THE EFFECT OF SEMANTIC CLUSTERING ON WEB SEARCH PERSONALIZATION

Garofalakis John and Oikonomou Flora
Computer Engineering and Informatics Department, Patras University, University Campus, Patras, Greece

Keywords: Semantic Clustering, Ontology, User Profiles, Dietary Profiles, Web Search, Personalization, Web Usage Mining, Information Search and Retrieval.

Abstract: The World Wide Web has become a huge data repository and it keeps growing exponentially, whereas the human capability to find, process and understand the provided content remains constant. Search engines facilitate the search process in the WWW and they have become an integral part of the web users' daily lives. However users are characterized by different needs, preferences and special characteristics, navigate through large Web structures and may lost their goal of inquiry. Web personalization is one of the most promising approaches for alleviating information overload providing tailored navigation experiences to Web users. This paper presents the methodology which was implemented in order to personalize a search engine's results for corresponding users' preferences and dietary characteristics. This methodology was implemented in two parts. The online part uses a search engines' log files and the dietary characteristics of the users in order to extract information for their preferences. With the use of an ontology and a clustering algorithm, semantic profiling of users' interests is achieved. In the online part the methodology re-ranks the search engines' results. Experimental evaluation of the presented methodology shows that the expected objectives from the semantic users' clustering in search engines are achievable.

1 INTRODUCTION

The World Wide Web is a system of interlinked computer documents (webpages) that can be accessed via the internet, which is a network of computers that is distributed all over the world and that communicate with one another. People use the World Wide Web for two main reasons: in order to communicate with other people or extract desired information. Based on the latter, we can assume that the World Wide Web is a huge data repository, a global library which is expanding at exponential rates.

However, the human capability to find, process and understand the provided content remains constant. The phenomenal growth of the number of new users who access the web and the huge amount of data that the web has, make the searching for particular information not only time consuming but also a daunting task. There comes the need for search engines which facilitate the search process and have become an integral part of the web users' daily lives.

Traditional search engines obtain the results by matching the words in the query with the document content and finally give out thousands of result pages of which only a handful are relevant. Moreover, users are characterized by different needs, preferences and special characteristics, and by navigating through large Web structures may lost their goal of inquiry.

Personalization is a popular remedy for the above problem. Personalization improves the search by considering the user's interests and characteristics and hence provides context oriented search results, which are the direct answers to the user's information need. Semantic searches on the other hand attempt to augment and improve the traditional search results by using semantic information from resources like ontology and thesaurus.

In this work, we propose a personalization method, which couples data mining techniques with the underlying semantics of the web content in order to build semantic clusters of user profiles. Regarding the semantic clusters, they actually comprise ontological subsets of a general ontology which was

John G. and Flora O..
THE EFFECT OF SEMANTIC CLUSTERING ON WEB SEARCH PERSONALIZATION .
DOI: 10.5220/0003094500600069
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2010), pages 60-69
ISBN: 978-989-8425-29-4

specifically designed for this project, representing the categories of interest and the characteristics for groups of web users with similar search tasks and dietary profiles. In this methodology apart from exploiting a specific user search history and the search history of other users with similar interests, we further exploit their individual traits. The user is assigned to relevant classes of common interest and dietary characteristics, so as to predict the relevance score of the results with the user goal and finally re-rank them. To this purpose, we exploit a reference ontology which was based on the Finnish Food Composition Database for the National Institute for Health and Welfare and the information for the ontology's construction is freely available online at http://www.fineli.fi/index.php?lang=en.

Specifically the methodology consists of two major parts: the offline and the online part. In the former three tasks can be described. First of all, we gather the user's search history and his/hers dietary profile and then we process his/hers activity, taking into consideration other users' activities and dietary profiles and constructing clusters of commonly preferred concepts. The final task of the offline part is to define ontology-based profiles for the active user based on the detected interests from his current activity, the interests from the semantic cluster in which he has been assigned from previous searching sections and the dietary characteristics of the current user and the users who belong to the aforementioned cluster. As far as the online part is concerned, two tasks can be described here. At first, we re-rank the web results combing the above information with the semantics of the delivered search engine results and then we constantly re-organize the conceptual clusters in order to be up-to-date with the user's interests and personal characteristics.

Our approach has been experimentally evaluated by utilizing the Google AJAX Search API and the results show that semantically clustering users in terms of using ontologies in search engines is effective.

