

Converting Desktop into a Personal Activity Dataset

Sergey Chernov, Enrico Minack

L3S Research Center
Hannover, Germany
chernov@L3S.de, minack@L3S.de

Pavel Serdyukov

University of Twente,
Enschede, The Netherlands
serdyukovp@cs.utwente.nl

Abstract

The current experiments on personalization in information retrieval are limited to the available collections of the real world data. While a number of publications exploited user interaction with Desktop, often these experiments are neither repeatable nor comparable. In this paper we elaborate on the need for logging the Desktop activity data and creating a common collection for Desktop search evaluation. We describe the design of such a dataset and necessary logging tools. We also outline the current state of our Personal Activity Track initiative towards creation of the Desktop search dataset. While this effort is currently targeting English-speaking users, it is also applicable to Russian and other languages.

1 Introduction

People tend to store more and more information on their personal computers. Consequently, a number of Desktop search tools were released by some major Web search engine vendors (Google, Microsoft, Yahoo etc.) to facilitate resource finding within this enlarging information management system. Yet the current modern search technologies are not entirely applicable on the PC Desktop environment, due to the big difference between Desktop and Web data. First, the data is structured in different ways. For example, there are no explicit links between Desktop resources, such that link analysis techniques cannot be used directly. Second, 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. Finally, the information seeking on a Desktop has different focuses than that on the Web. For example, people often seek for a previously known item on a Desktop, which makes the historical data rather important.

The development of search algorithms and tools for the Desktop requires an appropriate test collection accepted by the community [1]. However, no such dataset is available. The high privacy of user data and its heterogeneity across multiple desktops make the task of creating a dataset particularly aimed at the PC Desktop environment more challenging than in other more open milieus, such as the Web. Additionally, as many desktop resources are accessed within some given activity context, one must be able to reconstruct these contexts in order to exploit them for IR tasks (i.e., using metadata annotations, access links, etc.).

The creation of the testbed for experiments with personalized search is a difficult task, highly complicated because of privacy concerns. This paper describes the ongoing work towards a common dataset based on users' desktop information. We present a possible dataset design and ways for collecting the personal information. We also outline the discussion points for the future work.

2 Related work

Interesting research results were obtained in Personal Information Management (PIM) field in last years. This topic was developed within the information retrieval, database management, human-computer interaction and semantic Web communities. Recently, a number of interesting papers utilized Desktop data and / or activity logs for experimental evaluation. For example, in [2], authors used indexed Desktop resources (i.e., files, etc.) from 15 Microsoft employees of various professions with about 80 queries selected from their previous searches. In [3], Google search sessions of 10 computer science researchers have been logged for 6 months to gather a set of realistic search queries. Similarly, several papers from Yahoo [4], Microsoft [5] and Google [6] presented approaches to mining their search engine logs for personalization. In other relevant papers [7, 8] the temporary experimental settings were used, which made these experiments are neither repeatable, nor comparable. We want to provide a common Desktop specific dataset for this research community.

The most related dataset creation effort is the TREC-2006 Enterprise Track¹. Enterprise search considers a user who searches the data of an organisation in order to complete some task. The most relevant analogy between the Enterprise search and

Desktop search is the variety of items which compose the collection (e.g., in the TREC-2006 Enterprise Track collection e-mails, cvs logs, Web pages, wiki pages, and personal home pages are available). The most prominent difference between the two collections is the presence of *personal documents* and especially *activity logs* (e.g., resource read / write time stamps, etc.) within the Desktop dataset.

In this paper we present an approach we envision for generating such a Desktop dataset. We plan our 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. Current work extends our original proposal presented in [9].

3 Dataset design

3.1 File formats and metadata

The data for the Desktop dataset will be collected among the participating research groups. We are going to store several file formats: TXT, HTML, PDF, DOC, XLS, PPT, MP3 (tags only), JPG, GIF, and BMP. Then, each group willing to test its system would locally collect several Desktop dumps, using logging tools for a number of applications like Acrobat Reader, MS Office, Internet Explorer, Mozilla Firefox and Thunderbird, while the set of logged applications can be extended in the future. Loggers save the information which we describe in Table 1.

