

Contextual Information Retrieval Using Ontology Based User Profiles

(Tracking Number: 135)

Abstract

Present day search engines are far from perfect because they return results based on simple keyword matches 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 instance 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.

Keywords

Ontologies, contextual information retrieval, personalization, user profiles, implicit measures.

1. Introduction

The huge amount of information available on the Internet is widely shared primarily due to ability of Web search engines to find useful information for users. However, present day search engines are far from perfect. They return results based on simple keyword matches 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 are lacking 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, that is then used to deliver appropriate content and services; tailor-made to the user's needs [Bonnet 2001]. When applied to search, personalization would involve the following steps:

1. Collecting and representing information about the user, to understand the user's interests.
2. Using this information to either filter or re-rank the results returned from the initial retrieval process, or directly including this information into the search process itself to select personalized results.

Thus, the problem of search engine personalization has two broad dimensions:

1. How can accurate information about the user's interests be collected and represented with minimal user intervention.
2. How can this information about the user be used to deliver personalized search results?

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 a popular search engine Google [Google], making use of contextual information and compare the performance of our system.

2. Related Work

2.1 Semantic Web

One way to provide conceptual search is to explicitly state the meaning of the content in a Web page. Research in this area tries to address the problem by having the creators of the content explicitly specify the meaning associated with a page using a knowledge representation language. Examples of knowledge representation languages are Ontobroker [Decker, et al. 1998], RDF [Lassila and Swick 1999], OIL [Fensel, et al. 2000] and SHOE [Heflin, Hendler and Luke 1999]. Many efforts are also underway to construct domain specific ontologies that can be used by Web content providers. Although this approach has the potential to provide conceptual search within certain communities, considering the size and democratic nature of the Web, it likely that a large proportion of Web content will remain plain HTML.

2.2 Ontologies

An ontology is a specification of a conceptualization [Griber 1993]. Sophisticated ontologies incorporate logical relationships and membership rules. However, concept hierarchies can also be used as simple ontologies. [Labrou and Finin 1999] use Yahoo! categories as a concept hierarchy and classifies documents into it using an n-gram classifier. The OBIWAN project [Chaffee, Gauch 2000][Zhu, et al. 1999] uses the Open Directory project's [ODP] concept hierarchy as ontology. This ontology has been used to represent the content of Web sites and as the basis of user profiles [Gauch, Chaffee and Pretschner 2004] for personalized search and browsing.

2.3 Construction and Representation of User Profiles

Research in this area is directed towards building user profiles using non-invasive approaches. These user profiles are a representation of the user's interests. Most such as Wisconsin Adaptive Web Assistant (WAWA)[Shavlik, et al. 1999], Syskill and Webert [Pazzani, et al. 1996] and Chan [Chan 1999] build

profiles non-invasively by observing which Web pages users visit over a period of time. They generally use the profile to suggest related Web pages to the users as they browse. Widyantoro, Ioerger and Yen [Widyantoro, et al. 2000] have developed a three-descriptor representation to monitor user interest dynamics. This model maintains a long-term interest descriptor to capture user's general interests and a short-term interest descriptor to keep track of user's more recent faster changing interests. Goecks and Shavlik [Goecks, Shavlik 1999] learn user's interests by looking at more than just the pages themselves. They also observe and measure user mouse and scrolling activity in addition to user browsing activity.

2.4 Contextual Search

Rather than building long-term user profiles, contextual systems try to adapt to the user's current task. Watson [Leake, et al. 1999] monitors users' tasks, anticipates task-based information needs, and proactively provide users with relevant information. The user's tasks are monitored by capturing content from Internet Explorer and Microsoft Word applications. Stuff I've Seen [Dumais, et al. 2003], developed at Microsoft Research indexes the content seen by a user and uses the index to provide easier access to information already seen by the user and also to provide rich contextual information for Web searches.

