

# A Proposal for Desktop Retrieval Track

Sergey Chernov<sup>1</sup>, Pavel Serdyukov<sup>2</sup>, Gianluca Demartini<sup>1</sup>, Paul-Alexandru Chirita<sup>1</sup>,  
and Wolfgang Nejdl<sup>1</sup>

<sup>1</sup> L3S Research Center, University of Hannover, Expo Plaza 1, 30539, Hannover, Germany,  
chernov, demartini, chirita, nejdl@l3s.de

<sup>2</sup> Database Group, University of Twente, PO Box 217, 7500 AE Enschede, Netherlands,  
serdyukovpv@cs.utwente.nl

## 1 Motivation

Every day we work with our personal textual information, e.g. with emails, working documents or presentation slides. The amount of documents on a single PC grew up to dozens and hundreds of thousands. Some of them are used frequently, others are needed only once a year, but all of them are stored on our hard drive. When a user is searching for the specific contract sample, saved Web page or for the batch of email attachments from her colleague, she might receive hundreds of hits for the issued query keywords. Apparently, it is extremely difficult to find a single piece of relevant information within this information heap.

A number of desktop search tools were released by major search engine vendors (Google, Microsoft, Yahoo etc.) recently. They mostly rely on query keyword occurrences in the documents, cause unlike on the Web, there is a lack of explicit inter-document connections on the desktop. It makes hard to apply such hyperlink-based analysis algorithms as PageRank to refine document ranking. On the other hand, the volume of unstructured information is gradually moving towards semi-structured representation. For example, the address book contains different metadata fields for personal contacts, email messages can be searched by date, sender or title, and so on.

There is a Personal Information Management (PIM) research community working on the desktop search problems. The evaluation of specific desktop search task has not been properly addressed in research so far since there is no data available. Clear interest of the community in the public dataset is reflected in the report from the first PIM workshop in January 2005. Until recently, the main difficulty in creating a public desktop dataset was the privacy issue. Our proposal contains suggestions how to overcome this problem. We imagine the dataset which contains a mixture of different text documents in popular file formats, email messages, personal archives. In addition, we plan this new dataset to include *activity logs* containing the history of each file, email or clipboard usage. This dataset will bring a basis for designing and evaluating of special-purpose retrieval algorithms for different desktop search tasks.

L3S Research Center in Hannover is heavily involved in desktop search research. For several previous publications from L3S [2] [3] [4] the temporary experimental settings were used, which made these experiments are neither repeatable, nor comparable. Since desktop is a rich source of context information, Twente Database Group working on the subject of context-aware IR is also highly interested in the proposed testbed. We

propose to combine our joint effort and suggest to include Desktop Search Track in TREC 2007, with coordinators:

- Sergey Chernov <chernov@l3s.de>
- Pavel Serdyukov <serdyukovpv@cs.utwente.nl>
- Gianluca Demartini <demartini@l3s.de>
- Paul-Alexandru Chirita <chirita@l3s.de>

## 2 Related Work

In last years several top-quality papers utilized desktop data and/or browsing activity logs for experimental evaluation. For example, in [8] authors used indexed desktop information (from 10,000 to 100,000) from 15 Microsoft employees of various professions with about 80 queries selected from their previous searches. In [6] Google search sessions of 10 computer science researchers have been logged for the period of 6 months and these users accessed documents for the 10 of their own queries. In [7], the Web search and browsing activity of 4 students have been logged for the period of 2 months and these users accessed documents for 15 of their own past queries. Several papers from Yahoo [5], Microsoft [1] and Google [9] presented approaches to mining their search engine logs for search personalization. We want to provide a common public dataset for this research community.

The most related dataset creation effort is the TREC-2006 Enterprise Track<sup>3</sup> (TRECENT). Enterprise search considers a user who searches the data of an organisation in order to complete some task. In particular, it considers the email search task. But this effort is different, since they are searching only on a mailing list archive, not within personal email. We need *personal documents* and *activity logs* for the dataset, while the Enterprise Track does not have explicit activity logs, like when the file was opened or edited, which files were opened simultaneously, etc. We also want to have a mixture of documents, which is representative for an average desktop, with different file formats and relationships between them. For example, email messages may have attachments, Web pages may be stored in the folder with other pages, saved at the same search session, etc. We believe that properties of personal data and corresponding search tasks are considerably different from those of enterprise data in general.

