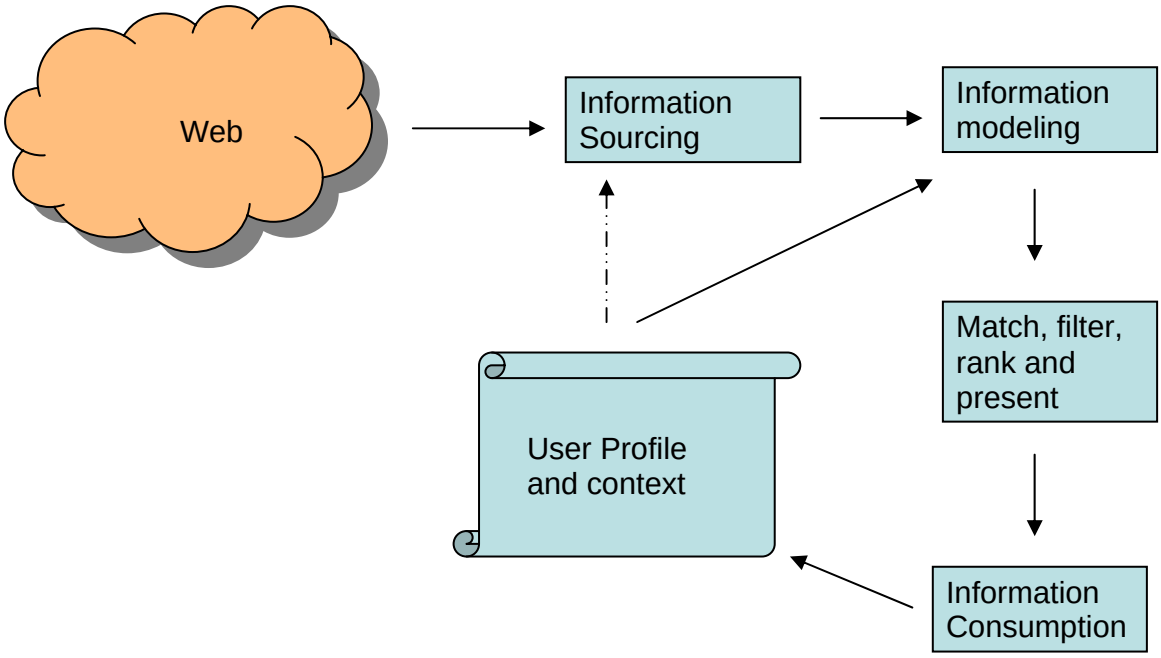

Figure 1 Using user profiles in information consumption

Figure 1 illustrates how a user profile might be used to personalize the information consumption process. The user profile is constructed by observing the information consumption patterns of the user such as the browser web page cache. The profile may also be bootstrapped using manual methods. It can be used to drive information sourcing (using API's provided by search engines or other aggregators such as de.li.cio.us) and to match and filter information obtained from information sources such as RSS feeds. The matched content is ranked and presented (using global and local information on the popularity and relevance of the content). Implicit and explicit feedback from the consumption behavior is used to update the profile which then drives information sourcing and filtering in future

Studies have shown that people are unwilling to explicitly specify their interests. A lot of personal data exists on user desktops; richer profiles can be built using this data. Hence, most prior research has focused on creating implicit user profiles [1, 4, and 9]. Most of the prior efforts in creating user profiles use frequently occurring document words to represent the profile. Profiles created in this manner suffer from the following problems

1. Irrelevant words – Words frequently occur in documents or web pages without being related to the contents of the page. For instance, a lot of web pages have the words "Home" and "page", however including these kinds of words in the profile is not useful.
2. Polysemy and synonymy – A word can have multiple meanings and multiple words can have the same meaning. A word based profile does not have sufficient context to disambiguate the meanings of individual words.

