Interaction Pool: Towards a user-centred test collection

Hideo Joho
Department of Computing Science, University of Glasgow, UK.
hideo@dcs.gla.ac.uk

Robert Villa
Department of Computing Science, University of Glasgow, UK.
villar@dcs.gla.ac.uk

Joemon M. Jose
Department of Computing Science, University of Glasgow, UK.
jj@dcs.gla.ac.uk

ABSTRACT

The advance of evaluation methodology is essential for the development of interactive systems that are based on the understanding of information seeking behaviour. This position paper presents a (rough) design of a community-based approach called the *interaction pool*, a repository of annotated interaction data that can be harnessed and shared by a research community interested in information seeking behaviour, interaction design, interface engineering, and realistic system evaluation. The design of such a repository was motivated by the need to develop a user-centred test collection which inherited the advantages of existing system-centred test collections while considering the characteristics of user-centred research and development.

1. INTRODUCTION

Evaluation of interactive systems and measuring their effects on information seeking behaviour are challenging. The comparison of different interface designs and interactive support systems are even more challenging. In Information Retrieval (IR), common test beds, called *test collections*, have been created and shared by the IR community, being used for extensive testing and comparison of retrieval algorithms over some decades.

While existing test collections have been an important asset for IR research, they are mainly designed for algorithmic evaluation, thus, user interactions and contexts of search are often simplified. Such a test collection is referred to as a *system-centred* test collection in this paper. This position paper is concerned with the design of test collections such that user interactions and search contexts are captured as part of the resource and shared by a community. We will refer to this as a *user-centred* test collection. We believe that such a test collection can facilitate the comparative evaluation of interactive systems and information seeking research while inheriting the advantages of existing approaches.

To maximise the benefits of a user-centred test collection, it is important to obtain feedback from the researchers in the relevant areas. For example, during the design of early test collections, Sparck-Jones and Van Rijsbergen [5] carried out a study to elicit the properties of test collections. While the specification of a test collection is not the main focus of this paper, we hope that this paper will set a tentative ground to discuss the properties of a user-centred test collection.

The rest of the paper is structured as follows. Section 2 summarises how existing test collections work and highlights their advantages and limitations. Section 3 presents a design of *interaction pool* which constitutes a central part of a user-centred test collection. Section 4 illustrates how the interaction pool can potentially facilitate the research on interactive systems and information seeking behaviour. Section 5 discusses several issues that are open for discussion in the context of a user-centred test collection. Finally, Section 6 concludes this paper.

2. SYSTEM-CENTRED TEST COLLECTION

A test collection usually consists of a document corpus, a set of topics, and a list of documents that are relevant to each of the topics (called *qrels*). The document corpus tends to be a static collection so that the performance measures are not violated by content changes. A topic is a description of a searcher's information need. A participant of a system-centred test collection then indexes the document corpus, performs a retrieval using the topic descriptions, and finally, submits the top N[1] ranked documents to the organiser. A *document pool* is then formed by using the top M[2] documents submitted by each of the participated systems (See Figure 1). The assessor of a topic judges the relevance of documents in the document pool, which become qrels.

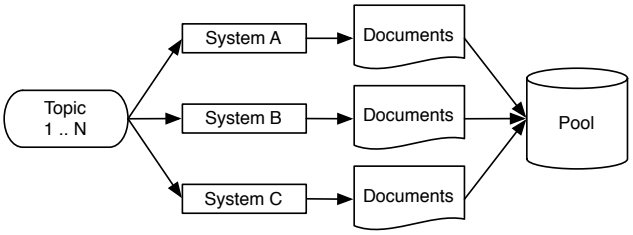

Figure 1: Pooling of documents.

The advantages of this approach is that participating systems use a common set of documents, topics, and qrels,

[1]e.g., 1000 docs
[2]e.g., 100 docs

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.

which makes the comparison among the systems fair and more reliable than tests performed in different conditions. By pooling the documents retrieved by different ranking algorithms, bias towards a particular system is minimised in the evaluation. This also makes it possible to assess a future system using the existing resource. Therefore, the use of a common data set and a pooling method is inherited and assumed in our design of a user-centred test collection.

