Personalizing Search Based on User Search Histories

Mirco Speretta
Electrical Engineering and Computer Science
University of Kansas
Lawrence, KS 66045

mirco@ku.edu

Susan Gauch
Electrical Engineering and Computer Science
University of Kansas
Lawrence, KS 66045
+1 (785) 864-7755

sgauch@ku.edu

ABSTRACT

User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles capture user information through proxy servers (to capture browsing histories) or desktop bots (to capture all activities on a personal computer). These both require participation of the user to install the proxy server or the bot. In this study, we explore the use of a less-invasive means of gathering user information for personalized search. In particular, we build user profiles based on activity at the search site itself and study the use of these profiles to provide personalized search results. In our study, we implemented a wrapper for Google to examine different sources of information on which to base the user profiles: queries and snippets of examined search results. These user profiles were created by classifying the information into concepts from the Open Directory Project concept hierarchy and then used to re-rank the search results.

User feedback was collected to compare Google's original rank with our new rank for the results examined by users. We found that queries were as effective as snippets when used to create user profiles and that our personalized re-ranking resulted in a 37% improvement in the rank-order of the user-selected results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models.

General Terms

Algorithms.

Keywords

User profiles, personalized search, conceptual search.

1. INTRODUCTION

Motivation

Personalization has been a very active research field in the last several years and user profile construction is an important component of any personalization system. Explicit customization has been widely used to personalize the look and content of many web sites, personalized search approaches focus on implicitly building and exploiting user profiles. Companies that provide marketing data report that search engines are utilized more and more as referrals to web sites, compared to direct navigation and web links [25] (i.e., StatMarket about WebSideStory product). As search engines perform a larger role in commercial applications, the desire to increase their effectiveness grows. However, search engines are affected by problems such as ambiguity and results ordered by web site popularity rather than user interests.

Conference'04, Month 1–2, 2004, City, State, Country.

Natural language queries are inherently ambiguous. For example, consider a user issuing the query "canon book". Due to ambiguity in the query terms, we will obtain results that are either religious or photography related. According to an analysis of 2 months of their log file data conducted by OneStat.com [21], the most common query length submitted to a search engine (32.6 %) is two words and 77.2% of all queries are three words long or less. These short queries are often ambiguous, providing little information to a search engine on which to base its selection of the most relevant Web pages among millions. A user profile that represents the interests of a specific user can be used to supplement queries, narrowing down the number of topics considered when retrieving the results. For the user in our example, if we knew that they had a strong interest in photography but little or none in religion, the photography-related results could be presented to the user preferentially.

Our approach is based on building user profiles based on the user's interactions with a particular search engine. For this purpose, we implemented GoogleWrapper: a wrapper around the Google search engine, that logs the queries, search results, and clicks on a per user basis. This information was then used to create user profiles and these profiles were used in a controlled study to determine their effectiveness for providing personalized search results.

The study was conducted through three phases:

1. collecting information from users. All searches, for which at least one of the results was clicked were logged per user.
2. creation of user profiles. Two different sources of information were identified for this purpose: all queries submitted for which at least one of the results was visited and all snippets visited. Two profiles were created out of either queries and snippets

3. evaluation: the profiles created were used to calculate a new rank of results browsed by users. The average of this rank was compared with Google's rank.

Many approaches create user profiles by capturing browsing histories through proxy servers or desktop activities through the installation of bots on a personal computer. These require participation of the user to install the proxy server or the bot. In this study, we explore the use of a less-invasive means of gathering user information for personalized search. Our goal is to show that user profiles can be implicitly created out of short phrases such as queries and snippets collected by the search engine itself. We demonstrate that profiles created from this information can be used to identify, and promote, relevant results for individual users.