Our work differs from the related projects in that we build ontology-based contextual user profiles rather than keyword vectors. We also focus on search rather than assistance during browsing in our application.

3. System Architecture and Implementation

3.1 System Architecture

The main components of our personalization system are as follows:

1. A system to non-invasively monitor user activity on his/her machine by capturing content from open Internet Explorer, MS-Office and MSN messenger documents. The information captured from this system can provide a good estimate about the interests of a user at a particular instance of time.
2. A classifier to classify the content captured by the system to generate a user-profile that gives a description of the user's current context.
3. A system to test and evaluate the performance of the entire system.

The activity of a user on his machine is continuously monitored by the Windows application. The content captured during this process is stored on the client machine. When the user submits a query, the content captured within a specific time is classified with respect to the ODP ontology. The classifier used in our system is based on the vector space model and the manually associated Web pages in the ODP collection are used as training data. A detailed discussion on the classifier can

be found in [Gauch, Chaffee and Pretschner]. 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. The user's contextual profile will be used to personalize/re-rank the results to suit the users interests. The process of personalizing the results using information from the user's contextual profiles is described in greater detail in the next section.

3.2 Personalizing Search Results Using Information from User's Contextual Profiles.

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

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

where

wt_{ik} = Weight of Concept_k in Context_i

wt_{jk} = Weight of Concept_k in document_j

The concept weights are calculated using the tf*idf formula used by the vector space model [Salton and 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:

Formula 1:

$$\text{Final Rank} = \alpha * \text{Conceptual Rank} + (1-\alpha) * \text{Keyword Rank}$$

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 .

In the next section, we describe experiments to evaluate the best source of information for the user's context, the number of concepts to use for the contextual

profile, the number of concepts to use for the document profile and how best to weigh the original rank versus the conceptual similarity (the value 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 API]. 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 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 3.1 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 to Google Wrapper. They were also asked to look through all the results returned by Google Wrapper before clicking on any result. The Google wrapper recorded which results on which they clicked, which we used as a form of implicit user relevance 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. In each experiment we report the average rank of the user-clicked result for our baseline system, Google and for our conceptual search engine.

4.2 Evaluation

4.2.1 Evaluation and Analysis of using information from Word Documents or Web Pages to provide contextual information to Web queries.

For the purpose of this experiment queries were analyzed by building contextual profiles from the content of word documents and Web pages separately.

Experiment 1: Building user's contextual profiles using context from 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 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.

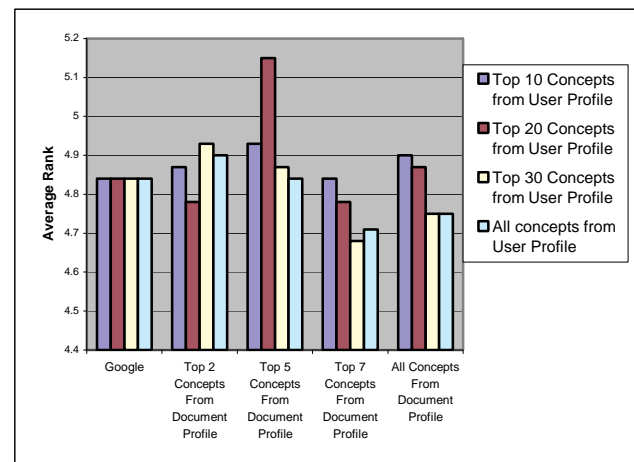


Figure 1: Contextual profiles based on Word document alone.

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 formula 1.

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 .