From the interaction point of view, however, the data stored in a system-centred test collection is the minimum set of interactions where a user submits a query and a system returns a set of (ranked) documents in response to that. The document pool, therefore, stores and evaluates the *outcome* of single search iterations harnessed by participants. However, in a study of interactive systems and information seeking behaviour, the *process* and *context* of search are of great interest. For example, search is often an iterative process which uses multiple queries and browsing of documents. Furthermore, context influences how a search session is developed and how document relevancy is perceived by searchers [3]. A system-centred test collection is not designed to store such data, although effort has been made to elicit some of the contexts inherent in test collections [1].

Another significant property of a system-centred test collection is that document relevancy is determined by a single assessor (who is often a topic creator). This is related to the lack of interaction in the design of system-centred test collections. As discussed above, the relevance of documents can vary over searchers and search contexts. In a user-centred test collection, therefore, the data should contain the document relevancy perceived by different searchers and different contexts. The interaction pool discussed in the next section is designed to address these issues of existing test collections.

3. DESIGN OF INTERACTION POOL

An interaction pool (See Figure 2) is an extension of the document pool where multiple iterations of search are stored. The interaction data such as the queries submitted by users, retrieved documents, click-through documents and their rank positions, next / previous result page viewing actions, are populated along with a timestamp in the interaction pool. Similar to a document pool, the interaction pool contains a range of interaction paths that would be recorded in participated studies which might use different search engines, interfaces, and support systems. This enables researchers to study the process of search harnessed by participants. As such, an interaction pool constitutes a central part of a user-centred test collection.

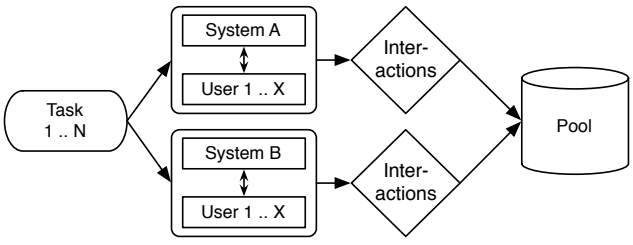

Figure 2: Pooling of interactions.

Participation in a user-centred test collection might occur as follows. First, a set of work/search tasks are defined. Participants carry out a user study using their own choice of systems and the given tasks. The interaction logs are recorded during the study and the data is submitted to the organiser to populate the interaction pool.

Another component considered in the design of an interaction pool is the search metadata. The metadata can consist of a work/search task description (providing a context of search as opposed to a description of what is relevant or not), a user's background, search contexts, system/interface descriptions, subjective assessments, and other information that allows us, for instance, to cluster the interaction data for a granular analysis of people's searching process.

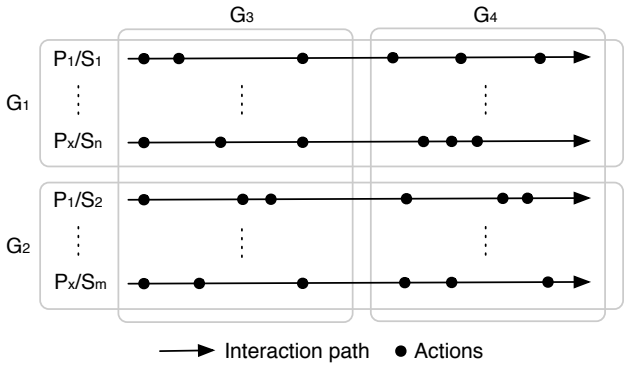

Figure 3: Grouping of interaction data.

Figure 3 illustrates the examples of grouping the interaction data. P denotes a participant ID and S denotes a search session ID. P_1/S_1 to P_x/S_m are a set of search sessions based on a task populated by participants. The horizontal arrows represent an interaction path where the dots indicates user actions occurred in the path. The first case (G_1 and G_2) groups the search sessions into two categories based on a facet or context of search environments. The facet/context can be anything as long as it can be extracted from the metadata (e.g., a user's role in an organisation, level of familiarity/interest with a search topic, search device used). The second case (G_3 and G_4) divides the interaction paths into two different stages of search sessions. In this way, one can analyse the search behaviour at the beginning to middle and middle to the end of search. These are just two examples and other usages of the interaction pool entirely depends on research interests.

