# I know I stored it somewhere - Contextual Information and Ranking on our Desktop

Wolfgang Nejdl and Raluca Paiu

L3S Research Center and University of Hannover
Deutscher Pavillon, Expo Plaza 1, 30539 Hannover, Germany
{nejdl,paiu}@l3s.de

## 1 Motivation

Future digital libraries will be distributed, and recent research has already explored some promising approaches focusing on distributed and peer-to-peer search and retrieval architectures, connecting distributed repositories efficiently and transparently. Another aspect, which has been less explored so far, is the role of the implicit personal repositories we all have on our desktops, and the efficiency we expect to gain from storing important resources in these "personal digital libraries". Ironically, in many cases, moving important resources closer to our workspace results in less retrieval efficiency.

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 some of the information we have available on our desktop to extend search beyond simple full-text search, and the algorithms we can build upon this information, implementing 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 represents contextual information, which is very useful for re-finding resources we already worked with. We will describe corresponding metadata for two selected scenarios. For the second aspect, we will focus on recent advances of PageRank-based ranking, extending models such as the ones describing our contextual information, and show how local and global ranking measures can be integrated in such a model.

These two techniques together show 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.

## 2 Representing Context Information

*Available desktop contexts.* Current desktop search prototypes fall short of utilizing desktop specific information, especially context information. Two 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.

Documents stored from emails and from web sites form our personal digital library, which holds the papers we are interested in. We should store contextual information for these documents to retrieve them efficiently, and to restore the original context we built up when storing these documents. Personalized ranking on the desktop should take this contextual information into account as well as the preferences implicit in this information. [4] discusses how people tend to associate things to certain contexts. So far, however, search infrastructures neither collect nor use this contextual information.

*Scenario specific annotation ontologies.* 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 use classes whenever we want to attach importance/rank on entities, attributes otherwise.

Metadata for the email context refer to the date an email was sent or accessed, the subject of the email and the email body. Emails are connected to their attachments through the 1:n *has_attachment* relation. The *reply_to* relation represents thread information. *Person* represents the sender of an email. Another important relationship represents the connection between emails / email attachments and the documents they are saved as, as we want to use all attributes originally connected to the email a document was attached to when we want to find it again.

For the browsing context, we annotate each page with additional information about its basic properties (URL, access date, etc), as well as more complex ones such as in- and out-going links browsed. An extended publication ontology makes use of additional knowledge about how CiteSeer pages are connected and what they represent. Publications are referenced by other publications and can cite others, they can have a publication date / year associated with them, as well as a conference or journal. Publications have authors and are stored as documents on the desktop.

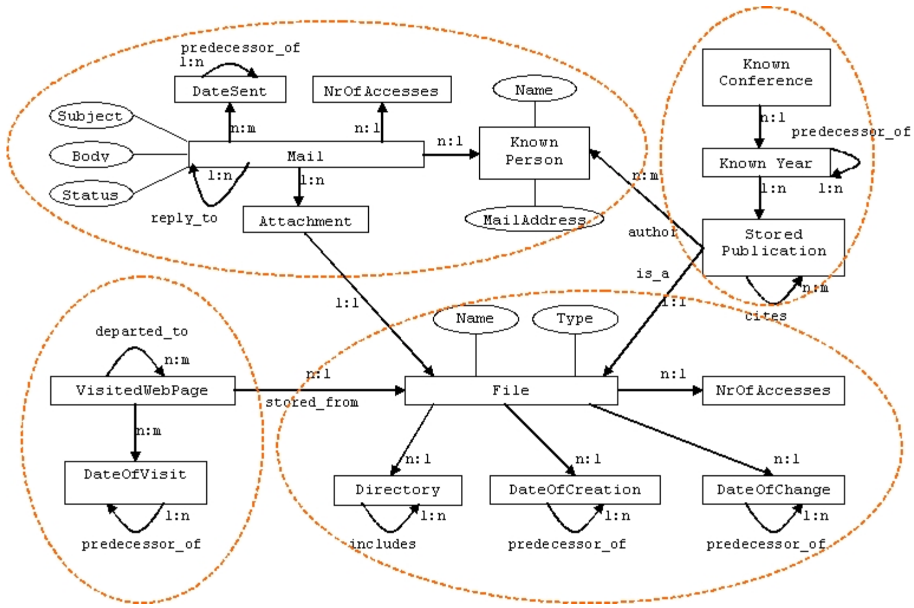


**Fig. 1.** Context ontology for our prototype

Other ontologies describe contexts like conferences, including reviewers, papers, meetings, authors, or private contexts like birthdays, including persons, locations, etc.

## 3 Authority Transfer and Ranking

*Authority transfer annotations.* Annotation ontologies should describe all aspects and relationships among resources, which influence the ranking. The identity of the authors for example influences our opinion of documents so "author" should be represented explicitly as a class in our publication ontology. Second, we have to specify how these aspects influence each others importance.