Permanent Information	Applied to
URL	HTML
Author	All files
Recipients	Email messages
Metadata tags	MP3
Has/is attachment	Emails and attachments
Saved picture's URL and saving time	Graphic files
Timeline information	
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.	Search fields in browsers
Clicked links	URL
Text selections from the clipboard	Text pieces within a file
Bookmarking time	Bookmarks in Internet browsers
Chat client properties	Status, sent filenames and links
Running applications	Task queue
IP address	User's address
Email status	Emails

Table 1: Timeline and Permanent Logged Information

3.2 Data gathering

As the privacy issue is very important here, we should address it already on the stage of data gathering. While some people are ready to share information with their close friends and colleagues, they do not like to disclose it to outsiders. In this case, there is a way to keep information available only for a small number of people: personal data is collected from participating

groups by local coordinators and pre-processed into the uniform XML format.

Every group can adapt its search prototypes to this format and submit binary files to the coordinators. Runs are then produced locally by a coordinator and results are sent back to the participants. This way, only trusted coordinators have access to the actual documents, while it is possible for all participants to evaluate their results. Similar scheme has been tested in TREC Spam Track, and it might be a necessary solution for future TREC tracks as well, whenever they involve highly private data (i.e. medical, legal, etc²).

Currently, the Windows XP version of the logger prototype is available for download from the Personal Activity Track Webpage³. The logger already supports some of the functionality described above, while it is regularly updated and a new release is scheduled for October 2007. In addition, we collect information from desktops of L3S employees, for a local dataset for internal experiments (see Section 7).

An important property of the proposed approach for dataset creation is that it can be applied to any language. For example, the members of Russian Information Retrieval Evaluation Seminar (ROMIP)⁴ can use this methodology and tools for creating a similar collection of personalized desktops with documents in Russian.

4 Search tasks

One of the current issues is a consensus in the community, what set of tasks to be evaluated. Among possible information retrieval tasks we envision: Ad Hoc retrieval, Folder Retrieval (i.e., ranking personal folders), Known-Item Retrieval, etc. The discussion is also open for Context Related Items Retrieval, using example items or keyword queries, Information Filtering, Email Management and related tasks. It is also interesting what kind of advanced search criteria users need. As a starting point, we show some examples of simple search tasks.

4.1 Ad Hoc Retrieval Task

Ad hoc search is the classic type of text retrieval when the user believes she has relevant information somewhere. Several documents can contain pieces of necessary data, but she does not remember whether or where she stored them, and she is not sure which keywords are best to find them.

4.2 Known-Item Retrieval Task

Targeted or known-item search task is the most common for the Desktop environment. Here the user wants to find a specific document on the Desktop, but does not know where it is stored or what is its exact title. This document can be an email, a working paper, etc. The task considers that the user has some knowledge about the context in which the document has been used before. Possible additional query fields are: time period, location, topical description of the task in which scope the document had been used, etc.

4.3 Folder Retrieval Task

It is very popular among users to have their personal items topically organized in folders. Later they may search not for a specific document, but for a 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 queries.

5 Queries

As we are aiming at real world tasks and data, we want to reuse real queries from Desktop users. Since every Desktop is a unique set of information, its user should be involved in both query development and relevance assessment. Thus, Desktop contributors should be ready to give 10 queries selected from their everyday tasks. This also solves the problem of subjective query evaluation, since users know best their information needs.

In this setting queries are designed for the collection of a single user, but some more general scenarios can be designed as well, for example finding relevant documents in every considered Desktop. It is thus possible to see the test collection as partitioned in sub-collections that represent single Desktops with their own queries and relevance assessments. This solution would be very related to the MrX collection used in the TREC SPAM Track, which is formed by a set of emails of an unknown person.

The query can have the following format:

- `<num> KIS01 </num>`
- `<query> Eleonet project deliverable June </query>`
- `<metadataquery> date:June topic:Eleonet project type: deliverable </metadataquery>`
- `<taskdescription>I am combining a 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>`

We include the `<metadataquery>` field so that one could 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 actual context of the query. As users know their documents very well, the assessment phase should go faster than normal TREC assessments. For the task of 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 query.

6 Discussion

There are several important questions which are not solved yet and they require an additional discussion within the community. We suggest the following directions of such a discussion.