3. Size of the profile – The size of the profiles built using words grow very fast, larger profiles reduce precision.
4. Words in the profile may represent a mixture of information, transactional and recreational needs of the user. For instance, the term camera might appear in the profile because a user read a review for a camera (in a transactional context). Using this in an informational context to recommend a news item might not be appropriate. The profile items need to be disambiguated and only the part of the profile relevant to the users' current need should be used

Given the above limitations of word based profiles, it is worthwhile considering other approaches for feature generation and concept representation in user profiles. Some authors have used a pre-existing ontology such as DMOZ [9] for representing the profile. The advantage of using DMOZ is that is an open-source voluntary effort, hence the onus of maintaining the ontology does not reside with one particular organization. There are a couple of problems in using DMOZ. The first is that the DMOZ tree is very large (over 600000 nodes). Most users will have only a fraction of interests represented by DMOZ and hence most nodes in the hierarchy will never be used. The second problem is that given a user document or web page, mapping it to a DMOZ node requires building classifiers for each node in DMOZ.

Gabrilovich [11] presented a method for feature generation for text documents using Wikipedia. This method takes a text document and maps it to a set of Wikipedia concepts. In this paper, we introduce a method for hierarchical profile creation that uses Wikipedia concepts for representing user interests in the profile. Wikipedia contains most of the topics of interest to a vast majority of humanity. A Wikipedia concept is far richer than a group of words in conveying a user interest. By creating a user specific hierarchy, we make the profiles concise.

Our main contributions in this paper are

- We propose a method for creating hierarchical user profiles using Wikipedia concepts
- We evaluate this profile on the parameters of profile stability and precision at different levels of the profile hierarchy.
- We propose a method for distinguishing informational and recreational interests in the profile from the commercial interests.

2 Related work

There are two approaches to building a hierarchical profile. One is to use an existing ontology such as the Open Directory. The profile is built by mapping or classifying user documents to an existing ontology. In a way, the ontology defines (and restricts) the vocabulary of the profile. Once the profile is built, tree distance measures can be used to measure the relatedness of two nodes in the profile [9]. However, building a profile for a single user using an ontology can be an expensive proposition (the DMOZ ontology is about 590000 nodes and it is unlikely that a single user will have more than a fraction of

interests represented in DMOZ). To get around this, Chirita et.al [9] require the user to input nodes in DMOZ that are of interest to them. However, users do not like to give these kinds of inputs. Moreover, the DMOZ hierarchy also represents the collective belief of a number of people and may not have enough detail to capture specific interests of a user.

The alternative is to build a hierarchy from scratch. Godoy and Amandi [3] present an algorithm that uses both implicit and explicit indicators of user interest to construct a hierarchical profile. Nodes in the profile are term vectors and leaves are words representing user interests. The algorithm uses cohesiveness with respect to the cluster centroid to assign new words to clusters.

Kim and Chen [2] describe an approach to construct a hierarchical profile (which they call User interest hierarchy). More general interests of the user are represented by the most frequently occurring words in the user collection. Words at the top of the hierarchy represent general interests of the user and words at the leaves represent specific interests. Xu et.al. [5] present a similar approach, where high frequent terms for a document is first identified (e.g. Sports, soccer). Then similar terms are identified using a Jacquard measure and these terms become part of a branch in the hierarchy. Then parent-child relationships are constructed to identify the root node of the hierarchy.

There has also been work on labeling web pages as having commercial intent [6]. However, there has not been much work on labeling nodes in the profile as it is hard to do with word based representations. There has also been some work on evaluating and improving user profiles after they have been built [8].

3 The proposed method

Creating a user profile using Wikipedia involves three steps. First the web pages (and other documents) are mapped to a set of Wikipedia concepts. Then a hierarchical profile is constructed from these concepts. Finally, the concepts in the profile are tagged in two ways. One type of tag describes whether the concept is of transactional or recreational interest. The other type of tag is a measure of how recent is the user interest in that concept.

3.1. Mapping user visited web pages to a Wikipedia concept.

A page is mapped to a Wikipedia concept as follows [11]. First all the Wikipedia topics and the content of the topics is indexed (we used Lucene for indexing). To map a document to a concept, we query the index with the contents of the web page. The titles of the documents that are returned as the query results constitute the mapping of the document to the Wikipedia concepts. The process is illustrated in figure 2.