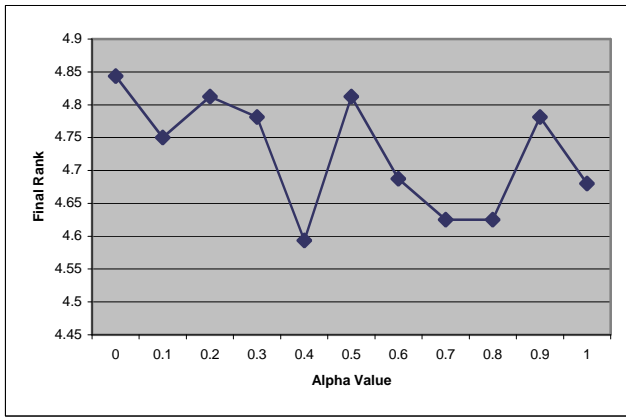


Figure 2: Effect of Alpha on Word-based Contextual and Google ranking

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: Analysis of queries by building user profiles using content from Web pages alone.

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.

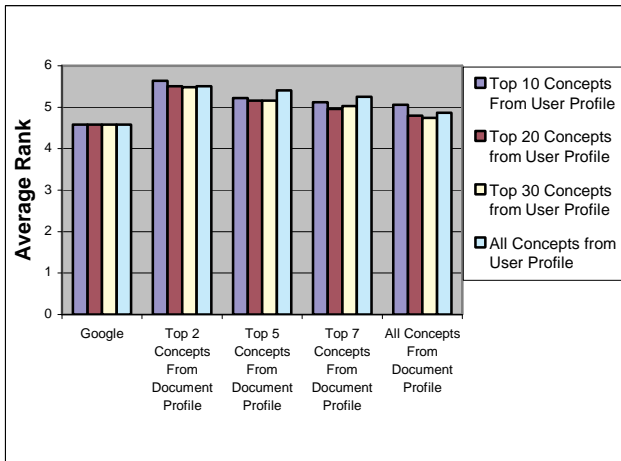


Figure 3: Contextual Profiles based on Web content alone.

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 .

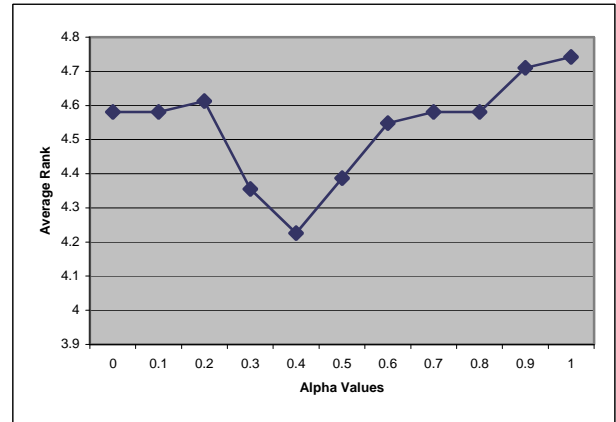


Figure 4: Effect of Alpha on Web-based Contextual and Google ranking

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 above 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

Experiment 3: Building contextual profiles by combining content from Word documents and Web pages.

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

$$\text{Final Profile} = \beta * \text{Word Profile} + (1 - \beta) * \text{Web Profile}$$

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 either 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. Figure 5 shows the effect of varying α and β on the combined profile.

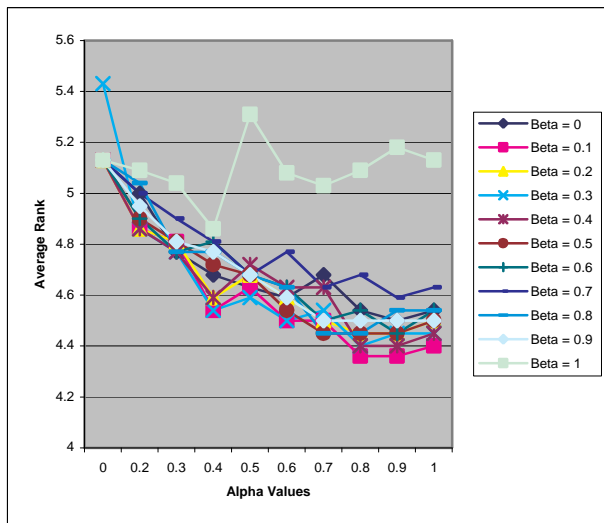


Figure 5: Effect of α and β when Contextual Profile is built from both Web and Word Documents

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. To study the effect of α and β independently the following two graphs that show the effect of α and β separately on the final rank were plotted.