The remainder of the paper is structured as follows: Section 2 discusses related work. In Section 3 we present the motivation for creating this project. In Section 4, we describe the reference ontology that our approach uses and we outline the semantic annotation of web results to the ontology classes. Furthermore, a presentation of how the user profiles are defined over the reference ontology and how the semantic clusters are formed is documented. In Section 5, we propose a technique for web search personalization combining profiles of semantic clusters and users' special dietary characteristics with the emerging profile of the active user. In Section 6, we mention the technologies which were used for this project. Finally in Section 7, we exhibit our experimental results and in Section 8 conclusions and further work is presented.

2 RELATED WORK

Our work touches various fields. In the following we review related work in personalization, cluster formatting and personalized semantic search.

2.1 Personalization

As the amount of information on the web continuously grows, it has become increasingly difficult for web search engines to find information which satisfies users' individual needs. Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Many recent research efforts have focused on this area. Most of them could be categorized into two general approaches: Re-ranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine (Pitkow, et al, 2002).

Moreover significant studies have been conducted for personalization based on user search history. A general framework for personalization based on aggregate usage profiles is presented in (Mobasher, et al, 2000). This work distinguishes between the offline tasks of data preparation and usage mining and the online personalization components. (Qui, et al, 2006) suggest learning user's preferences automatically based on their past click history and show how to use this learning for result personalization. (Yabo, et al, 2007) presented a scalable way for users to automatically build rich user profiles which summarize a user's interests into a hierarchical organization according to specific interests. Another approach to personalization is presented in (Radlin ski, et al, 2006) where three methods are proposed to increase the diversity of the top N search results which are returned by a search engine. Their approach involves re-ranking the top N search results such that documents likely to be preferred by the user are presented higher. Furthermore they presented a number of methods to collect diverse results for a given query using past query reformulations. Additionally research on personalizing search results (Dou, et al, 2007; Shen, et al, 2005; Teevan, et al, 2005) has found that implicitly gathered information such as browser history, query history, and desktop information, can be used to improve the ranking of search results on a per-user basis.

Ideas from all of these works where used in our approach in order to create a personalized system.

2.2 Cluster Formatting

In order to enhance the performance of personalized search a lot techniques suggest the use of clusters of users with similar interests and characteristics. (Teevan, et al, 2005) found that the performance of the personalization algorithm they studied improved as more data became available about the target user. This finding confirms the suggestion that additional data from similar people may be useful in enhancing personalization systems.

(Dou, et al, 2007) compared several personalization strategies and found that the use of click-through data and k-nearest neighbor collaborative filtering was a promising approach. (Almeida & Almeida, 2004) used Bayesian algorithms to cluster users of an online bookstore's search service into communities based on links clicked within the site and found that the popularity of different links within different communities could be used to customize the search result rankings.

VisSearch system developed by (Lee, 2005) uses data mining to uncover patterns in users' queries and browsing in order to generate recommendations for users with similar queries. Some recommender systems, such as the movie recommender system PolyLens created by (O' Conner, et al, 2001) attempted to generate recommendation lists for groups of users. (Smyth, 2007) suggested that click-through data from users in the same "search community" (e.g., a group of people who use a special-interest Web portal or who work at the same company) could enhance search result lists. Smyth provided evidence for the existence of search communities by showing that a group of employees from a single company had a higher query similarity threshold than general Web users. (Freyne & Smyth's, 2006) I-SPY system expanded the notion of search communities to include related communities, measuring intercommunity similarity based on the degree to which communities' queries and result click through overlap.

(Mei and Church, 2008) found that geographic location might serve as a reasonable proxy for community, since they observed that grouping users into classes based on the similarity of their IP addresses could improve search results.

(Teevan, et al, 2009) found that groupization, a personalization technique that combines personal and group content improves Web rankings for different groups/queries. (Smith, et al, 2009) managed to develop unambiguous clusters of URLs from clickthrough data from the MSN search query log excerpt (the RFP 2006 dataset) whereas (Tesic, et al, 2007) proposed a cluster-based sampling method and data modelling of the semantic context in fusion with text and visual search baselines in order to boost search performance for a diverse range of query topics. Moreover, (Nguyen, et al, 2009) proposed a different way of clustering in order to facilitate personalization; they enriched short snippets with hidden topics from huge resources of documents on the Internet, and managed to cluster and label such snippets effectively in a topic-oriented manner without concerning the whole Web pages.

In our work we used ideas from the aforementioned clustering techniques in order to create effective clusters for personalizing the web search results.

