ECIR 2011

Dublin, Ireland
April 18, 2011

EVALUATING PERSONAL SEARCH

Workshop of the 33rd European Conference on Information Retrieval

Organised by
David Elsweiler
Liadh Kelly
Jinyoung Kim

Preface

These proceedings contain the papers presented at the ECIR 2011 Workshop on Evaluating Personal Search, Dublin, Ireland, 18 April, 2011.

Personal Search (PS) refers to the process of searching within one's personal space of digital information, e.g., searching one's desktop or mobile phone for required data items or information. While some recent advancements have been made in this domain, research acceleration is hindered by the lack of established or standardized baselines and evaluation metrics, and lack of commonly available test collections. There is a clear consensus within the research community of the need for standardized repeatable evaluation techniques in the PS space, perhaps in the form of a TREC track for example. However, there are a number of significant challenges associated with this, not the least of which is the fact that the data associated with this domain is personal to the individual, multimedia in nature, and different users will have different forms of collections, differing information needs and different memories of required information. The aim of this workshop is to bring together researchers interested in working towards standardized evaluation approaches for the personal search space. Due to the large space that this covers, as a first step towards overall standardized personal search evaluation the workshop focuses on evaluation for the textual elements within personal desktop collections and known item keyword queries for these elements.

We would like to thank ECIR for hosting the workshop. Thanks also go to the program committee and paper authors, without whom there would be no workshop.

April 2011

David Elsweiler
Liadh Kelly
Jinyoung Kim

Organisation

Program Chairs

David Elsweiler (University of Erlangen, Germany)
Liadh Kelly (Dublin City University, Ireland)
Jinyoung Kim (Umass Amherst, USA)

Program Committee

Leif Azzopardi (University of Glasgow, UK)
Daragh Byrne (University College Dublin, Ireland)
Robert Capra (University of North Carolina, USA)
Yi Chen (Dublin City University, Ireland)
Sergey Chernov (University of Hanover, Germany)
Bruce Croft (University of Massachusetts, USA)
Ronald Fernandez – University of Santiago de Compostela, Spain
Cathal Gurrin – Dublin City University, Ireland
Karl Gyllstrom – Katholieke Universiteit Leuven, Beligium
Donna Harman – NIST, USA
David Hawking – Funnelback, Australia
Sara Javanmardi – Bing, USA
Gareth Jones – Dublin City University, Ireland
Noriko Kando – National Institute of Informatics, Japan
Diane Kelly – University of North Carolina, USA
David Losada – University of Santiago de Compostela, Spain
Ian Ruthven – University of Strathclyde, Glasgow, UK
Alan Smeaton – Dublin City University, Ireland
Jaime Teevan – Microsoft Research, Redmond, USA
Paul Thomas – CSIRO ICT Centre, Australia

Table of Contents

Presentations

Ways we can improve Simulated Personal Search Evaluation

David Elsweiler
Department of Computer Science (8 AI),
University of Erlangen, Germany
david@elsweiler.co.uk

David Losada
Department of Computer Science, University of
Santiago de Compostela, Spain
david.losada@usc.es

1. INTRODUCTION

Analysing how people perform personal search and evaluating the performance of a Personal Search algorithm in a controlled and repeatable way represent an important, but extremely difficult problem for researchers. In Personal Search Evaluation everyone has a unique collection of personal documents, which makes it difficult to compare the performance of one user against another. A second problem is that much of the information within individual collections is private so devising tasks for these collections is also a challenge. Even after overcoming these problems, there is still the issue of repeatability. An individual's relationship with his information changes constantly and the way he interacts is context-dependent. This means that any user study performed is almost impossible to re-perform under the same conditions.

A few methods have been proposed to address these issues. For example, Elsweiler and Ruthven [4] suggested a method of task creation for user-based laboratory re-finding experiments. Chernov and colleagues [2] proposed that researchers volunteer their own personal data to create a shared test collection for research purposes. Kim and Croft [5] use pseudo-desktop collections that have similar properties to personal collections to avoid privacy issues and utilise a simulated querying approach [1] to facilitate automated experiments for known-item tasks.

We believe that this third approach represents the best opportunity to run controlled and repeatable experiments to test retrieval models for Personal Search. That being said, this method, as has been applied to date, suffers from a number of limitations. It is oversimplified and is, consequently, unlikely to replicate user behaviour realistically. In this position statement we outline our views on the weaknesses of the approach and propose ways to improve the process.

2. OVERVIEW OF STATE OF THE ART

The pseudo-collections available in the community include three collections generated from TREC Enterprise track dataset [5], where prominent individuals were identified from the W3C mailing list. Documents were established for these people by taking the emails sent or received by these individuals on the mailing list. These mails were complemented by querying a web search engine with the name, organization and specialization of each target individual to obtain web pages and documents related to that person. A further collection was described in [6], where documents of various types were collected from many public sources in a particular Computer Science department. This collection contains emails from the department mailing list, news articles and blog postings on technology, calendar items of department announcements, web pages and office documents crawled from the department and research group web sites.

Strategies for building simulated queries have been proposed for known-item web page search [1] and for desktop search [5]. Essentially, they are based on randomly selecting a document (known-item) from the collection and algorithmically selecting query terms from the target document. This leads to the automatic generation of simulated queries and relevance judgments.

In the following sections we outline our thoughts on how the various aspects of this process may be improved. More specifically we offer suggestions to improve the query simulation process, the item selection process, and the collections used. We also discuss how we may evaluate the quality of the simulation.

3. IMPROVING QUERY SIMULATION

We posit that the query simulation process used in previous work may not reflect real life. The approaches used to date either randomly select terms from the documents to create queries of an allocated length or they draw terms independently based on how discriminative the terms are (using tf_idf-like weights). We believe this approach is overly simplistic and does not reflect the way queries would be generated in real life. This process does not take into account, for example, that:

- people may be more or less likely to choose query terms from different fields of a document (e.g., the subject or sender field of an email)
- spelling mistakes may be present
- queries may consist of phrases rather than just independent terms
- re-finding queries regularly contain named entities [3]

- queries may contain words not actually present in the document
- queries may be context- or situation-dependent. For example, the characteristics of the user or situation surrounding the task may influence the kind of queries submitted

We argue that to make the simulation process as accurate to real-life behaviour as possible the above aspects need to be accounted for. Our suggestion would be to seed the simulation with real query characteristics extracted from controlled or naturalistic user studies. For example, from a user study evaluating the use of a desktop search tool, e.g. [3], we can learn about how long queries tend to be, the document fields against which they are submitted to, the presence of spelling mistakes, etc. Further, a controlled laboratory-based evaluation, such as performed by [4], would allow researchers to control user and contextual variables to establish query profiles for different situations. This would offer the potential to test the hypothesis that query characteristics change in different scenarios and different algorithms may be offer better support in differing situations as a result.