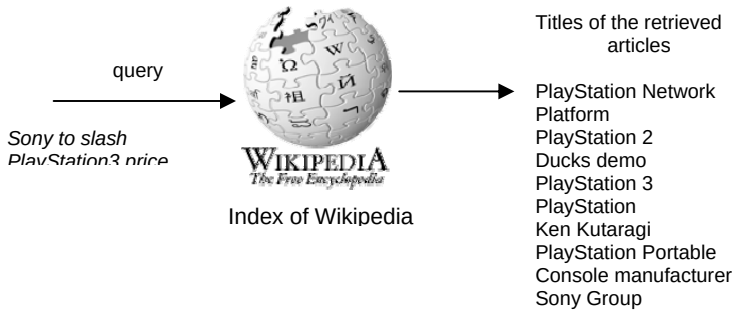

Figure 2 Querying the Wikipedia index

The brute force method of feeding the entire document content (after filtering out the document headers and html tags) suffered both from poor precision and poor performance. This is the well known long query problem. Gabrilovich [11] compute extract bigrams, however this would be very expensive to do on a user desktop. We devised a method to improve the precision and performance by choosing a subset of words from the document to map to a Wikipedia concept. We first computed the average word length and average support (number of occurrences) for all words in the document. We then chose only those words that had length greater or equal to the average word length in the document and frequency equal or greater than the average word frequency.

The selected words are fed as a query to the index of Wikipedia. The top 20 results returned by Lucene that exceed a cosine threshold similarity (0.1 in our experiments) are selected. Another Lucene index is constructed for each document; in this index the document title and the Wikipedia concept titles that matched the document are stored.

3.2. Constructing the hierarchical profile from the Wikipedia concepts

This is done using an algorithm similar to [5]. Each web page in the user cache is mapped to Wikipedia concepts. Each concept is then merged with a similar concept, made a child of another concept or remains as an independent concept. We set the minsup parameter in the algorithm of [5] to 4 and the delta parameter to 0.6. Figure 3 shows part of the hierarchical profile.

