Contextual Search Using Ontology-Based User Profiles.

Conference Paper · January 2007

Source: DBLP

CITATIONS
69

READS
161

3 authors, including:

Susan Gauch
University of Arkansas
128 PUBLICATIONS **3,804** CITATIONS

SEE PROFILE

Contextual Search Using Ontology-Based User Profiles

Vishnu Challam
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
vishnuc@microsoft.com

Susan Gauch
EECS Department
University of Kansas
Lawrence, KS 66045
sgauch@ittc.ku.edu

Aravind Chandramouli
EECS Department
University of Kansas
Lawrence, KS 66045
aravindc@ittc.ku.edu

Abstract

Search engines, generally, return results without any regard for the concepts in which the user is interested. In this paper, we present our approach to personalizing search engines using ontology based contextual profiles. In contrast to long-term user profiles, we construct contextual user profiles that capture what the user is working on at the time they conduct a search. These profiles are used to personalize the search results to suit the information needs of the user at a particular instant of time. We present the results of experiments evaluating the effect of the original versus conceptual ranking and the use of multiple sources of information to build the contextual profile. We were able to achieve a 15% improvement over Google in the average rank of the result clicked by a user when contextual information extracted from open Word documents and Web pages was used to re-rank the results.

1. Introduction

The huge amount of information available on the Internet is widely shared primarily due to the ability of Web search engines to find useful information for users. However, most search engines display results without any concern for the information needs of the user at a particular instance of time. The query "Wild Cats" returns the same results to a person searching for wild animals and a sports fan searching for information about his favorite team. Search engines lack a personalization mechanism that would understand the information needs of the user at a particular instance of time and return custom results. Personalization broadly involves the process of gathering user-specific information during interaction with the user, which is then used to deliver appropriate content and services; tailor-made to the users needs (Bonett, 2001). In this paper, we present our approach to personalizing Web search engines using ontology based contextual user profiles. In contrast to long-term user profiles, we construct contextual user profiles that capture what the user is working on at the time they conduct a search. We post-process the results of the Google search engine (Google, 2006a), making use of contextual information and compare the performance of our system with that of Google.

2. Related Work

Personalized search has seen a significant amount of research over the last few years. A key aspect of personalized search involves collecting information about the user's needs. This information can be collected by asking the users to specify their interest explicitly (Glover et al., 1999; Google, 2006b) or automatically using non-invasive approaches (Chan 1999; Shavlik et al., 1999). Usually, this information stored as user profiles represent the long-term users interest. However, there have been efforts to provide contextual search based on the user's current task (Budzik & Hammond, 2000). Researchers have also worked on combining both long-term and short-term user's interest (Widyantoro et al., 2000). More recently personalized search techniques have been investigated that build user profiles based on user browsing history (Gauch et al., 2004) or search histories (Speretta & Gauch, 2005). Our work is similar to (Budzik & Hammond, 2000) in that we use information from the user's current task to construct contextual

profiles. However, instead of using keyword profiles that are more sensitive to variations in keywords, we build a more robust profile based on weighted concepts selected from an ontology.

3. System Architecture and Implementation

The activity of a user on his machine is continuously monitored by a Windows application that captures content from open Internet Explorer, MS-Office and MSN messenger documents. The content captured during this process is stored on the client machine. This content is used to build a user's contextual profile.When the user submits a query, the content captured within a specific time is classified with respect to the ODP ontology (ODP, 2006). A detailed discussion on the classifier can be found in (Gauch et al., 2004). The classifier represents the user's contextual profile for the time window as a weighted ontology. The weight of a concept in the ontology represents the amount of information recently viewed or created by the user that was classified into that concept. When the user issues a query, their recently stored context is classified to create the user's contextual profile. This contextual profile is uploaded to the server along with the query. The query is submitted to the search engine and the titles, summaries and ranks of the top 10 results are obtained. The results are re-ranked using a combination of their original rank and their conceptual similarity to the user's contextual profile. The search result titles and summaries are classified to create a document profile in the same manner as the user's contextual profile. The document profile is compared to the contextual profile to calculate the conceptual similarity between each document and the user's context. The similarity between the contextual profile and the document profile is calculated using the cosine similarity function

$$sim(context_i, doc_j) = \sum_{k=1}^{N} wt_{ik} * wt_{jk} \quad (1)$$

where

wt_{ik} = Weight of Concept$_k$ in Context$_i$ and wt_{jk} = Weight of Concept$_k$ in document$_k$

The concept weights are calculated using the tf*idf formula used by the vector space model (Salton & McGill, 1983). The documents are re-ranked by their conceptual similarity to produce their conceptual rank. The final rank of the document is calculated by combining both key word rank and conceptual rank using the following weighting scheme

$$Final\ Rank = \alpha * Conceptual\ Rank + (1-\alpha) * Keyword\ Rank \quad (2)$$

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 pure keyword based ranking. If α has a value of 1, keyword based ranking is ignored and pure conceptual rank is considered. Both the conceptual and keyword based rankings can be blended by varying the values of α.