4. IMPROVING ITEM SELECTION

In current implementations of the query simulation process items in the collection are chosen at random to create known-items. However, previous work has shown that only a small number of personal documents tend to be re-found [7] and that various document properties, such as whether or not it has been re-found before and the time that has lapsed since last access will influence whether or not it will be later re-found.

Further, current approaches treat documents independently, i.e, they do not consider the fact that they may be related and this may influence the likelihood that they will be re-found. If these kinds of properties could be built into the simulation process, we hypothesize that a much more realistic framework for evaluation could be achieved.

We propose to perform longitudinal, naturalistic investigations to establish predictors that documents will be re-used, i.e. document properties that make them more likely to be re-found. This could be achieved by using statistical modelling techniques, such as logistical regression.

5. IMPROVING PSEUDO COLLECTIONS

Due to the inherent difficulties in establishing an appropriate collection for this kind of work, with existing pseudo collections the main criteria has been on establishing any collection that looks like a personal collection, i.e. it is semi-structured and contains information largely created by or associated with one person. While this is a good starting place, we have to investigate whether this is really enough.

The first issue to address is collection size. The existing collections are very small. Second, it is important to ensure that pseudo collections cover a similar breadth of topics as real email collections. Third, the distributions of meta data e.g. senders in email collections should be comparable in real and artificially created collections.

6. EVALUATING THE SIMULATION

Evaluating how the methods suggested above affect the ecological validity of the process is again difficult.

In the literature, query simulations are often evaluated against manual queries (e.g. [1], in the context of known-item web search). Usually, given a pseudo collection, we do not have manual queries and, therefore, this limits the way in which we can assess the quality of the simulated queries. The few attempts done to evaluate the simulations in a pseudo collection environment were based on rather artificial ways to produce hand-written queries from the pseudo collection [5]. Therefore, we strongly argue that a proper method to evaluate simulated queries for pseudo collections is still to be found. Achieving this challenging objective would be a significant advance in this field.

7. CONCLUSIONS

In our view, the pseudo desktop collection approach with simulated queries is the best option to achieve a realistic, controlled and repeatable test environment for Personal Search. In this paper, we have enumerated a number of paths on which simulated evaluation for Personal Search needs to make progress.

8. ACKNOWLEDGMENTS

This work was partially funded by the Alexander von Humboldt Foundation. Further, the second author acknowledges the financial support given by "Ministerio de Ciencia e Innovación" through project grant TIN2010-18552-C03-03

9. REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog, *Building simulated queries for known-item topics: an analysis using six european languages*, Proc. ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 455–462.

[2] S. Chernov, P. Serdyukov, P. Chirita, G. Demartini, and W. Nejdl, *Building a desktop search test-bed*, ECIR'07: Proceedings of the 29th European conference on IR research (Berlin, Heidelberg), Springer-Verlag, 2007, pp. 686–690.

[3] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins, *Stuff i've seen: a system for personal information retrieval and re-use*, Proc. ACM SIGIR '03:, 2003, pp. 72–79.

[4] D. Elsweiler and I. Ruthven, *Towards task-based personal information management evaluations*, Proc. ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 23–30.

[5] J. Kim and W. B. Croft, *Retrieval experiments using pseudo-desktop collections*, CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management (New York, NY, USA), ACM, 2009, pp. 1297–1306.

[6] ———, *Ranking using multiple document types in desktop search*, Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval (New York, NY, USA), SIGIR '10, ACM, 2010, pp. 50–57.

[7] S. K. Tyler and J. Teevan, *Large scale query log analysis of re-finding*, Proc. WSDM '10, 2010.

Pseudo-Desktop Collections and PIM: The Missing Link

Daniel Gonçalves
INESC-ID / IST / TULisbon
Av. Rovisco Pais, 1
1049-001 Lisbon, Portugal
+351 213100248
daniel.goncalves@inesc-id.pt

ABSTRACT

Personal Information Management has for a long time faced a serious problem: validating its results. By dealing with personal information, it is hard to collect performance and quality metrics, and to have a ground case against which possible solutions might be compared. Some efforts have been made to create canonical sets of data that might be used as the basis for such tests. We discuss to what extent are those data sets adequate for PIM, and how they might be improved. We argue that they capture only a limited part of the information in play in real scenarios, and while useful have a restricted applicability. Much meaning is provided by the users themselves, making it hard for information sets not annotated with such meta-data to suffice.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: [Information Search and Retrieval]

General Terms

Design, Standardization, Theory

Keywords

Personal Information Management, Reusable test collections, pseudo-desktop, desktop search, information retrieval

1. INTRODUCTION

The area of Personal Information Management (PIM) is concerned with the study of how people manage their information, from organization to retrieval. There have been many attempted solutions to those problems, and all have ultimately faced the same problems: Adequacy and Evaluation.

Regarding Adequacy, there is a wide range of ways in which individual users manage their information. Thus, while it is possible to test a new system for correctness with a custom-made or well known set of information (often the researcher's own), the doubt remains about whether the solution will work in the general case. Tests done with limited numbers of datasets are anecdotal at best, and there is a risk of over-specialization. Thus, extensive user tests must be performed, leading to the Evaluation problem.

When performing user studies for interactive systems, it is customary to ask users to perform a set of pre-determined tasks. Usability metrics such as task completion times and number of errors are then measured and used as a basis for discussion. In other areas, such as information retrieval, retrieval methods are applied to known (and often pre-classified) datasets. This enables the calculation of measures such as precision and recall. When evaluating PIM solutions, neither is possible. Since we're trying to evaluate solutions that deal with the users' own information, even if we ask the users to perform a same task, *the tasks they end up doing are not the same*. Specific tasks are dangerous. We cannot, for instance, ask all users to find "a document they wrote about dogs". Many won't have such a document. Only more general tasks are possible, such as asking users to find "the last document they wrote and sent to someone", but then the steps that might need to be done might differ wildly from user to user. One might have done so yesterday, the other a couple of months ago, in different settings, for different purposes, etc. Also, it is impossible for researchers to know the users' information and thus know for sure if a particular task has succeeded (was a document not found because of system failure or because it wasn't there to be found in the first place?).

All this makes the validation of PIM systems hard, and begets the creation or definition of a meaningful, representative, set of personal information that can be used as the canonical basis for testing. Such a set would solve the Adequacy problem, and alleviate the Evaluation problem, by allowing researchers to know the data beforehand. While replacing user studies is impossible, such sets might suffice to find meaningful preliminary results and as a way to compare solutions. This was attempted by Kim and Croft [3]. The authors produced three sets of pseudo-desktop data and associated queries. Their goal was to provide information that might be used to evaluate desktop search systems.