2. BACKGROUND

2.1 Ontologies and Semantic Web

According to Gruber [11], an ontology is a "specification of a conceptualization". Ontologies can be defined in different ways but they all represent a taxonomy of concepts along with the relations between them. In the context of the World Wide Web, ontologies are important because they formally define terms shared between any type of agents without ambiguity, allowing information to be processed automatically and accurately. *OntoSeek* [12] is an example of system based on ontologies. Utilizing information sources such as product catalogs and yellow pages it applies conceptual graphs to represent both queries and resources.

The expression "Semantic Web" was introduced by ETAI (Electronic Transactions on Artificial Intelligence) in 2000 to describe the extension of the Web to deal with the meaning of available content rather than just its syntactic form. Many XML-based projects such as Resource Descriptor Framework (RDF), Notation 3 (N3), and OWL started from there and each aims to define a syntax capable of describing and/or manipulating ontologies. One of the main bottlenecks in the evolution of the Web along these lines is the amount of manual effort usually required to create, maintain, and use ontologies. Our approach shares many of the same goals as the Semantic Web, however we focus on automatic techniques wherever possible.

2.2 Personalization

Personalization is the process of presenting the right information to the right user at the right moment. In order to learn about a user, systems must collect information about them, analyze the information, and store the results of the analysis in a user profile. Information can be collected from users in two ways: explicitly, for example asking for feedback such as preferences or ratings; and implicitly, for example observing user behaviors such as the time spent reading an online document. Explicit construction of user profiles has several drawbacks. The user provide inconsistent or incorrect information, the profile built is static whereas the user's interests may change over time, and the construction of the profile places a burden on the user that they may not wish to accept. Thus, many research efforts are underway to implicitly create accurate user profiles [6][7][22].

User browsing histories are the most frequently used source of information about user interests. Trajkova and Gauch [26] use this information to create user profiles represented as weighted concept hierarchies. The user profiles are created by classifying the collected Web pages with respect to a reference ontology. Kim and Chan [15] also build user profiles from the same source, however they use clustering to create a user interest hierarchy. The collected Web pages are then assigned to the appropriate cluster. The fact that a user has visited a page is an indication of user interest in that page's content. Extending this idea, Chan describes a metric to estimate the level of user interest; for example the percentage of links visited on a page or URL presented in bookmarks.

To achieve effective personalization, profiles should distinguish between long-term and short-term interests and include a model of the user's context, i.e., the task in which the user is currently engaged and the environment in which they are situated [19]. Several systems have attempted to provide personalized search that are tailored based upon user profiles that capture one or more of these aspects.

In the OBIWAN project [8], search results from a conventional search engine are classified with respect to a reference ontology based upon the snippets summarizing the retrieved documents. Documents are re-ranked based upon how well their concepts match those that appear highly weighted in the user profile.

PERSIVAL [18] is a system that provides personalized search on specific medical libraries. Rather than building a user profile, PERSIVAL allows users to augment queries by providing contextual information such as a patient record. PERSIVAL then extracts concepts from the context and uses them to expand the query. The patient record is also used to filter the search results, removing information that is not related to the specific case described in the context. They have extended their personalized search to also be applied to multimedia information.

Competitive Intelligence Spider and Meta Spider [5] are part of a client-based application that collects and organizes Web documents on the user's machine. Spiders may gather information directly from Web sites or through search engines. Collected documents are then analyzed and noun phrases are extracted to create a personal dictionary for the user to guide future searches. The noun phrases are also used to organize the documents and a graphical map of the results is generated. Users can personalize the search explicitly by selecting specific Web sites, the number of Web pages to collect, and the noun phrases used in the final map of results.

The Personal Search Assistant [14] is an application that a background process that collects information on behalf of a user by submitting queries to various search engines. Results are stored on the local machine and are analyzed so that they can be organized conceptually. The user manually creates a conceptual database that is input to a personal agent responsible for building a user profile. The profile is used to filter the results of later searches.