2.3 Personalized Semantic Search

The enormous increase of information in the web led the information retrieval community to strive towards changing the concept of "good for all" to "good for everyone". This in turn popularized personalized semantic search engines and semantically enhanced recommendation systems, with some related work in (Pretschner, et al, 1999; Gauch, 2003; Sheth, et al, 2002; Liu, et al, 2002; Liu, et al, 2004; Liu, 2002; Kim & Chan, 2005; Bose, et al, 2006; Sieg, et al, 2007). (Zhuhadar & Nasraoui, 2008) presented an approach for personalized search in a e-learning platform, that takes advantage of semantic web standards (RDF and OWL) to represent the content and the user profiles. In this approach, however, the authors used a taxonomy to semantically characterize their context whereas in this paper we used an ontology which contains a greater amount of information. (Sendhilkuman & Geetha, 2008) designed a personalized search index with the use of an ontology which provided them with a conceptual relation between the search keywords and the pages which matches the user's information need.All of these approaches in combination with the (Garofalakis, et al, 2008, 2009) led to the development of our presented system.

3 MOTIVATION

Why did we need a Semantic Cluster-based Search Engine? As already mentioned the World Wide Web grows exponentially where as the human capability to find, process and understand the provided content remains constant. Moreover, users are characterized by different needs and characteristics may face problems during their navigation in the Web and may lose their goal of inquiry. Web personalization is a promising solution which alleviates the problem of information overload. Furthermore, semantic searches augment and improve the traditional search results by using semantic information from resources like ontology and thesaurus. Additionally, health-related searches constitute a 4.5 % of all the searches on the Web which means that about 6.75 million health-related searches per day in Google alone are being conducted (Eysenbach & Kohler, 2003).

Taking all of these under consideration, we thought of creating a semantic cluster-based search engine which not only meets its corresponding users' preferences but it also considers their dietary characteristics.

4 SEMANTIC PROFILING

The general aim of this work is to introduce a method for personalizing the results of web searching. For this reason we focused on constructing user profiles implicitly and automatically, according to their interests, dietary characteristics and their previous behaviour on searching. At this direction we were based on the work described in (Pretschner, et al, 2003; Garofalakis, et al, 2009).

4.1 Reference Ontology

Our first goal was to create a reference ontology upon which we will base the user profiles. The profile of each user will be represented by a weighted ontology, depicting the users' interest for every class of the reference ontology and the users' dietary characteristics. We decided to create a new ontology from scratch, which was based on the Finnish Food Composition Database for the National Institute for Health and Welfare. In Figure 1 there is a depiction of some of the concepts of the constructed ontology.

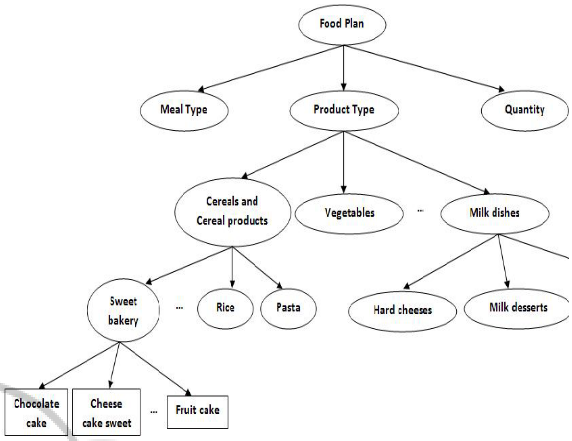

Figure 1: Part of the constructed ontology.

The ontology created is actually a directed acyclic graph (DAG). Since we wish to create a relatively concise user profile that identifies the general areas of a user's interests and dietary characteristics we created our reference ontology by enhancing with lots of food information. Thus, we not only created classes and instances as it has been done in previous works (which were presented in Section 2), but properties and methods were also implemented making the ontology more rich in context. The ontology in addition to the instances and the methods, constitute a knowledge base very thorough in information which can be used to effectively answer competency queries from the developed search engine. The ontology was created by Protégé, the free, open source ontology editor and knowledge-base framework and the language used for development was OWL.

4.2 Semantic Annotation

The construction of the profile, i.e. the weighted ontology, for every user includes two parts; the semantic annotation of user's previous choices and the dietary profile characterization which will be presented in Section 4.3. The semantic characterization of user choices is based on the methodology proposed in (Garofalakis, et al, 2009). Therefore, the user's previous choices are analyzed into keywords extracted from the visited web pages and the keywords are semantically characterized.