2. REPRESENTATIVENESS

The test collections described in [3] contain information divided into five categories: HTML pages, Emails, Word, PDF and Powerpoint files. All with the exception of emails contain around 1,000 items (emails are an order of magnitude more). This is not necessarily representative of a real personal information collection. Previous studies [2] have found a different distribution for the different item types. Another factor that could be taken into consideration is the relatively high variability in personal information collections. Three classes of users were identified, and it would be advantageous if the three pseudo-desktop collections reflected those types. Another important omission are multimedia files. Images and video have a growing importance in the users' lives. The sets should reflect this. Finally, there is organization information missing, folder hierarchy being the most important absence. This information reflects how users organize their information and are important to understand their real needs.

3. SUITABILITY FOR PIM RESEARCH

It is out opinion that, in their present form, the sets might be useful for specific retrieval-related solutions, but are in general unsuited for use with PIM tools. Our objections relate to three related key aspects: Lack of Autobiographic Information, Lack of Meaning, and Lack of Ground Truth.

3.1 Lack of Autobiographic Information

Personal information doesn't exist in a void. It has been previously handled by users, for a reason, in a context. This context goes beyond the computer, into the users' personal and professional lives. It is related to an extended set of autobiographic information, implicitly present in the users' minds. A user might know a document was written around his son's birthday, or during a relative's illness. This information is not part of the documents, but important to users and will determine how they and other personal information are remembered. By automatically generating pseudo-desktop collections, all autobiographic information is missing. It will be possible to use the collections to test techniques for which only the data in the documents is relevant, but not those for which the context is important. Autobiographic information often determines how and why tasks are performed. Furthermore, it is possible to design solutions where it plays a central role, by allowing users to manage their information in *personally relevant ways*.

3.2 Lack of Meaning

Consider an email message. Everyone might look at its sender, `johndoe@somewhere.com` and know that someone at that address is the recipient of the message. From the point of view of the user that sent it, the recipient isn't that email address, but rather (for instance) John Smith, a person with a shared context, other times referred as "Johnny" or even "boss". There is more meaning than can be gleaned from the email message itself (although it might to some extent be inferred from the entire data set). Also, when discussing "the project" in a message to that person the user might know that, in that context, "the project" is actually "Project Foo", on which both work. When designing a retrieval tool, it might make sense for searches for emails to "the boss" to return those to johndoe and "Johnny", etc. Without the users' knowledge, it will be very hard for a system to know they all represent the same person. Having such meaning would also allow us to test how solutions address the well-known Fragmentation Problem [1].

3.3 Lack of Ground Truth

It would be interesting to have data classified according to personal criteria. In traditional retrieval solutions, the set of documents is often manually classified to allow measures such as precision and recall to be computed. This also allows task success to be evaluated. In the context of PIM things are more complex. If a user requests documents about "Subject X", what should be returned? Most likely, not only those that actually contain the words "Subject X", but also others, related to that subject in some way (not to mention multimedia files for which there is no textual information at all). Paraphrases, synonyms, related people and subjects, might all be needed to take into consideration. Again, the user is often the only one that can provide this information, not only complex, but also of a subjective nature. The actual results that would satisfy the user might even change according to the context *at retrieval time*. Having this kind of ground truth would be necessary to evaluate PIM solutions.

4. WILL A SOLUTION EVER EXIST?

The Lack of Autobiographic Information looks at the wider context in which the information is used, and is extrinsic to the data set. The Lack of Meaning reflects the need to have an overall integrated view of all the information. The Lack of Ground Truth is related to how users view their data.

These problems point to the way to create information sets useful and reusable for the evaluation of PIM tools. First and foremost, real information from real users must be collected. An updated study to identify archetypical user classes must be performed, and a different user selected for each class. The collected information must include a wealth of data sources (files, email, calendar, contacts, etc). There are major privacy issues to be addressed. Most can be solved by anonymizing the data, consistently exchanging real names and addresses by simulated ones. A deeper level of anonymization might be necessary, handling project names, places and other sensitive information. This is the simplest part of the creation of the information set.

The users' cooperation would be necessary for the next steps: annotating the information with subjective metadata. The users would need to use a special purpose tool to enter autobiographic information. Also, they would be asked to annotate the documents themselves (and other information), minimizing the Lack of Meaning problem. Finally, they would be asked to classify their documents according to high-level tasks and subjects, addressing the Ground Truth problem (using tags instead of hierarchies, as the same information item might have different uses and meanings). Part of this might be done automatically. For instance, if two email messages are sent to `"John Smith <jsmith@gmail.com>"` and `"John Smith <johns@hotmail.com>"`, the system can make the educated guess that both are the same person. But still this would need to be checked and complemented by the user. The process would be iterative, to fine tune the result. Also, by monitoring the users' everyday use of their information, a set of representative tasks and queries should be collected.

It would be a labor intensive process, but result in information sets that can be understood even in the absence of the user, and used in a rich set of situations where personal information and its surrounding context are relevant.

5. CONCLUSIONS

The creation of pseudo-desktop collections is a worthy goal. Such sets might be very important in providing a testbed for repeatable, comparable experiments, and greatly facilitate the validation of PIM tools. Current versions lack key elements related to the users and the context in which the information is used, which will have to be included for the sets to be of use in a broader context.

6. ACKNOWLEDGMENTS

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and FCT project EDUCARE (PTDC/EIA-EIA/110058/2009).

7. REFERENCES

[1] Bergman, O., Beyth-Marom, R. and Nachmias, R.. The project fragmentation problem in personal information management. In Proc. CHI '06: pp. 271–274. ACM Press, New York, NY, USA, 2006. ISBN 1-59593-372-7.

[2] Gonçalves, D., Jorge, J., An Empirical Study of Personal Document Spaces.. In DSV-IS'03. LNCS v2844. pp. 47-60. Springer-Verlag, 6-9 June 2003, Funchal, Portugal

[3] Kim, J. and Croft W. B. (2009) Retrieval experiments using pseudo-desktop collections. In CIKM'09, pp1297–1306. ACM.

Simulating Memory Recall in Personal Search

Claudia Hauff
Web Information Systems
Delft University of Technology
c.hauff@tudelft.nl

Geert-Jan Houben
Web Information Systems
Delft University of Technology
g.j.p.m.houben@tudelft.nl

ABSTRACT

One of the biggest obstacles to research in desktop search is the lack of publicly available test collections. To alleviate this problem, two pseudo-desktop corpora were created by Kim et al. [4, 5]. An aspect these corpora do not address is the user's imperfect memory recall. The longer an information object that is to be re-found has not been seen, the more likely it is that the user's query is only partially correct. In this poster, we propose two basic models to take such partial memory recall into account.