Table 1: Aggregate relevance judgements

	Single	Interaction pool			
Doc	assessor	P_1/S_1	P_1/S_2	...	P_x/S_m
D_1	Rel	Rel	Rel	...	Rel
D_2	Non-rel	Non-Rel	Rel	...	Non-Rel
⋮	⋮	⋮	⋮	⋱	⋮
D_n	Rel	Non-Rel	Non-Rel	...	Rel

The interaction pool can also be used to store multiple relevance judgements of retrieved (or click-through) documents. In system-centred test collections, relevance assessments are usually carried out by a single assessor for each topic (See Table 1). In the interaction pool, however, document relevancy is no longer uniform and can vary across

the populated search sessions. Similar to the grouping of interaction data, the aggregated relevance judgements give us an advantage of investigating the effectiveness of interactions and systems from different facets/contexts. Since not all documents are retrieved by every sessions, the annotations such as *shown*, *clicked*, *rel/non-rel* can be associated with retrieved documents.

4. BENEFITS OF INTERACTION POOLS

The previous section described a sketch of an interaction pool that can be harnessed and shared by participants of a user-centred test collection. This section illustrates how such a resource can potentially facilitate the research on information seeking behaviour, user interface/interaction design, and system evaluation.

Information seeking behaviour.

The interaction pool can offers an opportunity to verify existing information seeking models that might have been developed through an ethnographic study. Researchers can access to the pool to analyse whether or not a modelled behaviour can be found in the interaction of various systems, users, and search contexts. While the interaction data in the pool are likely to be based on controlled environment experiments, the search metadata should reduce the level of uncertainty involved in the interpretation of information needs behind the seeking behaviours, compared to, for example, the analysis of search engines' query logs.

In a different scenario, the pool itself might become the rich source of investigation. For instance, one can attempt to mine behavioural patterns from the interaction data. As illustrated above, it will be easy to partition the data based on the annotated metadata, or re-organise the data set to highlight a certain facet/context of search.

Interaction/interface design.

For those who are interested in evaluating the usability or effectiveness of a new search interface, the interaction pool offers realistic user input for benchmarking, whether a user study or simulated study [4] is carried out in the investigation. For example, researchers can extract real queries formulated by the users of the interaction pool and use them as the input to a simulated study of a new interface. Since the users are likely to have a different interpretation of the information need of a given task, their queries are more realistic and diverse than those arbitrary formulated from a task description. The click-through documents in the interaction path can also be exploited as a user feedback trail or path during the task. Overall, the interaction pool can provide extra information for more realistic and controlled simulation of users in the study.

When a user study is conducted independently, the results of the new interface can be compared to the average performance obtained from the interaction pool, or compared to a particular set of search sessions selected by the facets/contexts given in the metadata. For example, one can measure if the new interface allows users to complete a task faster than the average performance in the pool. The subjective assessments can also be compared to other participants' data. When the interaction data in the pool can be used as a baseline performance, then participants can reduce the resources required to carry out a user study (e.g., time and number of subjects). Given that a user study tends to be an expensive process, the interaction pool can reduce evaluation effort. Like a system-centred test collection, we would expect that the experimental resources such as the tasks, document collections, and user's interaction data can be re-used by or support a future interactive IR study.

System evaluation.

The interaction pool offers a new challenge for those who are interested in system evaluation. While a system-centred test collection is designed to determine, for example, if System A is better than System B based on N *topics*, a user-centred test collection is designed to find the difference between the two systems based on N *contexts*. In particular, the notion of uniform document relevancy is no longer compulsory. The interaction pool allows researchers to control how document relevancy is determined by a given facet or context of a search environment.

In the simplest example, the qrels of a task can be generated as many individual search sessions, and the performance of systems can be measured by those individual qrel sets. When a certain facet/context is given, aggregated relevance judgements can be used to measure the system performance for contextual relevance. Since the path of user actions is stored in the pool, one can test the performance of relevance feedback techniques based on a range of interaction patterns.

5. OPEN ISSUES

The requirements and specifications of a user-centred test collection are still under developement. As such, there is a number of open issues. The following are some of the issues that emerged from the preparation of this paper.

Legal/Ethical issue Sharing interaction data imposes an additional element to consider when legal and ethical issues are concerned. This might be as simple as adding a section in a consent form noting that the collected data will be anonymised and shared by the research community. The issue might be more complex for industrial participants. A collective effort needs to be made by the community to share the data since conditions may vary across countries and companies.