The calculation of the semantic similarity between each keyword and each term of the ontology was computed by using semantic similarity measures with WordNet. In Wordnet, English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Several experiments were made in order to select the appropriate measure for our ontology. The measure that was applied is the PATH measure which calculates the relatedness by counting nodes in the noun and verb WordNet 'is-a' hierarchies.

$$\text{relatedness} = \frac{1}{\text{distance}} \qquad (1)$$

If two concepts are identical, then the distance between them is one; therefore, their relatedness is also 1. However, the assignment process is time-consuming; therefore we have implemented a caching policy to improve system response. The assignments of instance words are kept in cache, to minimize response time in case these words are met again. Every time that this process is executed the amount of previous choices that are semantically annotated are the users' choices that have not been annotated at the last execution of this step of the methodology. This saves time from the execution, since semantic annotation is a quite time consuming step of the overall method applied. As a result, the keywords and consequently, the users' choices are assigned to relevant classes of the ontology.

4.3 User Profiles' Formulation

In this step, our methodology uses the semantic annotations of the users' choices so as to construct the profile for every user. After the semantic characterization of user's choices to the ontology concepts our methodology moves on the profile creation.

From the web access logs kept in the web server our method extracts the user's previous choices, which have already been semantically annotated. Therefore, for every user we extract the concepts and the frequency of appearance from the previous choices that the specific user has made. In the end of the execution of this step, there is an accumulation of the preferences for every user and of the frequency for every concept, which is the initial weight, for every class (preference) in the ontology.

Additionally, for every user we construct his/hers dietary profile based on the data which the user enters to specific forms of the system. Based on these data, we calculate the Body Mass Index (BMI) of each user. The initial weight which was mentioned earlier in combination with each user's BMI constitute the overall weight for every class in the ontology.

Apart from the accumulation of the concepts for which the user has shown interest and his/hers dietary profile, we construct the vector that represents each user's profile. The vector's size is the number of concepts that the ontology consists of. The value of each element of the vector corresponds to the weight of the user interest for this concept. So we propose that, the weight for a concept i for the user u, is calculated as:

$$w_{iu} = \frac{cf_{iu}}{sum(cf_u)} \qquad (2)$$

where

cf_{iu}= the number of times that the concept i has been assigned to the user u.
$sum(cf_u)$ = the sum of the times that all the concepts of the ontology has been assigned to the user u.

For the concepts that the user has not selected any previous choice assigned to this concept the value is set to zero. So for a user u the profile is represented as follows:

$$p=<w_1^p, w_2^p, \ldots, w_n^p > \qquad (3)$$

Where n is the number of concepts in the ontology and

$$w_i^p = \begin{Bmatrix} weight\ (concepti\ , p)\_\ if\ concept\ i > 0 \\ 0, otherwise \end{Bmatrix} \qquad (4)$$

Therefore, it is obvious that the weight of each concept is the relative frequency of the concept among all concepts of the ontology. The sum of all weight is equal to one, representing the percentage of the user's interest for every concept. Moreover, for each user we create a file that has the profile vector.

4.4 Semantic Clustering

After creating each user profile, we move on to profile clustering. From the profile creation step, a profile for every user is stored in the database and a file with the user's vector weighted ontology is created. At this step of the methodology, the profiles of all the users that reacted with the search engine and have similar BMI values are accumulated and are clustered into clusters with similar interests.

The clustering algorithm that has been applied in the methodology proposed in the profile clustering step is the K-Means algorithm (Ma, et al, 2007). The K-Means algorithm accepts the number of clusters to group data into and the dataset to cluster as input values. It then creates the first K initial clusters (K= number of clusters needed) from the dataset by choosing K rows of data randomly from the dataset. Next, K-Means assigns each record in the dataset to only one of the initial clusters. Each record is assigned to the nearest cluster (the cluster which it is most similar to) using a measure of distance or similarity like the Euclidean Distance Measure, which was used in this module. K-Means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset. The preceding steps are repeated until stable clusters are formed and the K-Means clustering procedure is completed. Stable clusters are formed when new iterations or repetitions of the K-Means clustering algorithm does not create new clusters as the cluster centre or Arithmetic Mean of each cluster formed is the same as the old cluster centre. In the end of the execution of this step the users are grouped into clusters with similar interests and the clusters are stored to the database. Thus, a cluster profile is built, utilizing the average of preferences of all cluster members:

$$p_c=<w_1, w_2, \ldots, w_n> \quad (5)$$

Every time this step is executed, the clusters are constructed from the beginning and the users are grouped again. Thus, the clustering procedure is not based on the previous constructed clusters because we considered that the user's choices will alter periodically. The construction of the semantic users' profiles clusters is presented in Figure 2.