4. Experiments and Evaluation

4.1 Experiments

In order to test and evaluate the use of contextual profiles to personalize results from Web search engines, a wrapper around the popular Web search engine, Google, was built using the publicly available Google API (Google, 2006c). This wrapper program builds a log of the queries given by a user, the results returned by Google, the result on which the user clicked, and the summaries, titles and ranks of the results returned from Google. This log information was used to evaluate the performance of the system. For all experiments, the wrapper randomized the order

of the top 10 Google results before presenting them to the user so that the user would not be biased by the presentation order of the results. In order to evaluate the system, 5 users were asked to use the system to perform similar tasks. All 5 users were Computer Science graduates and were expert search engine users. Each was asked to use the system to help them write small essays on 6 different topics ranging from sports, to car purchasing, to jewelry. While the users were performing these tasks, the program described in section 3 was continually running in the background on their Windows machines, and capturing the content of the Web pages and the content typed into the Word documents. So that we could establish a context for the users, they were asked to at least start their essay before issuing any queries. They were also asked to look through all the results returned by Google Wrapper before clicking on any result. Result click-through was used as a form of implicit user relevance judgment in our analysis.

After the data was collected, we had a log of 50 queries averaging 10 queries per user. Of these 50 queries, 6 of them had to be removed, either because there were multiple results clicked, no results clicked, or there was no contextual information available for that particular query. The remaining 44 queries were analyzed and evaluated. Experiments were conducted to determine the number of concepts to be considered from the user's contextual profile, the number of concepts from the document summaries and the value of α for blending the conceptual rank and the original rank. The results from these experiments are presented in the next sub-section.

4.2 Evaluation

4.2.1 Representing the User's Context Using a Single Source

In this experiment, we compared the performance of queries performed with and without a contextual profile built from the content of Word documents and Web pages separately.

Experiment 1: Contextual queries using Word documents

After filtering queries for which Word context was available, there were 32 queries left for analysis. The queries were analyzed by trying different combinations of the number of concepts to use from user's contextual profiles and document profiles.

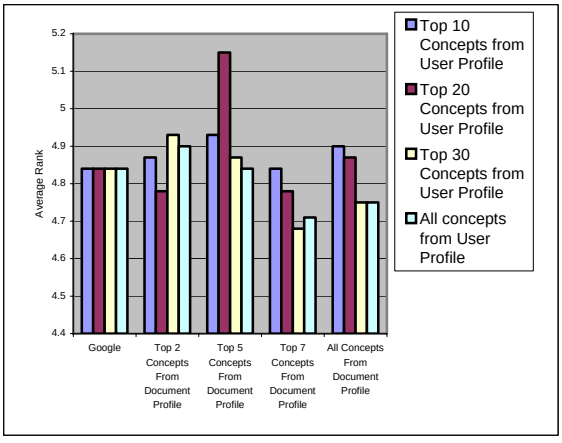

Figure 1: Profiles using Word documents

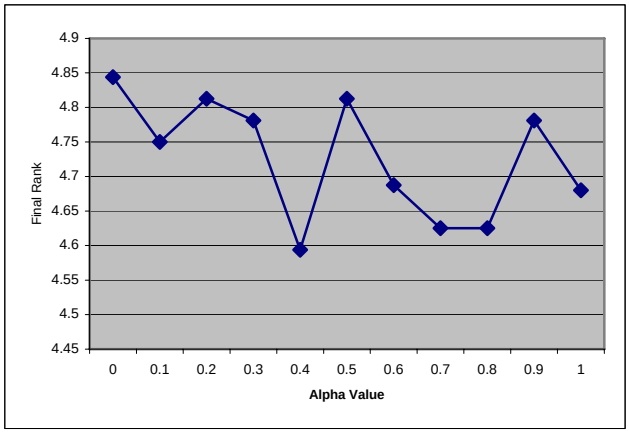

Figure 2: Effect of Alpha on Ranking using Word Documents

Figure 1 shows the average Google rank of 4.84 and average conceptual rank for the results clicked by the users. In this experiment, we varied the number of concepts used for the user's contextual profile and the document profiles. The best conceptual rank of 4.68 occurred when

using 30 concepts for the contextual profile and 7 concepts for the document profile. The final rank was calculated using equation 2. We then calculated the final rank for each document by combining the original and conceptual ranks and plotted the results obtained for various values of α. Figure 2 shows the final ranks obtained for various values of α. The best final rank of 4.59 was obtained when α had a value of 0.4. This is a 5.16 percent improvement over the performance of Google alone indicating that information from Word documents can be used to provide contextual information to improve Web queries.