1. INTRODUCTION

One of the biggest obstacles to research in desktop search is the fact that test collections, as they for instance exist for ad hoc Web search at TREC and other venues, are not publicly available due to privacy concerns. To alleviate this problem, two pseudo-desktop corpora were created by Kim et al. [4, 5]. These corpora contain several types of textual documents as they occur on personal desktops such as emails, files, calendar entries, as well as Web pages. The provided queries represent the most common personal search activity: re-finding of stored information objects. In our study, we focus on the test corpus that contains re-finding queries created by study participants in a human computation game [5] (*CS_collection*).

One aspect of re-finding that is currently not adressed in both corpora is the fact that users' memory recall is not perfect. The longer a sought-after document has not been seen, the more likely it is that the user has forgotten or wrongly remembers some of its attributes [3, 2]. The rate of recall has also been found to be attribute dependent, e.g. in the case of e-mails, the topic of an email has a higher recall rate than the sender [3]. Furthermore, users are more likely to remember the general meaning of a document, instead of specific details [6].

These findings motivate our proposal of studying simulated memory recall to model the user's imperfect recall of information objects. Specifically, we model the recall ability as a function of a document's access frequency as well as the preference for remembering the general content over specifics. We hypothesize, that such a model can change the relative effectiveness of retrieval approaches; an approach that performs well on perfect recall, may be outperformed by another approach on imperfect recall. If this is indeed the case, the imperfection of recall needs to be taken into account when investigating retrieval algorithms for personal search.

	2009 mathworks allan	*email survey for international students graduate employee organization*
Sender	allan	graduate employee organization
Date	2009	
Topic	mathworks	survey for international students
Type		email

Table 1: Manual annotation examples of two queries (top row) from the *CS_collection*.

2. METHOD

Query Annotation: The amount of detail we remember about an item depends on the *access frequency* as well as the attribute to remember: an item (an email, a webpage, etc.) that is regularly consulted or updated is known exceedingly well by the user, whereas an item that has not been looked at in a long time is likely not to trigger a detailed recollection of its attributes. Following the experiments on e-mail recall in [3], we distinguish three types: *hot* (recently accessed items), *warm* (items accessed within the last month) and *cold* (items not accessed within the last month).

In order to assess the impact of different attributes we manually annotated a subset of the queries in the *CS_collection*, namely those queries that aim at re-finding emails, news and calendar items. The following attributes were extracted from each query: (i) sender/author of the item, (ii) topic, (iii) date of sending/creation, and, (iv) the type information available (email, news, calendar). Two example queries of the *CS_collection* are given in Table 1; the date of sending/creating an item can be very general, e.g. *2009*, or very specific, e.g. *17 september*. The same holds for the sender, in some instances only the first name is used, in some instances the full name of the person or organization, such as *graduate employee organization* is given.

Simple Frequency Access Model: The simulation of hot, warm and cold queries proceeds as follows. We assume to know for each frequency access type the probability that the (simulated) user who issues the query remembers an attribute. Then, given an annotated query, for each attribute

	Num.	Topic	Date	Sender	Type
Email	187	178 (95%)	18 (10%)	59 (32%)	34 (18%)
News	94	83 (88%)	–	48 (51%)	8 (9%)
Cal.	49	49 (100%)	18 (37%)	–	9 (18%)
	330				

Table 2: Overview of the annotated queries (a subset of the *CS_collection* query set). News documents have no associated dates and calendar items have no associated sender.

available in the query, with probability p_{attr} we keep it in the final query and with probability $(1 - p_{attr})$ we remove it from the final query.

Partial Frequency Access Model: Removing entire attributes is overly simplistic, it is more likely that not the entire query attribute is forgotten, but only parts of it. For instance, instead of remembering the concrete day the email was received, the month might be remembered. Thus, in the second model, with probability $(1 - p_{attr})$ one or more information nuggets of an attribute are removed. In practice, we remove terms, one at a time. Once the first term of an attribute is removed, further terms are removed with probability p_{rm}. If p_{rm} is low, further removal of information nuggets is unlikely. To model the fact that specifics are more likely to be forgotten, the terms are removed in order of their discriminativeness [1]. In the example in Tab. 1, the topic attribute *survey for international students* will have terms removed in the order of [*survey,international,student,for*].

3. EXPERIMENTS

Query Annotations: We chose the query set of the *CS_collection*, as these are queries created by humans. We annotated the queries aimed at re-finding emails, calendar entries and news items as they are more likely to contain different attributes (such as sender, topic and date) than queries aimed at re-finding webpages or files. This left us with a total of 330 queries, most of them being re-finding emails queries[1]. The details of our annotated query set are shown in Table 2, including the percentage of queries that have a non-empty topic, date, sender (or author) and type attribute. In the case of email queries for instance, 95% of the queries contain a non-empty topic attribute; in cases where the topic attribute is missing, the queries always contain a non-empty sender attribute.

Most email queries contain as only non-empty attribute the topic (92 queries), while topic+sender (37) and topic+type (27) were also common. A similar picture emerges for news queries, queries with only the topic attribute (42) are the most common, followed by topic+author (33) and author only (11). Finally, the calendar queries also contain most often the topic as only non-empty attribute (26), followed by topic+time (14) and topic+type (5).

Of the 330 queries, we found 80 to contain one or more terms that are not contained in the target document; in almost all cases this were either spelling errors (*defencese* instead of *defense*) or non-delimited phrases (*september2009*) and not related to memory recall.

Results: We indexed all fields of the corpus with the Lemur Toolkit[2] with Krovetz stemming applied. Stopwords were not removed. As retrieval approaches we chose language modeling with Dirichlet smoothing ($\mu = 100$) with and without pseudo-relevance feedback (RM1)[3].

		Hot	Warm	Cold
LM	full recall	0.665	0.665	0.665
	simple	0.602	0.610	0.519
	partial $p_{rm} = 0.50$	0.630	0.634	0.576
	partial $p_{rm} = 0.25$	0.631	0.637	0.591
LM-PRF	full recall	0.679	0.679	0.679
	simple	0.605	0.614	0.529
	partial $p_{rm} = 0.50$	0.626	0.629	0.570
	partial $p_{rm} = 0.25$	0.639	0.633	0.591

Table 3: Results in mean average precision for language modeling (LM) and language modeling with pseudo-relevance feedback (LM-PRF).

For the purposes of this experiment, we relied on the recall probabilities reported in [3]. For hot, warm and cold email items respectively, the probability of remembering the time of sending, the sender and the topic is $p_{time} = \{0.71, 0.69, 0.57\}$, $p_{sender} = \{0.96, 0.71, 0.57\}$ and $p_{topic} = \{0.91, 0.93, 0.82\}$. We applied those probabilities to the news and calendar queries as well. The topic attribute is remembered best, the reported numbers show a slightly better remembrance for warm than for hot topics. One reason may be found in the fact that hot items are so new, that the user had no time yet to digest their content. It needs to be stressed, that we use these probabilities for exemplary reasons only.