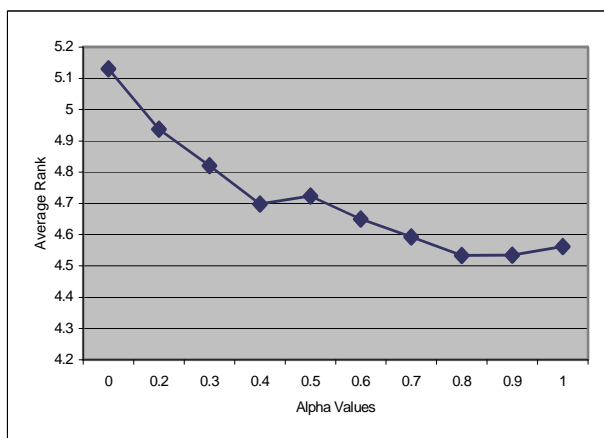


Figure 6: Effect of α on Final Rank when Contextual Profile is built from both Web and Word Documents

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.

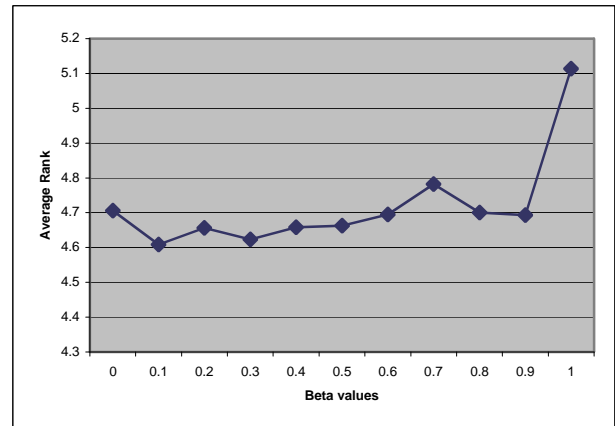


Figure 7: Effect of β on Final Rank when Contextual Profile is built from both Web and Word Documents

β values between 0.1 and 0.5 produce roughly comparable results, with the best value occurring with $\beta = 0.1$. 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 Word documents only resulted in a 5.16% improvement over Google and building a contextual profile using content from Web pages visited by the user resulted in a improvement of 7.86%. We found that 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.

In our experiments, most of the users were expert users of search engines, and the average query length was

around 4. Long queries tend to disambiguate themselves and result in better initial search results. It is possible that the improvements produced by the system would be more dramatic with shorter queries more common on the Web as a whole. In the system built, 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 various other sources such as chat transcripts, Excel spreadsheets, Power Point presentations etc. can be used to build the contextual profile and the effect of the content from these sources needs to be analyzed. Finally, a combination of the user's current context and long and short-term interests should be investigated.

References

[**Bonnet 2001**] Monica Bonett, Personalization of Web Services: Opportunities and Challenges. <http://www.ariadne.ac.uk/issue28/personalization/>

[**Chaffee, Gauch 2000**] Jason Chaffee, Susan Gauch. Personal Ontologies for Web Navigation. In *proceedings of the 9th International Conference on Information and Knowledge Management (CIKM), 2000*, pp 227-234.

[**Chan 1999**] Philip Chan: Constructing Web User Profiles: A Non-invasive Learning Approach. *KDD-99 Workshop on Web Usage Analysis and User Profiling*, pp. 7-12, 1999.