3. APPROACH

Our study investigates the effectiveness of personalized search based upon user profiles constructed from user search histories. GoogleWrapper is used to monitor users activities on the search site itself in order to gather individual user information such as queries submitted, results returned (titles and snippets), and Web pages selected from results retrieved. This per-user information is classified into a concept hierarchy based upon the Open Directory Project [20], producing conceptual user profiles. Search results are also classified into the same concept hierarchy, and the match between the user profile concepts and result concepts are used to re-rank search results.

We believe this approach has several advantages. User interests are collected in a completely non-invasive way, search personalization is based upon data readily available to the search engine, and the system effectiveness can be evaluated by monitoring user activities rather than requiring explicit judgments or feedback.

3.1 System Architecture

The architecture of our system consists of three modules:

1. GoogleWrapper: a wrapper for Google that implicitly collects information from users. Google APIs [10] and nusoap library [23] were used for the implementation. Users register with their email addresses in order to create a cookie storing their userID on their local machines. If the cookie was lost, GoogleWrapper notified the user and they could login to reset the cookie. When queries are submitted by users, GoogleWrapper logs the query and the userID and then forwards the query to the Google search engine. It intercepts the search engine results, logs them, re-ranks them, and then displays them to the user. When users click on a result, GoogleWrapper logs the selected document along with the user ID before redirecting the browser to the appropriate Web page.
2. The classifier from KeyConcept [9], a conceptual search engine, is used to classify queries, snippets for each user as well as the search engine results. This vector space model classifier implements a k nearest neighbors algorithm.

A set of scripts that process the log files and evaluates the per-user and overall performance. The log file is split between users and, for each user, further divided into training and testing sets.

3.2 User Profiles

User profiles are represented as a weighted concept hierarchy. The concepts hierarchy is created from 1,869 concepts in the top 3 levels of the Open Directory Project and the weights represent the amount of user interest in the concept. The concept weights are assigned by classifying textual content collected from the user into the appropriate concepts using a vector space classifier and the k- nearest neighbor algorithm. The weights assigned by the classifier are accumulated over the text submitted.

In earlier work [2], we constructed user profiles from Web pages browsed by the user, however, this study focused on using the user's search history rather than their browsing history, information more easily available to search engines. We evaluate the effectiveness of profiles built from user queries with those built from snippets (titles plus the textual summaries) of user-selected results.

Each query or snippet was classified, resulting in a list of concepts and weights in decreasing order of weight. Since the number of concepts per item to add to the profile was unknown, we did a preliminary analysis of the classifier results for 40 queries submitted by 8 different users [24]. By manually judging the top 10 classifier results as relevant or not, we determined that the top 4 concepts assigned per query were relevant 75% of the time, dropping dramatically after that. A similar analysis for snippets determined that the top 5 classifier results were reasonably accurate. The increase in number of accurate concepts is likely due to that fact that snippets are longer than queries and thus can be classified more accurately and/or fit in more concepts. Based on these results, all profiles reported in this study were built using the top 4 concepts returned from the classifier for queries and the top 5 concepts for snippets.

3.3 Personalized Search

When a user submits a query to the search engine, and the titles, summaries and ranks results are obtained. The top 10 results are re-ranked using a combination of their original rank and their conceptual similarity to the user's profile. The search result titles and summaries are classified to create a document profile in the same format as the user profile. The document profile is then compared to the user profile to calculate the conceptual similarity between each document and the user's interests. The similarity between the document profile and the user profile is calculated using the cosine similarity function

$$sim(user_i, doc_j) = \sum_{k=1}^{N} wt_{ik} + wt_{jk}$$

where

wt_{ik} = Weight of $Concept_k$ in $UserProfile_i$

wt_{jk} = Weight of $Concept_k$ in $DocumentProfile_j$

The documents are re-ranked by their conceptual similarity to produce their conceptual rank. The final rank of the document is calculated by combining the conceptual rank with Google's original rank using the following weighting scheme:

FinalRank = α * ConceptualRank + (1-α) * GoogleRank

α has a value between 0 and 1. When α has a value of 0, conceptual rank is not given any weight, and it is equivalent to the original rank assigned by Google. If α has a value of 1, the search engine ranking is ignored and pure conceptual rank is considered. The conceptual and search engine based rankings can be blended in different proportions by varying the value of α.

