

Desktop Search - How Contextual Information Influences Search Results & Rankings

Wolfgang Nejdl
L3S and University of Hannover
Deutscher Pavillon Expo Plaza 1
30539 Hannover, Germany
nejdl@l3s.de

Raluca Paiu
L3S and University of Hannover
Deutscher Pavillon Expo Plaza 1
30539 Hannover, Germany
paiu@l3s.de

1. MOTIVATION

Sophisticated web search technology usually allows us to find appropriate documents in a few seconds. Finding these documents on our desktop is surprisingly more difficult, at least if we have been storing documents for a few years or more. This is improving somewhat with the recent crop of desktop search engines, but even with these tools, searching through our (relatively small set of) personal documents with the recent beta of Google Desktop Search is inferior to searching the (rather vast set of) documents on the web with Google. The main reason for this is that one of the distinguishing features of Google - sophisticated ranking using PageRank and other features - is unavailable on our desktop. This position paper explores first how the contextual information inferred from the users' actions is used to extend search beyond simple full-text search. Second, we discuss how algorithms exploiting this information can provide efficient ranking of resources on our desktop.

Regarding the first aspect, we propose activity-based metadata and relationships as sources of additional information in desktop search. This information, in many cases representing contextual information, is very useful for re-finding resources we already worked with, improving the recall in desktop search. For the second aspect, we discuss how PageRank-based ranking algorithms can exploit this contextual information and show how local and global ranking measures can be integrated in such a model, achieving at the same time personalization of rankings.

Using contextual information thus shows considerable promise for extending efficient information access to our desktop, extending globally available information with user-centered activity-based information, and exploiting the unique information background we have available on our desktop. We are currently implementing first prototypes in the context of the open source Beagle project which aims to provide sophisticated desktop search in Linux.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

2. REPRESENTING CONTEXT INFORMATION

2.1 Available desktop contexts

Current desktop search prototypes fall short of utilizing desktop specific information, especially context information. Three of these missed opportunities include:

Email context. Documents sent as attachments lose all contextual information as soon as they are stored on the PC, even though emails usually include additional information about their attachments, such as sender, subject or valuable comments. We might discuss a paper with a colleague during a brainstorming session, and then afterwards send her the electronic version via email, together with a few helpful comments. After a while, our colleague might not remember details about the paper itself, but rather recall with whom she discussed it or which question was raised in the discussion and included as comment in the email or email thread. We would like to find the stored paper not only based on its content, but also associatively based on that context information.

Browsing context. Browser caches include all information about user's browsing behaviour, which are useful both for finding relevant results, and for providing additional context for results. When searching for a document we downloaded from the CiteSeer repository, we would like to retrieve not only the specific document, but also all the referenced and referring papers which we downloaded on that occasion as well.

Publication context. Research activities represent one of the occupations where the need for contextual information is very high. The most illustrative example is the publication itself: Where did this file come from? Did we download it from CiteSeer or did somebody send it to us by email? Which other papers did we download or discuss via email at that time and how good are they (based on a ranking measure or on their number of citations)? We might remember the general topic of a paper and the person with whom we discussed it, but not its title. These questions arise rather often in a research environment and have to be answered by an appropriate search infrastructure. Personalized ranking on the desktop should take this contextual information into account as well as the preferences implicit in this information.

2.2 Scenario specific annotation ontologies

We will use ontologies to specify which context information we want to represent in the scenarios we address and RDF metadata to encode this information. Figure 1 presents our current prototype ontology, which specifies context metadata for emails, files, web pages and publications, together with the relations among them, described in more detail in [2]. Conceptually, the elements in the rectangles represent classes, circles represent class attributes. We

example, if we view a publication important because it was written by an author important to us, we have to represent that in our context ontology. Another example are digital photos, whose importance is usually heavily influenced by the event or the location where they were taken. In this case both event and location have to be included as classes in our context ontology.

4. DESKTOP SEARCH SCENARIO

We have considered a set of test scenarios for validating our assumptions about the benefits of our metadata enhanced search engine. Let us look at one of them in more detail. We assume that Bob and Alice are two team members of a computer science research institute, both being interested in semantic web technologies. Alice is currently writing a paper about searching and ranking on the semantic desktop and she wants to find some good papers on this topic, which she remembers she stored sometime ago on her desktop. Let us look at the results she will find.

The next three subsections discuss the scenario in more detail. We describe the type of hits our extended Beagle engine returns, show how ranking information is used to order the search results and discuss how we can enhance them with additional context information.

4.1 Direct Hits vs. Metadata Hits

The hits returned by any normal search engine (direct hits) would be the ones that contain the searched keywords in the title or in the content of the publication, in our case publications **A**, **B** and **D** (“Searching and Ranking on the **Semantic Desktop**”, “Using Semantic Web Technologies to Build a **Semantic Desktop**”, “Semantically Rich Recommendations in Social Networks for Sharing and Exchanging Semantic Context”). Traditional search engines would not be able to retrieve the other two publications. Publication **C** is included as email attachment, and not indexed by Beagle. However, the corresponding email text contains the searched keywords, and therefore we can retrieve it as an indirect hit (from the metadata as for each email we automatically store the email text as metadata for all attachments). Publication **E** is not even stored on the computer but is among the cited publications by another stored publication (**A**) which contains the keywords in the title, and therefore will be returned as a metadata hit (“The Social **Semantic Desktop**”).

We extend the Best interface provided by Beagle to include the metadata hits together with the direct ones. The direct hits are shown as Beagle normally does, with the occurrences of the searched terms emphasized. For the indirect hits we display the resources whose associated metadata include the query terms. Our example shows the first 5 out of a total of 20 results. As depicted in Figure 3, the first hit is an email having as attachment a publication not stored on the computer but referring to the topic in the body of the email, which contains the searched keywords. The last result in this picture is another indirect hit but the resource that is displayed is neither explicitly nor implicitly (e.g. in the email attachment) stored on the computer. This is why the user is redirected to the results provided by Google when searching for the item that it is pointed to by the context information/metadata.