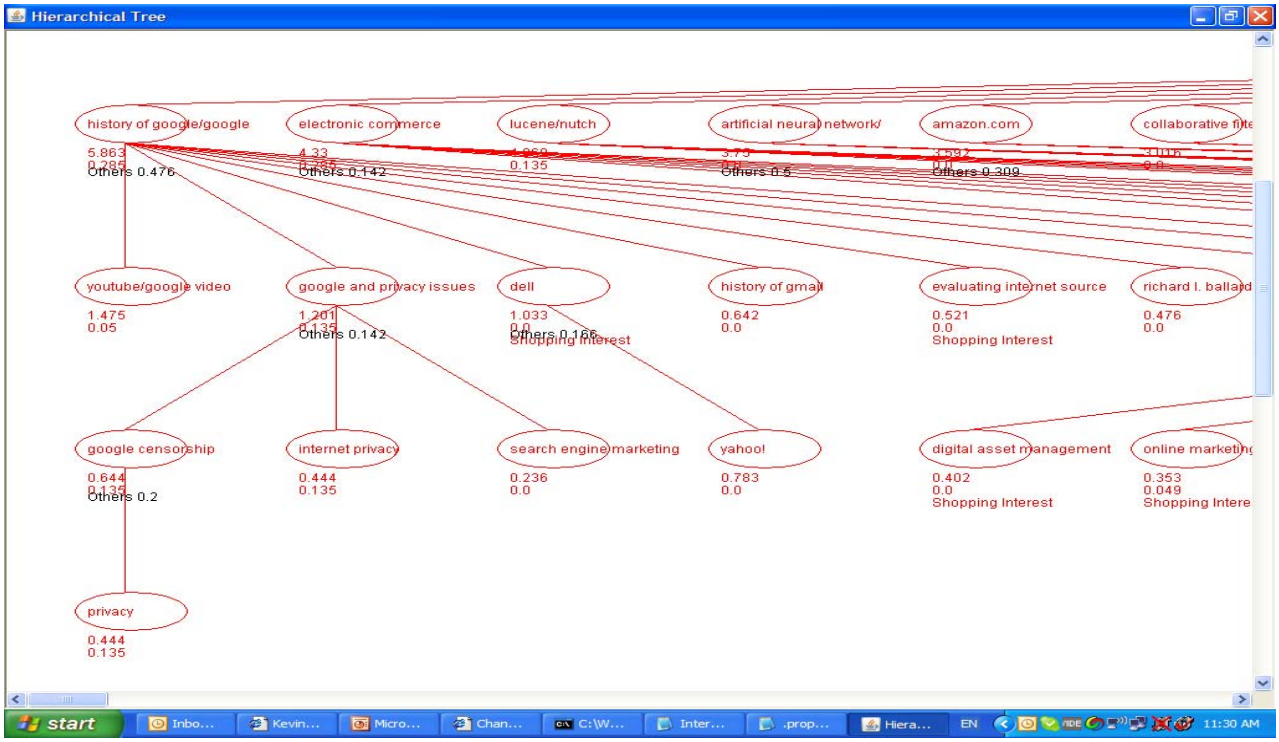

Figure 3 The hierarchical profile of Wikipedia concepts

3.3. Tagging Profile concepts

After the hierarchical profile is generated, the concepts in the profile are tagged in two ways

- as being of transactional interest or recreational interest.
- with the recency of the user interest in a concept

Some concepts may carry both tags. For instance, a user having photography as a hobby and who searched for cameras would have photography tagged as both a transactional and recreational interest. The concepts are also tagged according to the recency of interest.

For tagging transactional content, we crawled pages from shopping sites that allowed crawling. We then mapped the contents of each page to Wikipedia concepts and labeled those concepts as having transactional value. This gave a list of a shopping concepts, some of which we filtered manually as they did not pertain to shopping. After the filtering, we had about 7000 topics of transactional interest. For tagging recreational and hobby content, we just picked the topics under the recreational and hobby categories in Wikipedia. This yielded about 300 topics.

The recency of the user interest in a particular concept is based on the age of the pages supporting the concept.

$$\text{Re}cency = \sum_{\text{sup}\, porting\_pages} 1/e^{(todays\_date - datepagewasaccessed)}$$

The exponential decay ensures that recency of interest is significant only if a page mapped to the concept in the last week or so. This would allow a potential advertiser to target concepts of current interest to the consumer and to stop advertising after the interest wanes (this could happen if the user bought the item he was looking for).

4 Experiments

We have implemented the profiler in Java. The profiler uses web pages from the Internet Explorer (IE) Internet cache to construct the profile. To evaluate the profile, a collection of 3000 web pages covering a six month period from the web cache of the first author was used. All the pages were not used in profile construction, specifically pages with ad images, very small amount of text etc were excluded. This yielded about 600 pages that could be used

4.1 Stability of the profile

The first experiment we performed was to investigate the stability of the profile built using Wikipedia was compared to a word based profile. One definition of stability is from [8], namely the percentage of the top 50% words or concepts in the profile that change with additional data. We use a different measure of stability defined as the number of concepts in the profile with support greater than 5 that changed in an iteration of profile building over 100 web pages. We chose a threshold of 5 because our precision experiments (described later) showed that the resulting profile has high precision (exceeding 95 %) with this threshold. We constructed the profiles by considering pages in two ways: by their alphabetical names and by date.

Figure 4 below shows how stability evolved, when the web pages are considered in the order of the date of browsing (Stability_date) or in a random (alphabetical) order(Stability_alpha). As would be expected, the profiles become more stable as the number of pages in the profile increases. There is a slight dip in the stability when considered by date of browsing, this could be because of new user interests.