4. EXPERIMENTAL VALIDATION

We monitored the search activities of six volunteers for approximately six months. All queries submitted per user were divided into 40 training queries, those used to create profiles, and 5 testing queries, those used to evaluate the profiles. Five testing queries were selected and up to 6 profiles were evaluated by their effectiveness for personalizing the search results measured by

comparing the rank order of the user-selected results with and without re-ranking based on the profile.

We conducted a preliminary study and examined 100 randomly selected queries from 10 different users. We found that 94% of the user-selected results occurred in the first 3 Google results and no result after the tenth result (i.e., on the second page) was ever selected. The number of clicks on the first result was also much higher than the number of clicks on the second one. The same decrease was observed between the second and the third results. We concluded that user judgments were affected by Google's rank so, for this study, we randomized the search engine results before presentation to the user. Since all results selected occurred within the first page, we only randomized the first ten results. The user clicks thus collected were analyzed later to compare how Google ranked the selected result versus how our system would have ranked it based upon the user profile.

4.1 Experiment 1

The first variable we investigated was the number of training queries necessary to create a profile based upon the query text alone. As mentioned in section 3.2, we used the top 4 concepts returned by the classifier for each query. We created user profiles using training sets of 10, 20, 30, and 40 queries. A second variable studied was the number of concepts from the resulting profile to use when calculating the similarity between the profile and the document. We varied this number from 1 through 20. Based on earlier experiments [3], we used the top 7 concepts for each search result.

The resulting profiles were evaluated based upon the conceptual rank of the user-selected result, without any contribution from Google's original ranking. The conceptual rank was compared to the original rank of the selected result to see if there was any improvement. Fig 1 shows the comparison between Google's original rank and conceptual rank averaged over all queries. These results were generated using 30 queries to build the profile, the best observed result, for a varying number of concepts from the profile used for the similarity calculation. We observe that the best improvement of 33% occurs when using only the top concept, 3.4 versus 5.1. This difference was significant ($p = 0.01$).

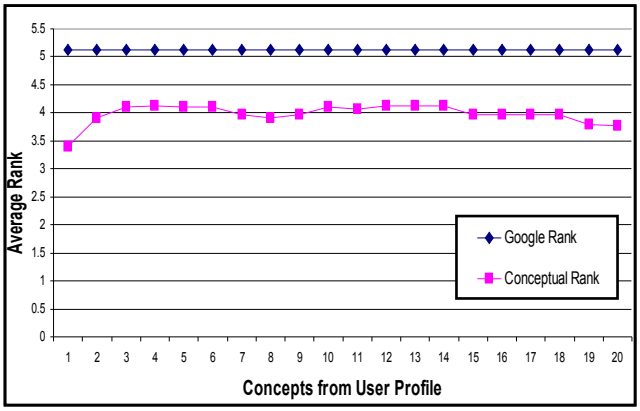

Fig 1: Google's Original Rank and Conceptual Rank Averaged over All Testing Queries. User Profiles are built using queries.

4.2 Experiment 2

This experiment repeats Experiment 1, the only difference being that the profiles were built using snippets rather than queries. Since text classified is, on average, longer than queries so we expected a bigger improvement. Once again, conceptual rank was used for evaluation and we used training sets of 10, 20, 30, and 40 snippets. As before, the best results occurred with 30 training snippets.