5 PERSONALIZATION ALGORITHM

The pre-processed user's choices, their semantic characterization and the users' clusters are used for processing and personalizing the results from a search engine. At this point every user that has reacted previously with the online search engine has been put in one cluster.

This cluster consists of users with similar interests and can be depicted as a weighted ontology such as the profiles. This weighted ontology will be presented as a vector, too. The personalized search includes the calculation of the similarity of each result returned by the search engine with the cluster's interests and its dietary characteristics. This calculation requires the execution of all the steps of the ontology-based user clusters for each result returned by the search engine.

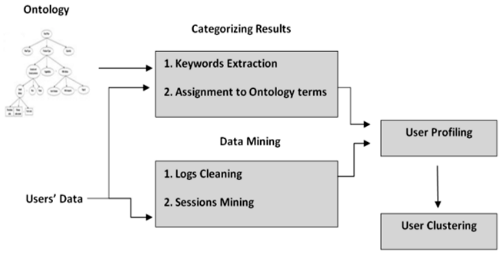

Figure 2: Clusters Creation by using semantic user's profiles.

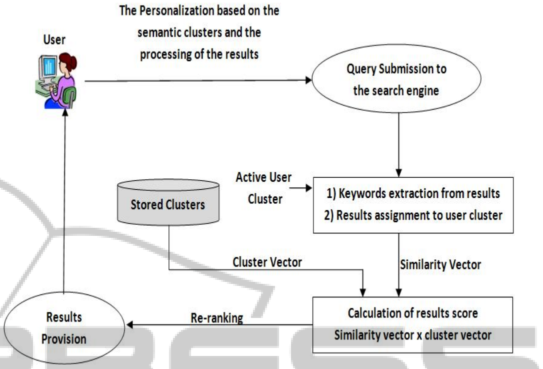
The Personalization based on the semantic clusters and the processing of the results
User
Query Submission to the search engine
Active User Cluster
Stored Clusters
1) Keywords extraction from results
2) Results assignment to user cluster
Cluster Vector
Similarity Vector
Results Provision
Re-ranking
Calculation of results score
Similarity vector x cluster vector

Figure 3: Personalization Algorithm.

Therefore, for every query that is set to the search engine the proposed methodology follows the following steps:

1. Extract keywords from users' previous choices.
2. Apply semantic annotation step not to the reference ontology but to a part of the ontology for which the cluster that the user belongs has a non-zero weight.
3. Calculate the value score value for each result.

The above three steps are executed for every result and the score value is kept in cache. Afterwards, the results of the search engine are organized for presentation to the user according to the score that has been calculated, beginning with the one with the highest score (Figure 3). The interaction with the search engine is made by using the Google AJAX search API which is referred to Section 7.

6 TECHNOLOGIES USED

The technologies which were used for the implementation of this work include the .NET framework and more specifically the C# programming language. For the database framework the SQL server was chosen. Those technologies were selected due to their high reliability, interconnectivity and portability. Moreover the .NET framework is a platform in which a lot of strange to it technologies can be applied effectively and enhance its operations.

In addition to the .NET framework, Perl programming language was used to create modules which performed the semantic similarity of the keywords with the ontology's terms. Perl was selected due to its capability of managing effectively strings of characters.

The applied ontology was created in the Protégé platform which supports the OWL (Web Ontology Language). Protégé was selected since it is a user friendly open source platform and it is systematically used by the scientific community.

Finally, AJAX search API was used in order to connect Google's search engine with our presented system. AJAX search API is actually a web service which is appropriate for keyword position tracking and can easily incorporated in a web page with the use of Javascript language.

7 TESTING AND EVALUATION

We developed a WWW search engine utilizing the Google AJAX Search API so as to test our methodology. The Search API returns the URL, the title and a short summary for each one of the eight results of the Google search engine. At first we run this limited search engine without personalizing the results but accumulating the users' choices. Next, we applied the method proposed and compared the results of the personalized representation with the non personalized representation.

For our experimental implementation, we used a database for storing users' BMI and their choices for every query applied in the limited search engine used for testing. Every time that a user enters the search engine there is an identification of the IP address, the agent and the domain name keeping off the multiple storage of a user in the database Moreover, the search engine stores in the database the query, user's calculated BMI profile which is based on the data the user inserted in a special system's form, and the choices of the user in every query. For every result that is clicked by the user the search engine stores the title, the URL and the short summary returned in the database. This database consists of the history of the requests and therefore is used as the web access logs in this methodology.