The results are reported in Table 3, averaged over 10 trials. While in the baseline (full recall, queries remain unaltered), the exploitation of pseudo-relevance feedback improves retrieval effectiveness (LM-PRF performs better than LM), with decreased memory recall the picture is less clear; e.g. for *warm* queries in the partial memory recall model, the LM approach slightly outperforms LM-PRF. Note, that these very small changes in retrieval effectiveness should be viewed with care, further experiments are required to determine if these changes are indeed significant.

4. CONCLUSIONS

In this poster we proposed two basic memory recall models to augment the available pseudo-desktop corpora in light of users' limited recall of information object attributes.

Possibilities for future work include the prediction of the type of document (hot,warm,cold), the user seeks. Based on the type, the best retrieval approach may be selected. Another option is to improve the memory recall model by taking false memories into account. As those wrong memories of an attribute are usually semantically related to the correct memory, determining the semantic relatedness of terms in the corpus and selecting a closely related term to replace the correct one would be a first step.

5. REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07*, pages 455–462, 2007.

[2] T. Blanc-Brude and D. L. Scapin. What do people recall about their documents?: implications for desktop search tools. In *IUI '07*, pages 102–111, 2007.

[3] D. Elsweiler, M. Baillie, and I. Ruthven. Exploring memory in email refinding. *ACM Trans. Inf. Syst.*, 26:21:1–21:36, 2008.

[4] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM '09*, pages 1297–1306, 2009.

[5] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *SIGIR '10*, pages 50–57, 2010.

[6] J. Sachs. Recognition memory for syntactic and semantic aspects of connected discourse. *Percept. Psychophys.*, 2:437–442, 1967.

[1] During the annotation process we removed duplicate queries & erroneous queries (e.g. by users understanding the human computation game wrongly)

[2] http://www.lemurproject.org/

[3] We did a grid search to find the best parameter setting for LM-PRF: $d = 100$ and $t = 100$ feedback documents and terms and a mixing weight of $w_{orig} = 0.9$.

A Strategy for Evaluating Search of “Real” Personal Information Archives

Gareth J. F. Jones
Centre for Digital Video Processing
School of Computing
Dublin City University, Dublin 9, Ireland
gjones@computing.dcu.ie

Yi Chen
Centre for Digital Video Processing
School of Computing
Dublin City University, Dublin 9, Ireland
ychen@computing.dcu.ie

ABSTRACT

Personal information archives (PIAs) can include materials from many sources, e.g. desktop and laptop computers, mobile phones, etc. Evaluation of personal search over these collections is problematic for reasons relating to the personal and private nature of the data and associated information needs and measuring system response effectiveness. Conventional information retrieval (IR) evaluation involving use of Cranfield type test collections to establish retrieval effectiveness and laboratory testing of interactive search behaviour have to be re-thought in this situation. One key issue is that personal data and information needs are very different to search of more public third party datasets used in most existing evaluations. Related to this, understanding the issues of how users interact with a search system for their personal data is important in developing search in this area on a well grounded basis. In this proposal we suggest an alternative IR evaluation strategy which preserves privacy of user data and enables evaluation of both the accuracy of search and exploration of interactive search behaviour. The general strategy is that instead of a common search dataset being distributed to participants, we suggest distributing standard expandable personal data collection, indexing and search tools to non-intrusively collect data from participants conducting search tasks over their own data collections on their own machines, and then performing local evaluation of individual results before central agregation.

1. INTRODUCTION

Personal information archives (PIAs) can include materials from many sources, e.g. content on personal computers and smartphones. The value of such archives can only be realised if they can be searched effectively. Development of suitable search technologies requires that their effectiveness be evaluated. Evaluation in information retrieval (IR) systems ideally includes the measurement of retrieval accuracy and users' satisfaction with the IR system. The former is particularly important for evaluation of IR algorithms, and is generally tested without actual user involvement, while the latter requires input from users. A standard IR evaluation collection includes a document collection, a test set of information needs expressed as search topics, and a set of judgments indicating the relevance of documents to each test topic. Use of this data in an interactive experimental setting enables standardized exploration of interactive IR [5]. However, for personal search, such a dataset is much more difficult to generate due to the heterogeneous nature of personal information space, practical challenges of collecting the data and, significantly, privacy concerns relating to the personal nature of this data. This latter issue creates problems for all aspects of evaluation for search of PIAs.

Current work on evaluation of PIA search is exploring the development of simulated personal Cranfield type search test collections [4]. However, this type of dataset only enables a limited range of research experimentation for PIA search [2]. For example, it cannot be used to explore how a user will query their own PIA or how they will interact with a particular search application. From the search perspective, the key difference between PIA search and standard search environments, is that only the owner of the PIA will be aware of the contents, and thus only they will be able to establish information needs which can be answered by the collection and to determine the relevance of returned content. In order to satisfy this requirement, real users are needed to perform test search tasks, preferably on their own personal data. This requires not only that a user participates in evaluation experiments, but also that they enable the archiving of their personal data and for it to be processed for use in a search system. In this paper we propose a strategy to support evaluation of PIA search based on real user data.

2. LIVING LABORATORY EVALUATION FRAMEWORK FOR PIAS

Our proposed PIA search evaluation methodology is similar to the idea of the *living laboratory* discussed in [3]. This is suggested in the context of evaluating information-seeking support systems which aim to assist users in carrying out open-ended search related tasks. The basic idea of the living laboratory is that rather than individual research groups independently developing experimental search infrastructures and gathering their own groups of test searchers, that an experimental environment is developed which facilitates sharing of resources. This might contain software for data collection, search and evaluation protocols, but also subjects who are available to participate in evaluation tests.

Within a living laboratory for PIA search evaluation researchers wishing to evaluate their technologies would participate in a collaborative evaluation effort. Common indexing and search components would be made available to individuals who agreed to take part in the evaluation exercise. This would then be used to gather PIAs locally and conduct search experiments as outlined in the following sections.

ECIR 2011 Workshop on Evaluating Personal Search April 18 2011, Dublin, Ireland

This proposal builds on the approach in existing work such as the *Stuff I've Seen* study described in [1]. In this investigation a desktop search system with rich user interface functionality was sent to about 230 participants to use as their daily desktop search tool. This tool was used to explore a series of questions on interactive desktop search, and collected a considerable amount of data for further analysis.