Research assets When the interaction data constitutes a fundamental asset in a study, it is conceivable that researchers are not willing to release such data to the community immediately. We need to consider how to achieve a win-win situation for the participants of test collection. Needless to say, participation in an interaction pool means that researchers can access a potentially large quantity and diversity of annotated interaction data which might be infeasible to obtain by a single researcher or research group.

Document collection In existing system-centred test collections, a static document collection (.e.g., web corpus) is often offered to participants. A static collection allows researchers to measure the performance of systems without the effects of content change. On the other hand, participants are responsible for indexing a common document set provided by a test collection. This might be too much effort for, or at least not the

main interests of, some of the participants of a user-centred test collection. An alternative choice is not to have any restrictions on the selection of document collections. Participants can use a search engine's API, for instance, to develop a new interface. In this case, we would need to devise the performance measures that are independent of document collections.

Work/Search tasks It is generally believed that studying people's searching behaviour in the context of tasks (e.g., work task or search task) is beneficial [2], and that a simple description of the search aim by asking users to find as many relevant documents as possible is not realistic. To attract many researchers to participate and contribute to the population of an interaction pool, we need to devise a set of work or search tasks that are realistic and interesting to the research community.

Annotation scheme Participation in a system-centred test collection such as TREC[3] is facilitated by the simple annotation adopted in the data submission. While the data in the interaction pool requires a more complex annotation scheme than a list of document IDs, we should aim to define a standardised scheme which is as simple as possible. A related issue is to formulate a core set of metadata and actions that need to be recorded for the population of an interaction pool.

Infrastructure When the core set of metadata and interaction data is defined, and an annotation scheme is specified by a community, then we would expect to have a repository server which enables participants to access to the pool through some sort of API.

Performance measures In a system-centred test collection, the performance of ranking algorithms is typically measured by precision, recall, and their variants. It is still not clear what performance measure is appropriate for interactive systems and their effects on information seeking behaviour. However, a user-centred test collection has the potential to employ the measures based not only on retrieval effectiveness (e.g., precision/recall) but also on interactions (e.g., number of actions, time to complete a task, etc.) as well as subjective assessments (e.g., "Would you use it if it's available on the web?").

An approach to establish the performance measures for a user-centred test collection can be to analyse the central dependent and independent variables frequently investigated in existing interactive IR studies.

Scale of data There are unknown properties in the current design of interaction pool: how many participants are needed to achieve a meaningful interaction pool; how many tasks should each participant carry out; how many subjects should each participant recruit for populating the pool. While the size and diversity matter in our design, a continuous co-ordination by a community is essential for the development of a successful test collection.

[3] `http://trec.nist.gov`

6. CONCLUSION

This position paper discussed a design of *interaction pool* aiming towards the development of a user-centred test collection. We illustrated how such a resource can support the evaluation of interactive systems. A number of open issues were also discussed. This paper is intended to stir the discussion of evaluation methodology as opposed to presenting a precise specification. We believe this workshop is an ideal forum to discuss such issues.

7. ACKNOWLEDGEMENTS

The authors thank to the anonymous reviewers for their constructive feedback on the paper. This work was supported by EPSRC (Ref: EP/C004108/1), and EU IST FP6 projects (Ref: 033715 (MIAUCE), 027122 (SALERO)). Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] J. Allan. HARD track overview in TREC 2005 high accuracy retrieval from documents. In E. M. Voorhees and L. P. Buckland, editors, *NIST Special Publication: SP 500-266, Proceedings of The Fourteenth Text REtrieval Conference*, 2006.

[2] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56(1):71–90, 2000.

[3] L. Freund, E. Toms, and C. Clarke. Modeling task-genre relationships for IR in the workplace. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 441–448, 2005.

[4] H. Keskustalo, K. Järvelin, and A. Pirkola. The effects of relevance feedback quality and quantity in interactive relevance feedback: A simulation based on user modelling. In *Proceedings of the 28th European Conference on Information Retrieval*, pages 191–204, London, UK, 2006. Springer.

[5] K. Sparck-Jones and C. J. van Rijsbergen. Report on the need for and provision of an 'ideal' information retrieval test collection. Technical Report British Library Research and Development Report 5266, University Computer Laboratory, Cambridge, 1975.