At next we apply the steps of the methodology proposed earlier for the creation of the semantic users' profiles clusters. In the web choice, i.e. the search engine's result, that has been selected we extract the keywords. For the experimental implementation the methodology for the keyword extraction is similar to the one proposed in (Dai, et al, 2002) for the keywords of the pages that have a link for a specific page. The keywords that are extracted for every URL are accumulated from the title of the URL and the short summary returned by the Google AJAX search API.

The title and the summary are parsed and are cleaned by the HTML tags and the stop words (very common words, numbers, symbols, articles) are removed, since they are considered not to contribute to the semantic denotation of the web page's content. The words that remain are considered the keywords for every URL since their number is small and no frequency is being taken into consideration. After the running of this step the keywords for every URL are stored to the database. At next the keywords are semantically characterized according to the way described in paragraph 4.2. Afterwards, the profiles of the users are created as analyzed in paragraph 4.3 and finally the users are grouped into clusters as referred in 4.4 according to the methodology proposed. In order to evaluate the proposed method and prove the efficient behavior of our personalization method, we performed some queries with polysemy and we used different types of users with different BMI values expecting the personalized results to be personalized according to the profile of the cluster that a user is set and to verify that our method can improve the results' ranking quality as desired. We applied the queries in the experimental implementation that returns the first eight results from the Google search engine through the Google AJAX search API. In one case we applied our personalization methodology whereas in the other case we extracted the results as they were returned by the search API. We evaluated the use of our automatically created user profiles for personalized search using the approach of ranking. A function is applied to the document query match values and the rank orders returned by the search engine. The relevant documents are moved higher in the results set and demote non-relevant documents.

Our experimental implementation was online for 1 month and twelve users have reacted with it. The choices that they have made for every query were stored in the database and their dietary profile was formed. The choices were processed and the user profiles were created. Next, we clustered the users in three clusters. The user that made the queries has already been put in a cluster and the reference ontology of the cluster upon which the score of the results will be based has been created.

The first example query which was applied in the search engine was the keyword "cookie". The word "cookie" not only has a twofold meaning, since it means the small, flat-baked treat or the text string stored by a user's web browser (also known as web cookie/ browser cookie/ HTTP cookie) but also constitutes a type of food which belongs to the cereal class according to the developed ontology. In this example a user with a dietary profile of class I obesity, interacts with the search engine. This type of user is advised to consume small quantities of such food categories (i.e. cookies) which provide his/hers body with enough energy without a lot of fat.

Based on all these, we expect that the search engine's results will refer both to food and computers. The user who gives this query in the search engine asks for information about cookie as a kind of food and expects results related to food. The search engine's role is twofold: on the one hand the returned results must be personalized according to the user's profile who is class I obese and on the other hand the results must be personalized based on the user's cluster which is a group of people who are interested in food and not in computers. In the following table we can see the results of the search engine for the query "cookie". The first column represents the order of the results of the AJAX search API without the application of the personalization methodology while in the second column we can see the order of the personalized results of the experimental application.

Next to each title we give in parenthesis the general concept of the result, which we have concluded after reading the summary. In the first column the result that the user searches is in place 5 and another result close to food but not the one that the user is interested in is in place 2. In the second column with the personalized results, the requested webpage is in place 2 and the similar to food webpage is in place 5. It is obvious that after the application of the personalization methodology that is proposed the results related with the user's interests are pushed to places closer to the top. The cluster into which the user belongs, as we have mentioned, has many interests that include food and this has been taken into consideration while calculating the score of each result pushing the results related with food in a higher place in the list of the results.

Table 1: Experimental results for query "cookie" for a user interested in a cookie related with food, with the cluster he belongs to has interest in food and the user has class I obesity.

Non Personalized Results	Personalized Results
HTTP cookie - Wikipedia, the free encyclopedia (computers)	The Unofficial Cookie FAQ (computers)
Cookie - Wikipedia, the free encyclopedia (food)	**Cookies - All Recipes (food)**
Cookie Mag: the Stylish Parenting Magazine for the New Mom: Home (people)	Enable Cookies: Search History and Settings –Web Search Help (computers)
What is cookie? - A Word Definition From the Webopedia Computer (computers)	Cookie Central (computers)
Cookies - All Recipes (food)	**Cookie - Wikipedia, the free encyclopedia (food)**
The Unofficial Cookie FAQ (computers)	HTTP cookie - Wikipedia, the free encyclopedia (computers)
Cookie Central (computers)	What is cookie? - A Word Definition From the Webopedia Computer (computers)
Enable Cookies: Search History and Settings – Web Search Help (computers)	Cookie Mag: the Stylish Parenting Magazine for the New Mom: Home (people)

Apart from this query, we have tested the proposed methodology in a second query, the keyword "egg". This keyword is not characterized by polysemy as was the case in the first example but is related with user's dietary profile. The user who performs this query is classified as underweight and the cluster to which he belongs has similar characteristics. Thus, the returned results which are related with food should be placed in higher places to the personalized search engine.