Experiment 2: Contextual queries using Web pages

After filtering queries for which there was no Web content available, there were 31 queries left. Figure 3 shows the average Google rank and average conceptual rank for the results clicked by the users. In this experiment, we varied the number of concepts used for the contextual profile and the document profile.

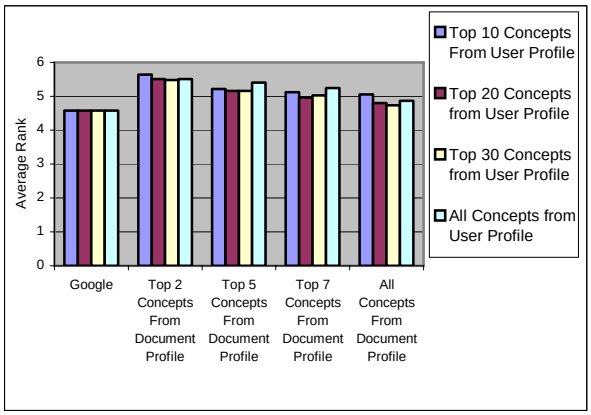

Figure 3: Profiles using Web content

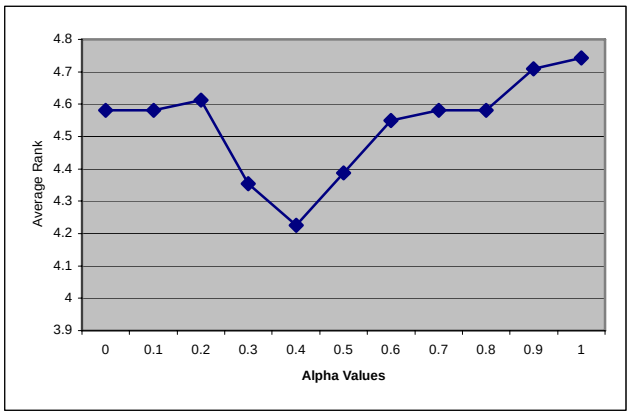

Figure 4: Effect of Alpha on ranking using Web Content

Based on the above analysis, it was found that when profiles are built using only the content from Web pages, the best average conceptual rank of 4.74 was obtained when top 30 concepts was used for the user's contextual profile and all concepts were used for the document profile . The final rank was calculated using these settings and formula 1 as before. Figure 4 shows the final ranks obtained for various values of α. Once again the best final rank of 4.22 was obtained when α had a value of 0.4. This is a 7.86% improvement over the original Google rank of 4.58. The results show that using information from either Web pages or Word documents to build contextual user profiles increases the performance of the system. The next series of experiments were conducted to analyze the performance of the system when content from different sources were combined to create the user's contextual profile.

4.2.2 Representing the User's Context Using a Combination of Sources

In this experiment, we compared the performance of queries performed with and without a contextual profile built from the combined content of Word documents and Web pages.

Experiment 3: Contextual queries using combined information

For these set of experiments, the final contextual profile was built based on the following formula:

$$Final\ Profile = \beta * Word\ Profile + (1 - \beta) * Web\ Profile \qquad (3)$$

where the Word Profile is the profile built from Word document content only and Web Profile is the profile built from Web page content only. When β is 0,the final profile is built using content from Web Pages only and when β is 1,the final profile is built using content from Word documents only. Varying the values of β between these two extremes will result in content from Web pages and Word documents being weighted differently. For the purpose of this analysis, the initial set of queries was filtered and only queries containing both Web pages and Word documents for contextual analysis were considered. Thus, 22 queries were analyzed. Based on the results from the previous experiments, the best conceptual rank is obtained when top 30 concepts from the contextual profile are considered, and the top 7 or all concepts from the document profile were considered. There was little drop off in accuracy for word based profiles when all document concepts were used, so when using a combined profile, we calculate the conceptual rank using the top 30 concepts from user's contextual profile and all concepts from the document profile. We varied α and β from 0.0 to 1.0 in increments of 0.1.