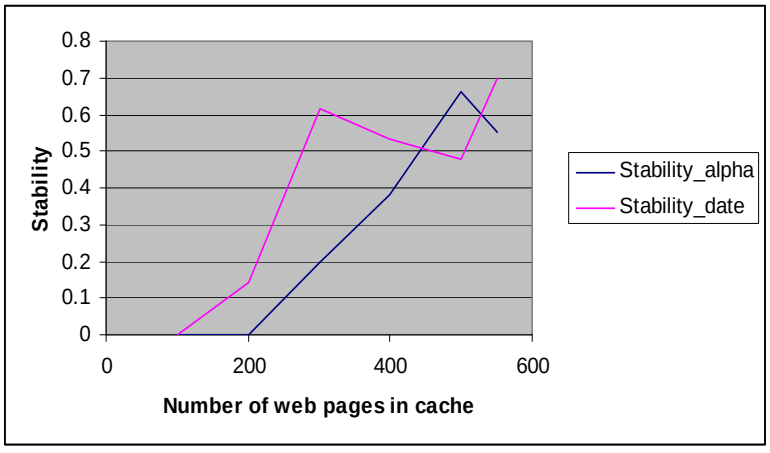

Figure 4 Stability of the profile

4.2 Profile Precision

We then evaluated the precision of the profile (figure 5). We separated out the profile concepts into three buckets: concepts with support greater than or equal to 5, concepts with support between 3-5 pages and concepts with support between 1-3 pages. Each term in the profile is rated as either relevant or not relevant by the user. As expected, the precision in the first bucket (for concepts with high support) was the highest.

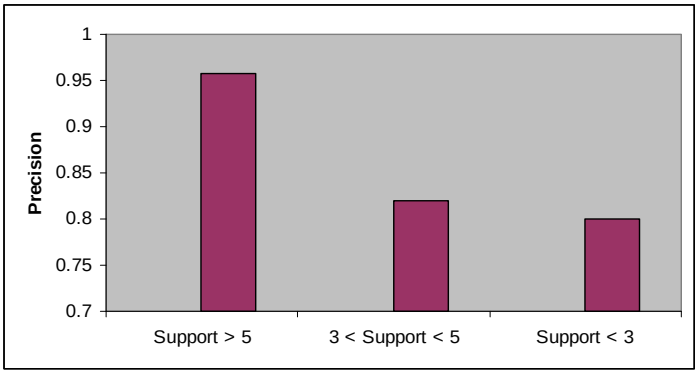

Figure 5 Profile precision

We next investigated the profile concepts at different levels in the hierarchy and their precision (figure 6). The maximum number of concepts were at level 2(34 %) followed by level 1 and level 3 (25 % each). The precision was maximum at level 3(about 97%); the precision at level 5 and 6 although 100% was not considered as there were few concepts at these levels in the hierarchy.

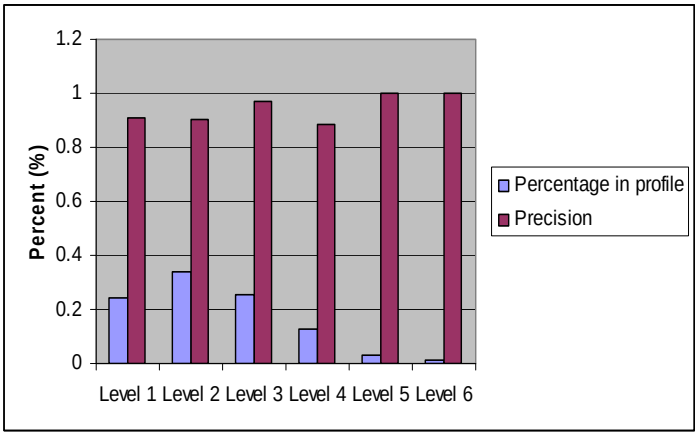

Figure 6 Profile precision at different levels of the hierarchy

6 Conclusions and future work

In this paper we have presented a method of constructing a user profile that uses Wikipedia. We found that good user profiles could be constructed using Wikipedia concepts (instead of words) as the profile representation. We also evaluated the resulting profile for stability and precision and found that the profiles are stable and concepts at the top of the hierarchy (that have high support) also have high precision. We also presented methods to tag concepts in the profile as being of transactional/recreational interest. Finally, we also presented a method to tag the recency of user interests in a concept. This could be used for both content filtering and advertising. One of the limitations of the current method is the large size of the Wikipedia index (about 1.4 GB). This restricts the use of the profiler on devices with low memory; we plan to work on reducing the size of the index.