## 3 Data

### 3.1 File Formats and Metadata

The data for the Desktop Track would be collected among the participants. We are going to store several file formats: TXT, HTML, PDF, DOC, XLS, PPT, MP3 (tags only), JPG, GIF, and BMP. Each group in the track will submit 1-2 desktop dumps, using software from L3S Research Center and Twente University. This software consists of logging tools for a number of applications listed in the Table 1.

The set of logged applications can be extended in future. Loggers save the information which we describe in the Table 2 and Table 3.

<sup>3</sup> <http://www.ins.cwi.nl/projects/trec-ent/>

Application	File Format
Acrobat Reader	PDF files
MS Word	DOC, TXT, RTF
MS Excel	XSL
MS Powerpoint	PPT
MS Explorer	HTML
MS Outlook	PST
Mozilla Firefox	HTML
Mozilla Thunderbird	MSF and empty extension, mbox format

**Table 1.** Logged Application and File Formats

Permanent Information	Applied to
URL	HTML
Author	All files
Recipients	Email messages
Metadata tags	MP3, WMA
Has/is attachment	Emails and attachments
Saved picture's URL and saving time	Graphic files

**Table 2.** Permanent Logged Information

### 3.2 Data Gathering

As the privacy issue is very important here, we would like to propose two options for possible information gathering.

#### 1. Optimistic approach

We assume that there are volunteers ready to contribute some of their personal information to the community, given that this information would be redistributed only among the Desktop Track participants via NIST and track coordinators. As a test case in L3S Research Center, we gave two laptops to students for half a year. They were able to use them for free, but the condition was that all the information on these laptops will be available for future research. They were also warned not to store highly private information like passwords or credit card numbers. As this approach worked well, we expect that all participating groups will find similar reasonable incentives to attract more volunteers.

#### 2. Pessimistic approach

While some people are ready to share information with their close friends and colleagues, they do not like to disclose it to outsiders. For this scenario there is a way to keep information available only for a small number of people. Personal information is collected from participating groups by the track coordinators and preprocessed into publicly available uniform XML format. Every group can adapt their search prototypes to this format and submit binary files to the track coordinators. Then runs are produced locally by the track coordinator and results are sent back to the participants. Then, only trusted coordinators (or NIST personnel) have access to the actual documents, while it is possible for all participants to evaluate their results. This scheme has not been tested

Timeline information	Applied to
Time of being in focus	All files
Time of being opened	All files
Being edited	All files
History of moving/renaming	All files
Request type: bookmark, clicked link, typed URL	HTML
Adding/editing an entry in calendar and tasks	Outlook Journal
Being printed	All files
Search queries in Google/MSN Search/Yahoo!/etc.	Explorer/Firefox search fields
Clicked links	URL
Text selections from the clipboard	Text pieces within a file and the filename
Bookmarking time	Explorer/Firefox bookmarks
Chat client properties	Status, contact's statuses, sent filenames and links
Running applications	Task queue
IP address	User's address and addresses user connects to
Email status	Change between received/read

**Table 3.** Timeline Logged Information

yet in TREC, but it might be a necessary solution for future tracks, which involve highly private data (i.e. medical, legal, etc.).

#### 4 Relevance Assessments and Evaluation

As we are aiming at real world tasks and data, we want to reuse real queries from desktop users. Since every desktop is unique set of information, then its user should be involved in both query development and relevance assessment. It requires that desktop contributors should be ready to give 10 queries selected from their everyday tasks. The query should have the following format:

- <top>
- <num> KIS01 < /num>
- <query> Eleonet project deliverable June< /query>
- <metadataquery>date:June topic:Eleonet project type:deliverable< /metadataquery>
- <taskdescription>I am combining new deliverable for the Eleonet project.< /taskdescription>
- <narrative> I am looking for the Eleonet project deliverable, I remember that the main contribution to this document has been done in June. < /narrative>
- < /top>

We include <metadataquery> field so one can specify semi-structured parameters like metadata field names, in order to narrow down the query. The set of possible metadata fields would be defined after collecting the desktop data.

The desktop contributors must be able to assess pooled documents 6 months after they contributed the desktop. Moreover, each query will be supplemented with the description of context (e.g. clicked/opened documents in the respective query session), so that users could provide relevance judgments according to the contemporary context

of the query. Users mostly know their documents, so the assessment phase should go faster than normal TREC assessments. For the known-item search the assessments are quite easy, since only one (at most several duplicates) document is considered relevant. For the adhoc search task we expect users to spend about 3-4 hours to do relevance assessment per one query.