- **Data and Privacy.** It is difficult to select appropriate data to build a testbed collection for experiments with personalization. There are several issues to be investigated in this concern like: (1) Privacy implications and data anonymization, (2) Storage and accessibility of test data, (3) Information sources (here, one of our major interests goes toward analyzing and discussing the logging of personal activities), etc. The discussion should also consider the personal data privacy problem both at the stages of data gathering and document relevance assessment. We would like to find out what the perfect collection is and what is the best way to interact with it? How the collection should be composed? Which information to include in the personal application activity logs? How to manage the privacy issues for the sharing the data?

- **Loggers and Test Applications.** This aspect is more focused on how we can collect necessary data and what kind of technical infrastructure should be implemented for PIM evaluation initiative. Among main questions we investigate which logging tools are already available, how they can be re-used for PIM evaluation and which experimental setup from existing evaluation initiatives can be adopted.

- **Measurement and Relevance Assessments.** Finally, a query format and the relevance metrics should be discussed. While there are already a plethora of metrics, do we need more novel measures or can adopt an existing one? We should agree on how should relevance assessments be performed. It would be interesting to formalize the user benefit from the PIM systems usage.

7 Dataset Use Case

One possible example of the dataset usage is the ongoing experiments carried out in L3S Research Center by Paul Chirita, Stefania Costache and Enrico Minack. The experiments address a possibility to increase search effectiveness with the personal information. Here we do not describe precisely the hypotheses they are testing, but rather give an idea how such a dataset can be of use for IR research.

They are trying to compare several ranking algorithms which take into account the information like usage-

based links between various desktop activity contexts. The evaluation is done over 11 users, who were running the logger for last 3-9 months. They take into account events like opening an email or pdf-file for reading, writing on a MS Word, txt or tex-file, or browsing through the Web-sites. They consider that switching between such resources being necessary for the user to achieve her working tasks, so they make use of this valuable context information in search result ranking algorithms. The search queries allow special syntax where user can specify not only plain text queries, but also specify that email, address book contact or mp3 file should be returned as an answer.

During the evaluation step, they processed the log-file and used different heuristics to extract links between resources that reflect relevance between them in a certain context. Each of the tested algorithms produced a large number of links under every tested parameter, so it was not feasible to evaluate the relevance of all the links by the users. Therefore, they identified all links in 9 categories, depending on the type of activity context identified at the Desktop level. Later, they randomly selected 5 links of each category, for each algorithm and algorithm parameter. This leads to 720 links evaluated by the users. The time effort per user was set from 1 to 1,5 hours. They took advantage of the fact that quality of a link between two resources is independent from the algorithm that extracts it, so if a link being evaluated for algorithm A was also extracted by algorithm B, later they can re-use the relevance assessment for algorithm B too.

The preliminary observations show that relevant links could be extracted but search quality is affected quite differently from user to user. In some cases it shows considerable improvements, in some it degrades the performance.

8 Conclusion

The creation of the testbed for experiments with personalized search is more challenging task than creating a Web search or XML retrieval dataset, as it is highly complicated by privacy concerns. This paper describes the ongoing work towards a common dataset based on users' desktop information. Here we presented a possible dataset design and means for collecting the personal information. Also, we outlined the discussion points for the future work and discussion within IR community.

9 Acknowledgment

We would like to thank Paul-Alexandru Chirita and Stefania Costache for their help in creating and using the dataset.

References

- [1] E. Voorhees. The philosophy of information retrieval evaluation. In Proc. of the 2nd Workshop of the Cross-Language Evaluation Forum (CLEF), 2001.
- [2] J. Teevan, S. T. Dumais, and E. 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.
- [3] F. Qiu and J. 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.
- [4] R. Kraft, C. C. Chang, F. Maghoul, and R. 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.
- [5] E. Agichtein, E. Brill, S. Dumais, and R. 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.
- [6] B. Yang and G. 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.
- [7] P.-A. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle++: Semantically enhanced searching and ranking on the desktop. In ESWC, pages 348–362, 2006.
- [8] P. A. Chirita, J. Gaugaz, S. Costache, and W. 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.
- [9] S. Chernov, P. Serdyukov, P.-A. Chirita, G. Demartini, and W. Nejdl. Building a desktop search test-bed. In ECIR '07: Proceedings of the 29th European Conference on Information Retrieval, pages 686–690, 2007.

¹ <http://www.ins.cwi.nl/projects/trec-ent/>

² <http://plg.uwaterloo.ca/~gvcormac/spam/>

³ <http://pas.kbs.uni-hannover.de/>

⁴ <http://www.romip.ru/>