One of the major concerns of user profiling is of privacy. Studies have shown that users are willing to trade profile privacy for other benefits [10]. There have been approaches to profile privacy by restricting access to the profile based on the support of a concept or by providing throttles based on the entropy exposed [5]. We also plan to investigate ways of automatically tagging concepts in the profile as private. In particular, we wish to investigate whether information sources such as user email (and other social network information like IM conversations) can be used to decide which concepts are private

Not all user preferences and interests are relevant in all situations. Hence it is necessary to activate different subsets of user preferences at runtime depending on the user context. Prior studies have represented context as changes to search queries/documents browsed [15]. Spreading activation techniques have been used to create contextual user profiles by amending the interest scores in the profile to represent the current context [14]. We plan to explore how to use context to construct profile views that are relevant to the current user context.

Finally, we plan to evaluate the utility of the profile in applications such as video sourcing, news filtering and search re-ranking [12]. We are also planning to conduct a study with advertisers to understand the value of the profile from an advertising perspective.

Acknowledgements:

I would like to thank Somnath Banerjee of HP Labs for his implementation of the Wikipedia querying software.

References

1. C.C. Chen, M.C. Chen, PVA: A self-adaptive personal view agent, Journal of Intelligent Information systems, 18:2/3, pp 173-194, 2002.

2. H.R. Kim and P.K. Chan, Learning implicit user interest hierarchy for context in personalization, Proc. of International conference on Intelligent User Interfaces (IUI), Miami, Florida, 2003.

3. D. Godoy and A. Amandi, User profiling for web page filtering, IEEE Internet computing, July-August 2005.

4. K. Sugiyama, K. Hatano, M. Yoshikawa, Adaptive web search based on user profile constructed without any effort from users, WWW 2004.

5. Y. Xu, B. Zhang, Z. Chen and K. Wang, Privacy enhancing personalized web search, WWW 2007.

6. H. Dai et. al, Detecting online commercial intention, WWW 2006.

7. N. Nanas, V. Uren and A.D. Roeck, Building and applying a concept hierarchy representation of a user profile, SIGIR 2003.

8. Trajkova J. and Susan Gauch, Improving Ontology based user profiles, Proceedings of RIAO 2004, University of Avignon, France.

9. P.A. Chirita , W. Nejdl, R. Paiu, C. Kohlschutter, Using ODP data to personalize search, SIGIR 2005.

10. A. Kobsa, Privacy enhanced personalization, CACM vol. 50, no. 8, August 2007

11. E. Gabrilovich and S. Markovich, Overcoming the brittleness bottleneck with Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, Proc. of the AAAI conference, 2006.

12. Z. Ma, G. Pant and O.R.L. Sheng, Interest based personalized search, ACM transactions on Information systems, Vol. 25, no. 1, February 2007.

13. Susan Gauch et.al, User profiles for personalized information access, Ch.2, “The adaptive web”, Springer LNCS 4321.

14. Ahu Sieg, Bamshad Mobasher, Robin Burke, Web search personalization with ontological user profiles, CIKM 2007.

15. Raghu Ramakrishnan and Andrew Tomkins, Towards a People Web, IEEE Computer, August 2007

16. Krishnan Ramanathan, Geetha Manjunath and Somnath Banerjee, Personalization tutorial at ACM Computer 2008, http://www.slideshare.net/rkrish67/personalization-tutorial-at-acm-compute-2008/

17. David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas, and Yannis Avrithis, Personalized Content Retrieval in Context Using Ontological Knowledge, IEEE Transactions on circuits and systems for video technology, vol. 17, no. 3, March 2007