## **5 Proposed Track**

### **5.1 AdHoc Retrieval Task**

Ad hoc search is the most classic type of text search when user believes that she has relevant information somewhere. She thinks that there are several documents which contain pieces of necessary information, but she does not remember whether or where she stored them and she is not sure which keywords are best to find them. After the user entered a query, she can be satisfied if she meets maximum number of relevant documents among top-k resulting documents.

### **5.2 Known-Item Retrieval Task**

Targeted or known-item search task is peculiar to Desktop Search environment to a quite considerable extent. This is the task where user wants to find a specific document on the desktop, but doesn't know where it is stored or what is its exact title. This document can be an email or a working paper. Two approaches for known-item querying are possible.

1. First one, like in case of ad hoc search, assumes that user specifies only keywords that she believes to be in the searched document. Simple structured queries using *title*, *body*, *author* parts could be allowed.
2. Second one considers that the user has some knowledge about the context in which the document has been used before. Possible additional query fields then are: time period, location, topical description of the task in which scope the document had been used.

After the user entered a query, she can be satisfied if the right document is retrieved at or near rank one.

### **5.3 Folder Retrieval Task**

It is very popular among users to have their personal information topically organized in folders. Later they may search not for the specific document, but for the group of documents in order to use it later as a whole - browse them manually, reorganize or send to a colleague. The retrieval system should be able to estimate the relevance of folders and sub-folders using simple keyword query.

## 6 Evaluation measures

An important decision to be made is about which evaluation measures we need to compute in order to present the results to the participants and assess the performance of the systems.

To compute the evaluation measures we plan to use the tool `trec_eval` in order to compute the measures common to all the TREC tracks such as Precision, Recall, Precision at 11 recall points, Mean Average Precision (MAP), P@N, Precision/Recall curve, R-Precision. For this reason the relevance assessments will be binary (A hit is relevant or not).

Then, as we recognize the ranking of the results as the most important thing to evaluate, we will compute a measure which considers the information about the ranking produced by the systems, namely `bpref`. Results would be macro-averaged among all the desktops in the dataset.

## 7 Conclusions

While the many of desktop search tasks in particular are similar to enterprise search, they still dramatically differ by the highly private nature of stored information. A desktop search can benefit from the rich user activity information, which is not included in any of current TREC tracks, as well as from the Internet metadata residing within certain types of desktop documents. The L3S Research Center together with Twente Database group propose to use our logging tools and a new data gathering scheme for building the first real snapshot of several desktops and include it into the TREC 2007.

## References

1. Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10, New York, NY, USA, 2006. ACM Press.
2. Paul-Alexandru Chirita, Stefania Costache, Wolfgang Nejdl, and Raluca Paiu. Beagle<sup>++</sup>: Semantically enhanced searching and ranking on the desktop. In *ESWC*, pages 348–362, 2006.
3. Paul Alexandru Chirita, Andrei Damian, Wolfgang Nejdl, and Wolf Siberski. Search strategies for scientific collaboration networks. In *P2PIR'05: Proceedings of the 2005 ACM workshop on Information retrieval in peer-to-peer networks*, pages 33–40, New York, NY, USA, 2005. ACM Press.
4. Paul Alexandru Chirita, Julien Gaugaz, Stefania Costache, and Wolfgang Nejdl. Desktop context detection using implicit feedback. In *In Proceedings of the Workshop on Personal Information Management held at the 29th ACM International SIGIR Conf. on Research and Development in Information Retrieval*. ACM Press, 2006.
5. Reiner Kraft, Chi C. Chang, Farzin Maghoul, and Ravi Kumar. Searching with context. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 477–486, New York, NY, USA, 2006. ACM Press.

6. Feng Qiu and Junghoo Cho. Automatic identification of user interest for personalized search. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 727–736, New York, NY, USA, 2006. ACM Press.
7. Bin Tan, Xuehua Shen, and ChengXiang Zhai. Mining long-term search history to improve search accuracy. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 718–723, New York, NY, USA, 2006. ACM Press.
8. Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456, New York, NY, USA, 2005. ACM Press.
9. Beverly Yang and Glen Jeh. Retroactive answering of search queries. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 457–466, New York, NY, USA, 2006. ACM Press.