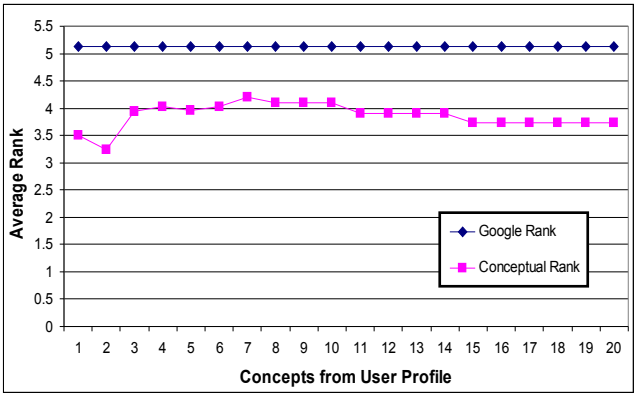

Fig 2: Google's Original Rank and Conceptual Rank Averaged over All Testing Queries. User Profiles are built using snippets.

Fig 2 shows the comparison between Google's rank and the conceptual rank as the number of concepts used from the profile is varied from 10 to 20. We observe that the best improvement of 37% occurs when using two concepts from the profile, 3.2 versus 5.1. This difference was significant ($p = 0.001$).

4.3 Experiment 3

In this experiment, we wanted to see if including the original rank returned by the search engine to the calculation of the final rank calculation could improve the overall results. We used the best conceptual rank found by Experiment 1, 30 training queries and 1 concept used from the profile. By varying the value of α in the FinalRank calculation, we varied the relative contributions of the conceptual and original rankings. Fig 3 shows that the best results are obtained when α is 1.0, i.e., when the original search engine rankings are ignored altogether. This is likely due to the fact that all the results on the first page match the query well and thus the distinguishing feature is how well they match the user's interests.

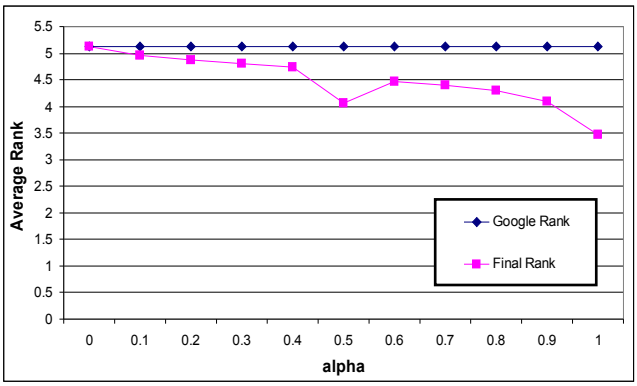

Fig 3: Plots Google's Original Rank and Final Rank Averaged over All Testing Queries. User Profiles are built using just queries.

We then examined the results best results on a query by query basis to get a more detailed understanding of the effects of our personalized search. For the 30 testing queries, 5 per user, re-ranked using a query-based profile, 10 (33%) showed an improvement, 17 (57%) were unchanged, and only 3 (10%) were negatively impacted. Thus, the personalized re-ranking helped 3 times as many queries as it hurt.

4.4 Experiment 4

Similar to the Experiment 3, we examined the effect of combining the search engine's original rank with the conceptual rank when calculating the final rank of a result. Based on Experiment 2, the conceptual rank was calculated using a profile built from 30 snippets and the top 2 profile concepts were used. Fig 4 shows the comparison between Google's rank and the final rank as α is varied from 0.0 to 1.0. Once again, the best results occur when α is 1.0, i.e., when the original search engine rankings are ignored altogether.

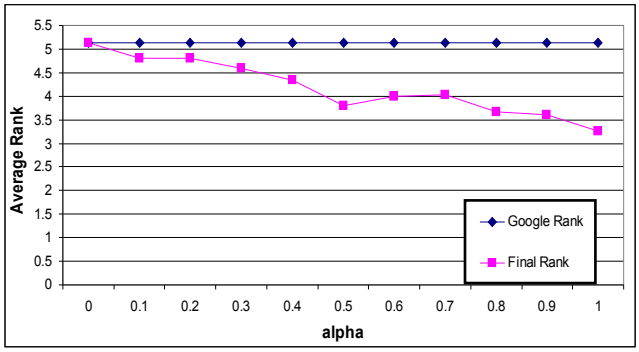

Fig 4: Plots Google's Original Rank and Final Rank Averaged over All Testing Queries. User Profiles are built using snippets.