People who are categorized as underweight should consume food rich calories. Eggs are characterized as such and contain lots of nutrients. Consequently, for this user results related with food should be ranked higher.

In the second table we can see the results of the search engine for the query "egg". Again, in the first column the results that the user searches are in places 3 and 6 whereas in the second column, with the personalized results, the requested webpages are now in places 1 and 2. It is obvious that after the application of the personalization methodology the results related with the user's interests are pushed to places closer to the top.

In both examples, it has been shown that the methodology given the relatedness of the results with the cluster's preferences and user's characteristics has pushed the desired results in

Table 2: Experimental results for query “egg” for a user who is underweight and the cluster he belongs to has similar characteristics and is interested in food.

Non Personalized Results	Personalized Results
Egg (food) - Wikipedia, the free encyclopedia	**Eggs, How To Cook Eggs, Egg Recipes, Egg Nutrition Facts, All (food)**
Egg (biology) - Wikipedia, the free encyclopedia	**WHFoods: Eggs (food)**
Eggs, How To Cook Eggs, Egg Recipes, Egg Nutrition Facts, All (food)	Incredible Edible Egg | Eggs - what's inside the shell?
Welcome to the American Egg Board	Welcome to the American Egg Board
Incredible Edible Egg | Eggs - what's inside the shell?	Egg (food) - Wikipedia, the free encyclopedia
WHFoods: Eggs (food)	Easter Eggs - Eeggs.com
Easter Eggs - Eeggs.com	Egg (biology) - Wikipedia, the free encyclopedia
egg - Kitchen Dictionary - Recipezaar	egg - Kitchen Dictionary - Recipezaar

places higher than the places they were put without personalization.

8 CONCLUSIONS AND FUTURE WORK

We presented a personalization methodology which is based on clustering semantic user profiles. The method calculates users’ BMI, analyzes and annotates semantically the web access logs. At next it organizes the users’ profiles based on their choices and their BMI, and groups the users into clusters. The personalization of the results returned by the search engine is done by an on-the-fly semantic characterization and the score of each result is calculated. The scores of the results are kept in cache and the results are reorganized and presented to the user according to this score putting the one with the highest score first. By the experimental implementation we showed that the personalized method proposed has notably possibilities to change the scene in personalization. Future work includes the use of Fuzzy K-Means (Qiang, et al, 2008) that allows the creation of overlapping clusters, so that a user may belong to different cluster profiles with different weights or the use of K-Max algorithm. Also, the development of a reference ontology with more levels and alteration in factors such as the score of each result taking into consideration the user’s preference with greater weight than the rest users of the cluster.

REFERENCES

Almeida, R. and Almeida, V., 2004, A community-aware search engine. In *Proceedings of the World Wide Web Conference 2004*, pp. 423-421

Bose, A., Beemanapalli, K., Srivastava, J., Sahar. S., 2006, Incorporating Concept Hierarchies into Usage Mining Based Recommendations. *Proc. of WebKDD 2006: KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), Philadelphia, PA.*

Dai, H., Mobasher, B., 2002, Using Ontologies to Discover Domain-Level Web Usage Profiles, In *Proceedings of the 2nd Workshop on Semantic Web Mining, PKDD 2002*

Dou, Z., Song, R., Wen, J.R., 2007, A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the World Wide Web Conference 2007*, pp. 581-590

Eysenbach, G., Kohler, Ch., 2003, What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet, In *AMIA, Annual Symposium Proceedings Archive,* pp. 225-229

Freyne, J. and Smyth, B., 2006, Cooperating search communities, In *Proceedings of the AH’06,* pp. 101-110

Garofalakis, J., Giannakoudi, T., Vopi, A., 2008, Personalized Web Search by Constructing Semantic Clusters of User Profiles, In *Proceedings of 12th international Conference, KES 2008,* pp. 238-247

Garofalakis, J., Giannakoudi, T., 2009, Exploiting Ontologies for Web Search Personalization, In *Web Personalization in Intelligent Environments,* pp.49-64

Gauch, S., 2003, Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3), pp. 219–234

Kim, H., Chan, P. K., 2005, Personalized ranking of search results with learned user interest hierarchies from bookmarks. *WEBKDD 05 Workshop*, pp. 32–43