4.2 Rankings

As Figure 3 shows, the results are displayed according to the computed resource rankings. The first hit has the highest rank among the resources in the result set. The rank values are also shown so that the user has an impression about how relevant the results are. In our case the most important result is the email Alice received from Bob including in its attachment a publication. Alice has exchanged her context with Bob, as well as with Caro-

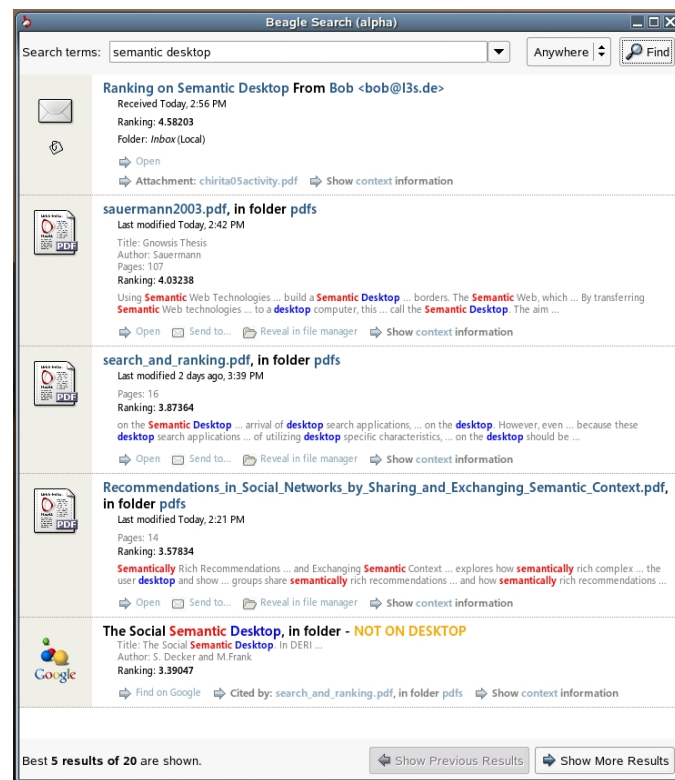


Figure 3: Beagle Main Window

line, Dan and Tom. All these persons have a high level of trust for Bob and therefore, by exchanging context information, in the resulting graph the node corresponding to Bob will have many incoming links. This translates into a high rank value for Bob. As we suggested in Figure 2, certain percentages of Bob’s rank will flow towards all nodes that Bob’s node points to. Since Bob is the sender of the email which was identified by Beagle as a match for the searched terms, the rank of this hit has a very high value.

We additionally observe that the last hit from this partial set of results, in spite of the fact that it is not stored on the desktop, has also a high rank because it is cited by the third hit of this query. Its high rank is also influenced by other resources this publication receives links from.

4.3 Metadata Visualization

Whenever a user clicks on the ‘Show context information’ link for a certain result, the corresponding metadata can be visualized, both for direct or indirect hits. A new window pops up displaying a list of details that correspond to the ontology related to the type of resource. Alice chooses to visualize the first result returned by Beagle, representing an email from Bob and having as attachment a publication. Since the interesting resource for this query is the PDF file in the attachment, the metadata window displays the annotations corresponding to publications together with other contextual information associated with it. The publication “Activity Based Metadata for Semantic Desktop Search” has 5 authors and for each of the authors we can further display the next level of metadata. For example, Alice can extend author S. Ghita and see other publications of this author. Additionally, she is able to see its referenced publications, the ones that cited it and that the publication was presented at the ESWC conference in 2005. Information

related to the provenance of this resource is also shown, the email it was saved from and its sender.

5. CONCLUSIONS AND RELATED WORK

This paper has explored two techniques - activity-based metadata and authority transfer annotations - as important contributions towards enabling efficient retrieval and ranking for the “personal digital repositories” building up on our computers. Activity-based metadata describe context information relevant for finding and connecting the resources we store on our desktop, authority transfer annotations exploit this context information to rank retrieved resources in a personalized way. Global ranking services like Google or Citeseer-derived ranking services can initialize these personalized ranking measures. Our prototype uses the open source project Beagle² as underlying desktop search infrastructure and extends its regular full-text indexing capabilities with contextual metadata and ranking.

Facilitating search for information the user has already seen before is also the main goal of the *Stuff I've Seen (SIS)* system, presented in [3]. Based on the fact that the user has already seen the information, contextual cues such as time, author, thumbnails and previews can be used to search for and present information. [3] mainly focuses on experiments investigating the general usefulness of this approach, though, without presenting more technical details. Based on SIS, [4] proposes a timeline-based visualization of search results over personal content. This basic timeline view is then augmented with public (holidays, news headlines) and personal (calendar appointments and digital photographs) landmark events. The main goal of this system is to facilitate browsing.

6. REFERENCES

- [1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, September 2004.
- [2] P. Chirita, R. Gavriiloaie, S. Ghita, W. Nejdl, and R. Paiu. Activity based metadata for semantic desktop search. In *Proceedings of the 2nd European Semantic Web Conference*, Heraklion, Greece, May 2005.
- [3] S. Dumais, E. Cutrell, JJ Cadiz, G. Jancke, R. Sarin, and Daniel C. Robbins. Stuff i've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th ACM SIGIR Conference*, Toronto, July 2003.
- [4] M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz. Milestones in time: The value of landmarks in retrieving information from personal stores. In *Proceedings of the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, Zurich, September 2003.

²<http://www.gnome.org/projects/beagle/>