2.1 Data Collection and Indexing

Evaluation of personal search requires a document collection and a search system. In terms of developing a personal collection there are two options: to index the data currently on the participant's computer, or to collect data incrementally over a period of time. In either case it will be necessary to install one or more applications on the computer to index the data for search applications.

To do this, open source IR toolkit projects may provide suitable backend technologies for indexing and retrieving for PIA search. However most of these systems are too constrained to traditional IR tasks, e.g. only handling one file form at a time. A PIA search application must be able to index very heterogeneous data sources, and thus existing toolkits may need some extension to support this . In practical terms data may be collected via plugins to existing data management clients before being made available to the indexing application.

2.2 Search System

In order to explore the question of search effectiveness and user search behaviour, we suggest the development of standard search systems which could then be distributed to participants for installation on their own computer. The search client would then search the PIA data index created on their computer. Use of a component based framework for the search system would enable different interface elements and retrieval algorithms to be used in alternative instantiations of the system, which would then enable comparison of their usefulness in search.

2.3 Experimental Tasks

Search topics within a standard IR test collection are typically defined in terms of specific topics known to be covered by the documents in the collection. Since the specific details of an individual's personal collection will not be known and will vary between collections, broader search tasks would need to be defined for our porposed experimental scenario, e.g. referring to meetings with unnamed friends, relatives or colleagues. Even with these more general task statements, the searcher may sometimes find that they are unable to recall any relevant content in their PIA to search for. How to develop suitable task descriptions would obviously have to be clearly defined, and useful lessons in doing this may be gathered from work in designing less specific exploratory search tasks for evaluating information-seeking support systems [3].

2.4 Evaluation

A key part of an IR test collection is the relevance information indicating which items in the collection are relevant to the searcher's information need. Retrieval effectiveness is typically measured using metrics such as precision, recall and various averages where there are multiple relevant items, and average rank and mean reciprocal rank where there is a single relevant known-item. In order to gather relevance data for PIA search, the searcher could be asked to assess the relevance of items retrieved in response to their search in response to each task. If this were undertake at the end of searching for each task, it should not interfere with their search behaviour.

Assessing all items in a collection for relevance is impractical. However, assessing only the items retrieved at high rank using one retrieval method may not give a reasonable indication of the effectiveness with which available relevant documents are being retrieved. To address these issues, *pooling* of results from runs using multiple retrieval methods is often used to construct better approximate relevance sets for standard IR test collections. For the PIA search case, the users topic statement could be applied to multiple retrieval algorithms; only one of these being used by the searcher. The responses of multiple retrieval runs could be pooled and shown to the searcher for assessment.

Interactive search effectiveness can be explored using various measures such as numbers of actions required to locate a required item type, time taken to complete a task, amount of relevant content found, and also potentially subjective feedback of the user's search experiences. Details of user action and responses to questions could be used to explore the cognitive processes undertaken in forming queries and performing a search.

3. CONCLUSIONS

In this paper we have proposed a strategy for PIA search evaluation using a living laboratory approach. The scenario is based on users maintaining their own PIA on their own computer, and using standardized tools to index and search their collection. All relevance assessment and evaluation is also carried out on their computer with only the computed evaluation metrics being returned for aggregation thus preserving privacy of experimental subjects' personal data.

4. ACKNOWLEDGEMENTS

This work was supported by grant CMS023 under the Science Foundation Ireland Research Frontiers Programme 2006. The authors are grateful to the reviewers for their very useful feedback.

5. REFERENCES

[1] S. T. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the ACM SIGIR 2003*, pages 72–79, Toronto, Canada, 2003.

[2] P. Gomes, S. Gama, and D. Gonçalves. Designing a personal information visualization tool. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction*, pages 663–666, Reykjavik, Iceland, 2010.

[3] D. Kelly, S. Dumais, and J. O. Pedersen. Evaluation challenges and directions for information-seeking support systems. *IEEE Computer*, 42(3):60–66, 2009.

[4] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM 2009*, pages 1297–1306, Hong Kong, China, 2009.

[5] P. Over. The TREC interactive track: an annotated bibliography. *Information Processing and Management*, 37(3):369–381, 2001.

Towards “Cranfield” Test Collections for Personal Data Search Evaluation

Liadh Kelly
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
lkelly@computing.dcu.ie

Gareth J.F. Jones
Centre for Digital Video Processing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

Desktop archives are distinct from sources for which shared “Cranfield” information retrieval test collections[1]have been created to date. Differences associated with desktop collections include: they are personal to the archive owner, the owner has personal memories about the items contained within them, and only the collection owner can rate the relevance of items retrieved in response to their query. In this paper we discuss these unique attributes of desktop collections and search, and the resulting challenges associated with creating test collections for desktop search. We also outline a proposed strategy for creating test collections for this space.

Keywords

Test collection creation, pseudo desktop collections.

1. INTRODUCTION

Research progress in development of retrieval techniques for the personal search space is hindered by the lack of common shared test collections. To conduct experiments researchers have largely needed to create their own test collections consisting of individuals data, queries and result sets. There are two problems with this approach: 1) the effort required to create these collections; and 2) the difficultly in gaining large volumes of subjects for such experiments. In other spaces (e.g., web search) standardized collections exisit, hence eliminating these problems. The difficulty for standardization in the personal search space, is the personal nature of collections and individuals resulting unwillingness to share these collections. We foresee two possible avenues for standarization in this space: 1) through a blackbox technique where participating institutes submit their retrieval algorithms for evaluation on the personal collections of other participating institutes using an agreed task formation approach, etc; or 2) through development of pseudo desktop collections, queries and result sets. In this paper we focus on development of pseudo collections for standardized IR evaluation in this domain.

2. TOWARDS TEST COLLECTIONS

To our knowledge the only existing work on pseudo desktop test collection creation is [2]. In this work the authors proposed amassing and created 3 pseudo desktop collections by extracting emails of 3 individuals prominant in the W3C collection and locating web pages, word documents, pdf files and powerpoint presentations related to these people by a web search query consisting of the persons name, organization and area of speciality (provided by TREC expert search track). They randomly chose known items from these collections and used a modification to the approach proposed by [1], for simulated query generation for web page re-finding, to generate simulted queries across multi-field personal items. This approach presents a promising new direction towards larger scale test collections creation for the desktop space and means to examine the utility of desktop retrieval approaches without the need for real users and their collections. However, these collections do not represent the diversity of real users collections, and hence may not provide a reliable way to evaluate the performance of retrieval algorithms intended for personal desktop collections. The created collections contain a limited number of item types and the same volume of each provided item type across the three collections (with the exception of emails). Given the personal nature of desktop collections, we can expect individuals to have different types of collections, with varying volumes and types of content, covering varying volumes of topics. Further the generated pseudo collections do not take account of the items individuals will actually want to retrieve from their collections. In addition, it is not known to what extent the query formulation approach used reflects what collection owners will actually recall about required items and hence the query terms they will use. Indeed the query generation approach of Azzopardi et al [1] which forms the core part of this multi-field query formation approach is acknowledged by its authors to require further analysis and refinement to exhibit more of the characteristics observed by individuals in web page re-finding.