Lee, Y-J., 2005, VizSearch: A collaborative Web searching environment, *Computers & Education,* 44(4), pp. 423-439

Liu, F., Yu, C., Meng, W., 2002, Personalized web search by mapping user queries to categories. *Proceedings of the eleventh international conference on Information and knowledge management*, pp. 558–565

Liu, F., Yu, C., Meng, W., 2004, Personalized Web Search for Improving Retrieval Effectiveness

Liu, J., 2002, Resource-bounded online search for dense neighbourhood on the Web, PhD thesis, Canada

Ma, Z., Pant, G., Sheng, O., 2007, Interest-based personalized search, *ACM Transactions Information Systems*, 25(1)

Mei, Q. and Church, K., 2008, Entropy of search logs: How hard is search? With personalization? With backoff? In *Proceedings of WSDM' 08*

Mobasher, B., Cooley, R., Srivastana, J., 2000, Automatic Personalization based on web usage Mining, In *Communications of the ACM,* 43(8), pp.142-151

Nguyen, C., Phan, X., Horiguchi, S., Nguyen, T., Ha, Q., 2009, Web Search Clustering and Labeling with Hidden Topics, In *ACM Transactions on Asian Language Information Processing*, 8(3), Article 12

O' Conner, M., Cosley, D., Konstan, J., Riedl, J., 2001, PolyLens: A recommender system for groups of users. In *Proceedings of ESCW'01,* pp. 199-218

Pitkow, J., Schuetze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T., 2002, Personalized Search. In *Communications of the ACM*, 45(9), pp.50-55

Pretschner, A., Gauch, S., 1999, Ontology based personalized search. Master's thesis, University of Kansas, Electrical Engineering and Computer Science

Pretschner, A., Cauch, S., Chafee, J., 2003, Ontology-Based User Profiles for Search and Browsing, In *Web Intelligence and Agent Systems*, 1 (3-4), pp. 219-234

Qiang, W., Yunming, Y., Huang, J.Z., 2008, Fuzzy K-Means with Variable Weighting in High Dimensional Data Analysis, In *Web-Age Information Management 2008, WAIM'08,* pp. 365-372

Qui, F., Cho, J., 2006, Automatic identification of user interest for personalized search. In *Proceedings of the 15th International WorldWide Conference,* Edinburgh, Scotland, U.K., ACM Press, New York

Radlinski, F., Dumais, S., 2006, Improving Personalized Web Search using Result Diversification, In *Proceedings of the SIGIR'06*, Seattle, Washington, USA, 6-11 August 2006

Sendhilkuman, S., Geetha, T. V., 2008, Personalized Ontology for Web Search Personalization, In *Compute 2008*

Shen, X., Tan, B., Zhai, C. X., 2005, Implicit user modeling for personalized search, In *Proceedings of CIMK'05*, pp. 824-831

Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., Warke, Y., 2002, Managing semantic content for the Web. *Internet Computing, IEEE*, 6(4), pp. 80–87

Sieg, A., Mobasher B., Burke, R., 2007, Ontological user profiles for representing context in web search. *Web Intelligence and Intelligent Agent Technology Workshops, 2007 IEEE/WIC/ACM International Conferences*, pp. 91–94

Smith, G., Brailsford, T., Donner, C., Hooijmaijers, D., Truran, M., Goulding, J., Ashman, H., 2009, Generating unambiguous URL clusters from Web search, *Proceedings of WSDM'09*

Smyth, B., 2007, A community-based approach to personalizing Web search. In *IEEE Computer,* 40(8), pp. 42-50

Teevan, J., Dumais, S. T., Horvitz, E., 2005, Beyond the commons: On the value of personalizing Web search. In *Proceedings of SIGIR'05*, pp. 449-456

Teevan, J., Ringel Morris, M., Bush, S., 2009, Discovering and Using Groups to Improve Personalized Search. In *Proceedings of WSDM'09,* pp. 15-24

Tesic, J., Natsev, A., Smith, J., 2007, Cluster-based Data Modeling for Semantic Video Search, In *Proceedings of CIVR'07,* pp. 595-602

Yabo, X., Benyu, Z., Zheng, C., Ke, W., 2007, Privacy-Enhancing Personalized Web Search, In *Proceedings of the International World Wide Web Conference IW3C2*, Banff, Alberta, Canada, 8-12 May 2007

Zhuhadar, L., Nasraroui, O., 2008, Personalized Cluster-based Semantically Enriched Web Search for E-learning, In *Proceedings of the ONISW'08 Conference,* pp. 105-111