Examining the results for the 30 testing queries re-ranked using a profile built from snippets, 12 (39%) showed improvement, 16 (53%) were unchanged, and only 2 (7%) were negatively impacted.

4.5 Experiment 5

To validate that best user profiles chosen by Experiments 1 and 2, can be used for queries other than those used to determine the settings, 2 queries that not previously seen were evaluated. We calculated the conceptual rank using both the query-based and snippet-based profiles and compared the conceptual rank to the original search engine rank. Table 1 summarizes these results, and verifies that we see comparable improvements for the validation queries as observed for the original test queries used to tune the profile creation algorithms.

Ranking Based On	Average Rank	Percent Improvement
Google (Original)	4.6	---
Conceptual (Queries)	3.0	34%
Conceptual (Snippets)	2.8	38%

Table 1: Comparison of Average Rank for Validation Queries

5. CONCLUSION AND FUTURE WORK

We built a system that creates user profiles based on implicitly collected information, specifically the queries submitted and snippets of user-selected results. We were able to demonstrate that information readily available to search engines is sufficient to provide significantly improved personalized rankings. We found that using a profile built from 30 queries produced an improvement of 33% in the rank of the selected result. A user profile built from snippets of 30 user-selected results showed a larger, but not significant, improvement of 37%. The snippet-based profile also improved more queries (12 versus 10) and hurt fewer (2 versus 3), so there is some indication that it is a slightly more accurate profile.

Our best results occurred when conceptual ranking considered only one concept from the query-based profile, and two from the snippet-based profile. This may be because the training and testing queries came from a relatively short window of time and users were working in a focused manner. However, the ranking improvements hold fairly steady across the evaluated range of 1 – 20 concepts used. For personalized results over a broader range of user queries, it would be safer to use more concepts from the profile.

The user profiles we used to build were based on a three-level deep concept hierarchy. We would like to examine the effect of using fewer or more levels of the ODP hierarchy as our profile representation. Also, the current concept hierarchy is static, and we would like to evaluate algorithms to dynamically adapt the hierarchy for specific users by merging and/or splitting concepts based upon the amount of user interest. Finally, we would like to combine the user profiles with the document selection process, not just the document re-ranking, to provide a wider set of relevant results to the user rather than just reorganizing the existing results.

6. REFERENCES

[1] D. Billsus, M.J. Pazzani. *A hybrid user model for news story classification*. In proceedings of the seventh international conference on User modeling, Banff, Canada, pp. 99 – 108, 1999.

[2] J. Chaffee, S. Gauch. *Personal Ontologies for Web Navigation*. In Proceedings of the 9th International Conference on Information and Knowledge Management. (CIKM), pp 227 – 234, 2000.

[3] V. Challam. *Contextual information retrieval using ontology based user profile*. Master's thesis, University of Kansas, Lawrence, KS, 2004.

[4] P.K. Chan. *A Non-Invasive Learning Approach to Building Web User Profiles*. KDD-99 Workshop on web usage analysis and user profiling, pp. 7 – 12, 1999.

[5] M. Chau, D. Zeng, H. Chen. *Personalized spiders for web search and analysis*. In Proceedings of the 1st ACM-IEEE Joint Conference on Digital Libraries, pp 79 - 87, 2001.

[6] C.C. Chen, M.C. Chen, Y. Sun. *PVA: a self-adaptive personal view agent*. In proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining 2001, San Francisco, California, pp. 257 – 262, 2001.