To highlight the differences that can be present across real users collections, consider the personal collections of 3 subjects gathered through logging on their laptop and PCs over a period of 20 months[2], shown in Table 1. These individuals fit a common user profile of being computer science post graduates at the same university. However, as can be seen, even for these similar subjects large differences exist in the volumes of different item types in the collections and in the number of re-accesses of the individuals. Extended

[1] Referred to as ‘test collections’ for remainder of paper.

[2] See http://www.cdvp.dcu.ie/iCLIPS for further details.

ECIR Workshop on Evaluating Personal Search, April 18, 2011, Ireland.
.

over the entire populous, with widely varying interests and requirements, we suspect much different variations would be noted not only in the volumes of information types contained within individuals collections, but also in the diversity and types of topics covered.

Personal collection owners will also have personal experiences and memories associated with the items in their archive, which will guide, depending on their information needs at given moments in time, the items they wish to retrieve from the archive and the query terms they will use in this retrieval process. Individuals will search their personal collections with different personal intentions, memories of required item and personal query generation styles.

We believe in creating pseudo desktop collections that the make up of real users collections need to be replicated as closely as possible to determine how sucessful retrieval approaches will be on real users collections containing varying types and volumes of data, with requirements for different types of item retrieval using different styles of query formation. In the next section we describe a means to gain an understanding of the make up of these collections.

3. STATISTICAL ANALYSIS

To successfully build pseudo desktop collections which represent 'real' users test collections a detailed understanding of the make up of real users desktop collections, items they retrieve from these collections and query formation styles is required. Part of such an analysis could take the form of observations, user studies, diary studies, etc, as are carried out in the PIM community. However, a detailed statistical analysis of the make up of the collections and querying behaviour of a large cross section of the populous is also required in order to move to a situation where real users collections can be replicated in a pseudo way.

To understand the make up of individuals desktop collections, statistics need to be built up on the volume of different information types in these collections, the volume of topics covered, the amount of similarity between items, etc. This analysis could potentially be conducted through a drive within the research community, with either clear guidelines on the statistics to gather or crawlers to automatically generate statistics from participants PCs provided.

We propose that required statistics for target result items would include: extension type of target item, distinctiveness of target item in collection as a whole, recency of last access to target item, etc. And that required statistics for user queries would include: query length, frequency of query terms in target item, frequency of query term in collection as a whole, etc. Similar to gaining statistics on the content of individuals desktops, a stand alone search application or a tool which plugs into individuals current search application (e.g., Google Desktop) could be provided to the research community to log statistics on the nature of queries performed and items retrieved on subjects computers.

Using gathered statistics we propose generating pseudo collections which mimic the characteristics of 'real' collections, described in the next section.

4. TEST COLLECTION CREATION

Using the statistics gathered for each individuals desktop contents, query format and items retrieved, we believe the techniques developed in [1] and [2] provide a strong foundation from which to build pseudo test collections which mimic the characteristics of 'real' test collections.

Type	**Subject 1 Total**	**Subject 2 Total**	**Subject 3 Total**
code files	590 (17)	183 (9)	2,220 (10)
excel	455 (8)	66 (4)	141 (9)
email	3,760 (1)	2,509 (2)	10,243 (2)
pdf	182 (5)	381 (3)	69 (4)
presentations	92 (10)	147 (3)	95 (19)
web	3,895 (3)	15,642 (2)	44,457 (3)
word	311 (7)	310 (6)	373 (13)
text files	381 (6)	81 (2)	308 (6)
other	7 (23)	32 (11)	40 (2)
TOTAL:	9,673 (4)	19,351 (2)	57,946 (3)

Table 1: Total number of distinct items. Average number of accesses to items provided in brackets.

We propose mimicking desktop content by using the statistics gathered on the make up of individuals desktop content to lay user profiles on top of an extension to the pseudo desktop collection creation approach proposed in [2]. In extending this approach, other information which could be mined in creating these collections includes the details provided by people on their homepage, e.g., many people provide lists of personal and work interests and details on co-workers (either explicitly or through inferred means, e.g., co-authorship of papers in the case of academics) on their homepages. We also envisage possibilities to extend the content gathering approach to include other item types and items generated from web content using exisiting summarization, extraction and rephrasing approaches, for example.

Having created pseudo desktop collections we propose extracting target result sets from each user's collection using the available statistics on what the 'real' user retrieves from their collection. To form the queries for the target items, a query generation process which uses the statistics on the 'real' users query formation for the given target item is required. We envisage the query generation approach proposed by [1] and refined to facilitate multi-field retrieval by [2], coupled with the information gained by our proposed statistical analysis would form a good starting point for development of a query generation process for this space.

5. CONCLUDING REMARKS

To drive the test collection creation approach presented in this paper and to both implement and evaluate its component part's against real collections will require formation of a consortium and possibly creation of TREC-like tracks.

The aim of this paper is to present a potential approach to move towards TREC-like collections for research in desktop search and to highlight the requirements of such collections. Thus generating debate and stimulating further progression on possible directions at the workshop.

6. REFERENCES

[1] L. Azzopardi, M. de Rijke, and L. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR'07*, July 2007.

[2] J. Kim and W. B. Croft. Retrieval experiments using pseudo-desktop collections. In *CIKM '09*, 2009.

A Three-stage Evaluation Model for Personal Information Access

Jinyoung Kim and W. Bruce Croft
Department of Computer Science
University of Massachusetts Amherst
{jykim,croft}@cs.umass.edu

ABSTRACT

Evaluation has posed a challenge in studying techniques for personal information access (PIA), and simulated evaluation can be used to address some of the issues. In this work, we list several classes of simulation techniques for PIA evaluation, and propose a three-stage model for PIA evaluation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: [Information Search and Retrieval]

General Terms

Algorithms

Keywords

Personal Information Access, Evaluation Model

1. INTRODUCTION

Personal Information Access (PIA) is a research area focusing on providing solutions to help people find or re-find their own information. The need for better PIA solution is clear as individuals have more and more information in their everyday lives. However, the research for building tools that support such needs has been stagnant mostly due to the challenges in evaluation [1].

Most of previous systems for PIA were often evaluated by an instrumentation-based user study—deploying the system in a real environment and having it evaluated by actual users. Although this kind of evaluation has its own benefits, it requires considerable resources. Moreover, the collections and usage logs from these studies are not open to other researchers because they include private information. We will use the term 'user study' to denote this instrumentation-based evaluation method.