[**Decker et al. 1998**] Stefan Decker, Michael Erdmann, Dieter Fensel, Rudi Studer, Ontobroker: Ontology based Access to Distributed and Semi-Structured Information. *Proceedings of W3C Query Language Workshop QL'98. (1998)*

[**Dumais et al. 2003**] S. T. Dumais, E. Cutrell, E., J. J. Cadiz, G. Jancke, R. Sarin and D. C. Robbins (2003). Stuff I've Seen: A system for personal information retrieval and re-use. *Proceedings of SIGIR 2003*.

[**Fensel et al. 2000**] Dieter Fensel, Frank van Harmelen, Ian Horrocks, Deborah, OIL: An Ontology Infrastructure for the Semantic Web. *IEEE Intelligent Systems, Vol. 16, No. 2, March/April 2001*.

[**Gauch, Chaffee and Pretschner**] Gauch Susan, Chaffee Jason, and Pretschner Alexander, Ontology Based User Profiles for Search and Browsing, Web Intelligence and Agent Systems (In Press 2004).

[**Goecks, Shavlik 1999**] Jeremy Goecks, Jude Shavlik: Automatically Labeling Web Pages Based on Normal User Actions. In *Proceedings of the IJCAI Workshop on Machine Learning for Information Filtering, Stockholm, Sweden, July 1999*

[**Google**] <http://www.google.com>

[**Google API**] <http://api.google.com>

[**Gruber 1993**] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2): 199-220, 1993

[**Heflin, Hendler and Luke 1999**] Heflin, J., Hendler, J., and Luke, S. SHOE: A Knowledge Representation Language for Internet Applications. *Technical Report CS-*

TR-4078 (UMIACS TR-99-71), Dept. of Computer Science, University of Maryland at College Park. 1999

[**Labrou and Finin 1999**] Yahoo! as an ontology: using Yahoo! categories to describe documents. *Proceedings of the eighth international conference on Information and Knowledge Management, Kansas City, Missouri, 1999*.

[**Lassila, Swick 1999**] O. Lassila and R. Swick, Resource Description Framework (RDF) Model and Syntax Specification. *World Wide Web Consortium recommendation. 22 February 1999*.

[**Leake et al. 1999**] Leake, D., Scherle, R., Budzik, J., and Hammond, K. (1999). Selecting Task-Relevant Sources for Just-in-Time Retrieval. In *Proceedings of the AAAI-99 Workshop on Intelligent Information Systems*. AAAI Press, Menlo Park, CA, 1999

[**ODP**] The Open Directory Project (ODP), <http://dmoz.org>.

[**Pazzani et al. 1996**] M. Pazzani, J. Muramatsu, and D. Billsus, "Syskill & Webert: Identifying interesting Web sites," in *Proceedings of the 13th National Conference on Artificial Intelligence (AAA196)*, 1996, pp. 54--61.

[**Salton and McGill 1983**] G.Salton, M.J.McGill. *Introduction to Modern Information Retrieval, McGraw hill, New york, 1983*.

[**Shavlik et al. 1999**] Jude Shavlik, Susan Calcari, Tina Eliassi-Rad, Jack Solock: An Instructable, Adaptive Interface for Discovering and Monitoring Information on the World Wide Web. *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, pp. 157 - 160, Redondo Beach, CA.

[**Widyantoro 2000**] Dwi H. Widyantoro, Thomas R.Ioerger and John Yen, Learning User Interest Dynamics with a Three-Descriptor Representation. *Journal of the American Society for Information Science*, 52(3): 212-225, 2000.

[**Zhu et al. 1999**] Xiaolan Zhu, Susan Gauch, Lutz Gerhard, Nicholas Kral, Alexander Pretschner. Ontology-Based Web Site Mapping For Information Exploration. In *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM), 1999*, pp 188-194.

[**Zhu, Gauch 2000**] Zhu Xiaolan, Gauch Susan: Incorporating Quality Metrics in Centralize/Distributed Information Retrieval on The World Wide Web, Proc. of the 23rd International ACM SIGIR Conf. SIGIR '00 July 2000, Athens, Greece, pp. 288-295.