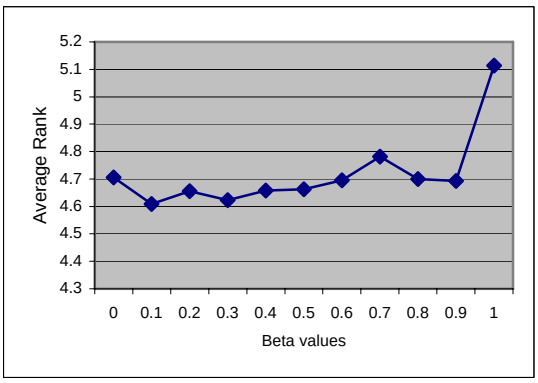

Figure 5: Effect of α on Final Rank

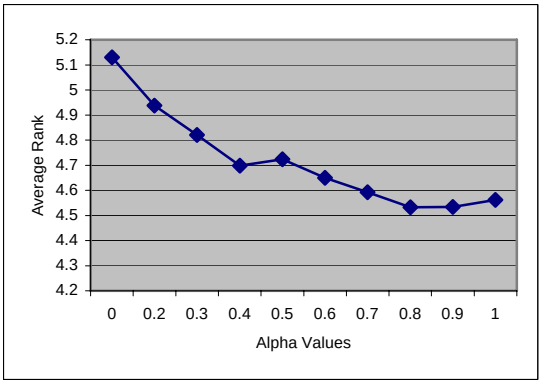

Figure 6: Effect of β on Final Rank

The best final rank of 4.36 is obtained when α has a value of 0.8 and β has a value of 0.1. This is a 15% improvement over the Google rank of 5.13. A β value of 0.1 means that 10% of the user's contextual profile is built from the Word content versus a 90% contribution from the Web documents. The α value of 0.8 indicates that the final rank is based 80% on the conceptual rank and only 20% on Google's original rank. Figure 5 and 6 shows the effect of α and β independently. The high value of α indicates that the conceptual rank should be given more weight than the search engine's rank. This may be because we are re-ranking among the top 10 results only, and they may match the user's query equally well. The primary distinguishing factor is therefore their conceptual similarity to the user's context. β values between 0.1 and 0.5 produce roughly comparable results, with the best value occurring with β = 0.1. The increased importance of Web content maybe because the Word documents created were very short and although we normalized for length in both cases, they just may not have contained enough content to build an accurate profile as compared to more comprehensive Web pages. If there was more content available from the Word documents a higher value of β might have been observed.

5. Conclusions and Future Work

In this paper we demonstrated that content captured from user activity can be used to build contextual user profiles and that these profiles can be used to improve Web searches.

Experiments were done to study the importance of the content from various sources, and the importance of conceptual ranking during personalization. Building a contextual profile using content from Web pages visited by the user resulted in larger improvements when compared to building a contextual profile using content from Word documents alone and when combining various sources, they should be weighed differently to build a better profile. When the content from Web pages and Word documents were weighed differently an improvement of 15% over Google was achieved. We also found that within the top 10 results of Google, re-ranking should be done giving more weight to the conceptual similarity between the user's contextual profile and the document than the original rank order. For our initial experiments, the contextual profile was built based on the most recent document of each type only. Studies need to be done to determine the best time window within which documents captured should be included in the contextual profile. Also, content from other sources such as chat transcripts, Excel spreadsheets etc. will be used to build the contextual profile. Finally, a combination of the user's current context and long and short-term interests will be investigated.

6. Acknowledgements

This work was partially supported by NSF ITR 0225676 (SEEK).

References

Bonett, M. (2001) Personalization of Web Services: Opportunities and Challenges. http://www.ariadne.ac.uk/issue28/personalization/

Budzik, J & Hammond, K.J. (2000). User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, (pp. 44--51), New Orleans, Louisiana

Chan, P. (1999) Constructing Web User Profiles: A Non-invasive Learning Approach. In *KDD-99 Workshop on Web Usage Analysis and User Profiling*, (pp. 7—12).

Gauch, S, Chaffee, J & Pretschner, A. (2004) Ontology Based User Profiles for Search and Browsing, *Web Intelligence and Agent Systems*, 1(3-4), 219-234.

Glover, E, Lawrence, S, Birmingham, W & Giles, C.L. (1999). Architecture of a metasearch engine that supports user information needs. In *8th International Conference on Information and Knowledge Management*, (pp. 210—216), Kansas City, Missouri, November.

Google Search Engine (2006a) http://www.google.com

Google Personalized Search (2006b) http://www.google.com/psearch/

Google API (2006c) http://api.google.com

The Open Directory Project (ODP) (2006) http://dmoz.org.

Salton, G & McGill, M.J. (1983). Introduction to Modern Information Retrieval, McGraw hill, New York.

Shavlik, J, Calcari, S, Eliassi-Rad, T & Solock, J. (1999). An Instructable, Adaptive Interface for Discovering and Monitoring Information on the World Wide Web. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces,* (pp. 157—160), Redondo Beach, CA.

Speretta, M & Gauch, S (2005). Personalized Search based on User Search Histories. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*,Compiegne University of Technology, France.

Widyantoro, H, Ioerger, T & Yen J (2000). Learning User Interest Dynamics with a Three-Descriptor Representation. *Journal of the American Society for Information Science*, 52(3), 212--225.