In this position paper, we explain how simulation-based evaluation techniques can address such issues, and propose a three-stage model of PIA evaluation based on the stages of development for a research project.

2. SIMULATED EVALUATION FOR PERSONAL INFORMATION ACCESS

Before discussing simulated evaluation for PIA in detail, we start by defining several components of PIA evaluation. The first component is the *collection* of documents with corresponding metadata. Another component is the *task* of information access, which typically includes known-item finding, topical search, and so on. The last component is the *interaction* between the system and the use. The left side of Figure 1 summarizes these three components.

We then introduce simulated evaluation for PIA. Simulated evaluation in general refers to a class of evaluation techniques where a component of evaluation is replaced with simulated parts, and we can classify simulated evaluations based on the component being substituted for. The right side of Figure 1 summarizes how each of these components can be replaced with simulated components.

Firstly, we can consider replacing the *collection* for evaluation with a simulated collection. Since we need to simulate personal information for PIA evaluation, techniques has been suggested [3] where they collected a set of documents which are related to a person, and topically coherent. Replacing the collection can eliminate most of the privacy concerns, since user's personal information is no longer used.

However, it brings several issues in PIA evaluation. Firstly, the connection between a user and one's collection is lost. Also, some of metadata associated with documents cannot be properly collected (e.g., folder hierarchy). These limitations can be problematic depending on the research question under consideration, and real users' collections should be used in such cases. Section 3.3 will provide an example for this case.

Another way of simulation is replacing the task with an artificial task. For the case of known-item finding, which is known to be the most typical of PIA tasks [1], we only need to choose a target item (document) and assume that the user is trying to find the document. For the case of topical search, one can build a list of hypothetical search topics as a substitute for actual tasks.

Finally, one can simulate the interaction between the system and the user. Since the nature of interaction depends on the task, the detailed technique for simulating the interaction should vary according to the task being simulated. For the case of known-item finding using keyword search, where the primary artifact of interaction is user's search queries,

Evaluating Personal Search Workshop in ECIR'11 Dublin, Ireland
Copyright 2011 ACM 978-1-4503-0099-5/10/10 ...$10.00.

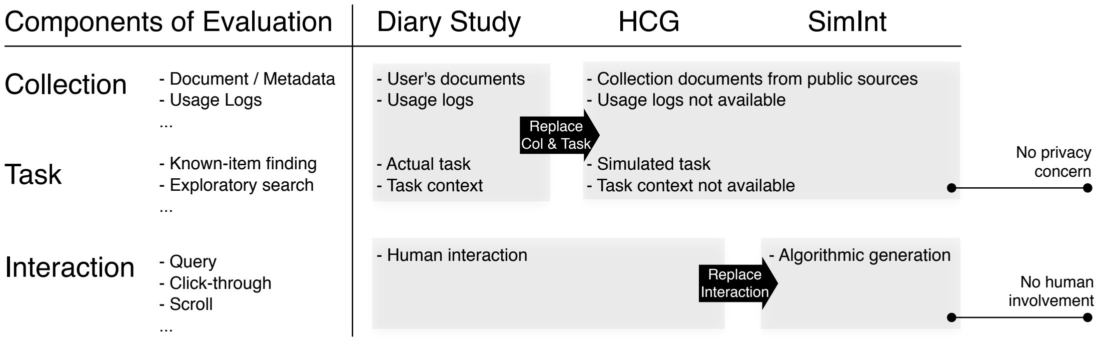

Figure 1: Components of PIA evaluation and the comparison of three stages of evaluation.

this simulation can be done by taking terms from a target document based on some distribution.

3. A THREE-STAGE EVALUATION MODEL FOR PERSONAL INFORMATION ACCESS

Greenberg et al. [2] argues that user study should be employed with a caution, adding that the choice of evaluation methodology must arise from and be appropriate for the actual problem or research question under consideration. We believe that the same principle holds true for PIA evaluation. User study and simulation techniques each has different pros and cons, and the choice or the combination of these techniques should depend on the nature of research questions at hand.

Here we introduce a three-stage evaluation model for personal information access. It is a compilation of existing evaluation methods based on different stages of a research project, where each stage is designed to verify and refine research questions with the different level of development.

3.1 Stage 1 : Simulated Interaction

At an early stage of project, researchers would have only rough hypotheses on the problem, with no facility (e.g., prototype software) to verify their ideas. Simulated interaction is shown to be useful at this stage by allowing them to make a rough estimate on the relative performance of different algorithms [3]. Kim and Croft [3] suggested the pseudo-desktop technique where they used a simulation of collection, task and interaction in the context of known-item finding task.

While simulated interaction in itself may not be sufficient for the final validation of research ideas, it can be the first step by which initial hypotheses are verified and the experimental infrastructures can be prepared.

3.2 Stage 2 : Human Computation Game

With the initial validation of research ideas through Stage 1, researchers can perform a user study using simulated tasks in a simulated collection. This user study can optionally take the form of a game, which motivates participants to complete the task under certain constraints in a competitive setting.

Another paper by Kim and Croft [4] performed a user study based on this idea, where they evaluated several retrieval methods for desktop search based on simulated known-item finding tasks. User studies of this kind are less costly than diary studies since they do not require client-side instrumentation and can be done within a short time. Another benefit is the sharing of the resulting data, since they do not use any private information.

3.3 Stage 3 : User Study

While the evaluation method at Stage 2 can be used to perform evaluations with reasonable human involvement, it is inadequate for some class of research problems because some aspects of the collection and task are hard to be modeled. For instance, evaluating retrieval methods which exploits user's task context would require actual user involvement. In this case, a long-term user study which involves the instrumentation of software to users' system may be required. Our suggestion is not to eliminate this kind of user study completely, but to avoid it when simulation techniques can be an alternative.

4. CONCLUSION

In this paper, we described several classes of simulated evaluation techniques for personal information access, and proposed a three-stage evaluation method. Instead of performing an expensive user study from the beginning of a project, simulated evaluation techniques can be employed to refine initial research ideas gradually. One can start by simulating all components of an actual task, gradually replacing simulated parts with real parts while refining research ideas.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *SIGIR '07*, pages 23–30, New York, NY, USA, 2007. ACM.

[2] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceeding of the SIGCHI conference*, CHI '08, pages 111–120, New York, NY, USA, 2008. ACM.

[3] J. Kim and W. B. Croft. Retreival experiments using pseudo-desktop collections. in Proceedings of CIKM'2009, Hong Kong, China, pages 1297–1306, 2009.

[4] J. Kim and W. B. Croft. Ranking using multiple document types in desktop search. In *In Proceedings of SIGIR '10*, pages 50–57, New York, NY, USA, 2010. ACM.