ObjectRank [1] has introduced the notion of authority transfer schema graphs, which extend schemas similar to the ontologies previously described, by adding weights and edges in order to express how importance propagates among the entities and resources inside the ontology. These weights and edges represent the authority transfer annotations, which extend our context ontologies with the information we need to compute ranks for all instances of the classes defined in the context ontologies.[1]

Figure 2 depicts our context ontology plus its authority transfer annotations. For example, authority of an email is split among the sender of the email, its attachment, the number of times that email was accessed, the date when it was sent and the email to which it was replied. So, if an email is important, the sender might be an important person, the attachment an important one and/or the number of times the email was

[1] In contrast to ObjectRank, we do not compute a keyword-specific ranking, but a global one.

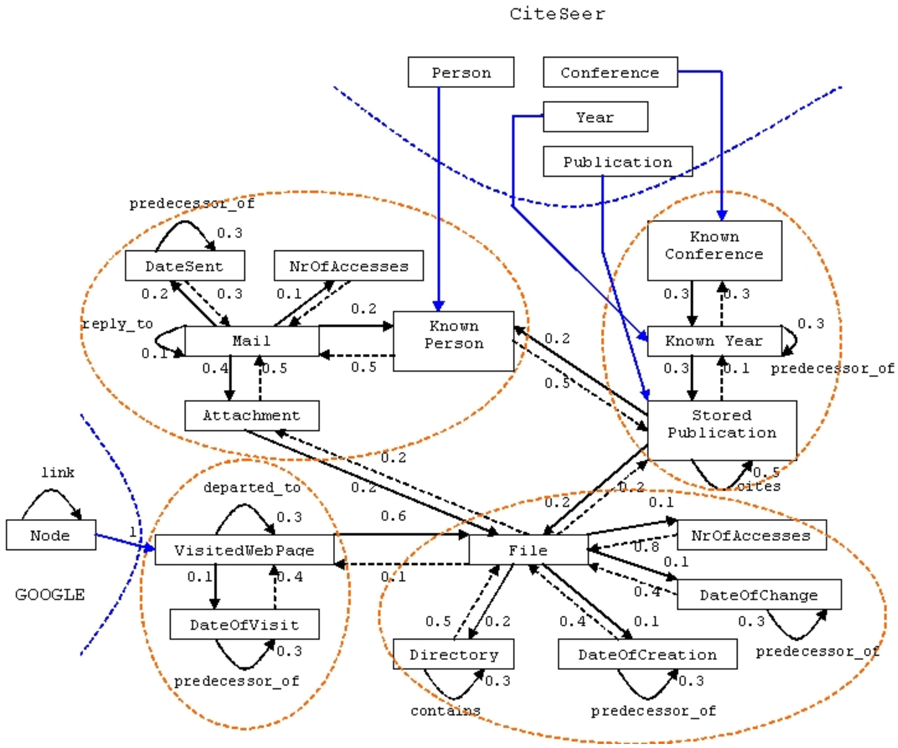


**Fig. 2.** Authority transfer annotations, including external ranking sources

accessed is very high. Additionally, the date when the email was sent and the previous email in the thread hierarchy also become important. As suggested in [1], every edge from the schema graph is split into two edges, one for each direction. This is motivated by the observation that authority potentially flows in both directions and not only in the direction that appears in the schema (if we know that a particular person is important, we also want to have all emails we receive from this person ranked higher). The final ObjectRank value for each resource is calculated based on the PageRank formula.

The ontology representing our web browsing context says that a visited webpage is important if we arrived at the current one from an important page, if the file under which it is stored is important, or if the date when the page was visited is important. For the CiteSeer context, publications transfer part of their authority to other papers they cite, to their authors, to the files under which they are stored, and to the year when the paper was published. As we can see, citing important papers doesn't make a paper important.

*Personalized Preferences and Ranking.* Different authority transfer weights express different preferences of the user, translating into personalized ranking. The important requirement for doing this successfully is that we include in a users ontology all concepts, which influence her ranking function. For 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.

*Personal vs. External Ranking.* For the computation of authority transfer, we can also include additional external ranking sources, connecting global ranking computation and personalized ranking of resources on our desktop. These external ranking sources are used to provide the seed values for the calculation of the personal ranking. Our prototype ontology includes two global ranking services, one returning Google ranks, the second one ranks computed from the CiteSeer database.

## 4 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 help 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 Beagle[2] 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, thubnails 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, [5] 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.

## References


1. A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, Toronto, September 2004.
2. P. Chirita, R. Gavriloaie, S. Ghita, W. Nejdl, and R. Paiu. Activity based metadata for semantic desktop search. Technical report, L3S, December 2004.
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 *SIGIR*, Toronto, July 2003.
4. Teevan J., Alvarado C., Ackerman M. S., and Karger D. R. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *CHI*, Vienna, April 2004.
5. M. Ringel, E. Cutrell, S. Dumais, and E. Horvitz. Milestones in time: The value of landmarks in retrieving information from personal stores. In *INTERACT*, Zurich, September 2003.


[2] http://www.gnome.org/projects/beagle/