[7] M. Claypool , P. Le, M. Waseda, D. Brown. *Implicit Interest Indicators*. Proceedings of the 6th international conference on Intelligent user interfaces (ACM), pp 33 - 40, 2001.

[8] S. Gauch, J. Chafee and A. Pretschner. *Ontology-Based Personalized Search and Browsing, Web Intelligence and Agent Systems*. Vol. 1 No. 3-4, April 2004, pp.219-234.

[9] S. Gauch, D. Ravindran, S. Induri, J. Madrid, and S. Chadalavada, *Internal Technical Report ITTC-FY2004-TR-8646-37*, Information and Telecommunication Technology Center, University of Kansas

[10] Google APIs. http://www.google.com/apis.

[11] T. R. Gruber. *A translation approach to portable ontologies.* Knowledge Acquisition, Volume 5, Issue 2 (June 1993) Special issue: Current issues in knowledge modeling, pp. 199 – 220.

[12] N. Guarino, C. Masolo, G. Vetere. *OntoSeek: Content-Based Access to the web.* IEEE Intelligent System, Volume 14, no. 3, pp. 70 – 80, 1999.

[13] G. Jeh and J. Widom. *Scaling personalized web search*. In Proceedings of the Twelfth International World Wide Web Conference. In Proceedings of the twelfth international conference on World Wide Web, Budapest, Hungary, Pages: 271 - 279, 2003.

[14] P.R. Kaushik, K. Narayana Murthy. *Personal Search Assistant: a configurable personal meta search engine.* Fifth Australian World Wide Web Conference, Southern Cross University, Lismore, Australia, 1999.

[15] H.R. Kim, P.K. Chan. *Learning implicit user interest hierarchy for context in personalization.* In Proceedings of the 8th international conference on Intelligent user interfaces, Miami, Florida, USA, 2003, pp. 101 - 108.

[16] S. Lawrence. *Context in Web Search*. IEEE Data Engineering Bulletin, Volume 23, Number 3, pp. 25 – 32, 2000.

[17] F. Liu, C. Yu, W Meng. *Personalized web search by mapping user queries to categories*. In Proceedings of the 11th International Conference on Information and Knowledge management, McLean, Virginia, USA, 2002, pp. 558 - 565

[18] K.R. McKeown, S. Chang, J. Cimino, S.K. Feiner, C. Friedman, L. Gravano, V. Hatzivassiloglou, S. Jhonson, D.A. Jordan, J.L.Klavans, A. Kushniruk, V. Patel and S. Teufel. *PERSIVAL, a system for personalized search and summarization over multimedia healthcare information.* Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, pp 331 – 340, 2001.

[19] S. Mizzaro, C. Tasso. *Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web*. Lecture Notes in Computer Science. In proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, pp 306 – 316, 2002.

[20] The Open Directory Project (ODP). http://dmoz.org.

[21] OnStat.com. *Most people use 2 word phrases in search engines according to OneStat.com*. http://www.onstat.com, Feb 2004.

[22] M. Pazzani, J. Muramatsu and D. Billsus. *Syskill & Webert: identifying interesting Web sites*. In proceedings of the 13th National Conference on Artificial Intelligence (AAAI96), pp 54 – 61, 1996.

[23] PHP Library for Google APIs. http://dietrich.ganx4.com/nusoap/.

[24] M. Speretta. *Personalizing Search Based on User Search Histories*. Master's thesis, The University of Kansas, Lawrence, KS, 2004.

[25] StatMarket.com. *Search guiding more web activity*. http://www.statmarket.com, 2003

[26] J. Trajkova and S. Gauch, *Improving Ontology-Based User Profiles.* RIAO 2004, Vaucluse, France, April 26-28, pp. 380-389.

[27] Y. Yang. *A study on thresholding strategies for text categorization*. In proceedings of the 24th Annual International ACM SIGIR Conference on Research and development in Information Retrieval, pp 137 – 145, 2001.