Evaluation Experiments and Experience from the Perspective of Interactive Information Retrieval

Ross Wilkinson Mingfang Wu

CSIRO ICT Center,
Melbourne, Australia
{Ross.Wilkinson, Mingfang.Wu}@csiro.au

Abstract. It has long been a tradition of evaluating information retrieval systems with very simple user models and very simple tasks: the task is to retrieve relevant documents to a user need described by a query. TREC, the Text REtrieval Conference sponsored by NIST, raised the bar by providing large scale collections, well defined user needs, independently judged documents, and a specified form of success. Groups from around the world all tackled this same task that allowed wide analysis of just what factors influenced system performance.

Yet there was concern, as system performance improvement did not always lead to human performance improvement, so a concerted effort to study how people interact with information retrieval systems was undertaken in the Interactive Track of TREC. This paper describes this track, some of the experiments that we have undertaken in this track, and highlights some of the real problems in such evaluation.

There are two key issues that we have often observed in interactive information retrieval. The first issue is that human preference is often not correlated with human performance. Consequently, evaluation that relies solely on either form of evaluation is not reliable. The second issue is that genuine improvements are very difficult to validate, as system variation tends to be dominated by task variation and user performance variation. Consequently, the statistical power of these experiments, often already very expensive to conduct because of user participation, can be quite low. Thus we argue for staged experiments where only very "obvious" system performance gains are explored.

In the end, simple performance measures have proved less helpful than deeper analysis of just how people interact with their information systems.

1 From System-oriented Evaluation to User-oriented Evaluation

Information Retrieval (IR) research has long been driven by evaluation. The evaluation was largely on the system part (or so called batch mode evaluation). The basic testing environment is to build a test collection which includes 1) a collection of documents to search on, 2) a set of queries to search for, 3) and the relevance judgement about each document with respect to each query. A testing system is then fed with a set of queries and scored according to its ability to retrieve relevant documents.

In the early stages (of the field), each research groups built their own small test collection (with hundreds of documents and a few queries). As a result, it was hard to compare systems across research groups and generalize the experimental results. The

TREC [7][14] is a major initiative to address these issues. The goal of TREC is to "encourage research in text retrieval based on large test collection", and to "provide a common task evaluation that allows cross-system comparisons". Compared to other previous test collections (such as Cranfield test collections [6]), the TREC test collection enlarges greatly the size of document collection, provides more realistic queries, and uses more realistic relevance judgments made by independent assessors. In addition, TREC also provides specialized test collections for different test tasks (tracks), such as: filtering, high precision, question answering and so on.

This kind of evaluation is designed to allow one parameter to be manipulated at a time - this is good for comparison of individual components embedded within systems. The evaluation criteria are usually: precision (the proportion of retrieved documents that are relevant) and recall (the proportion of relevant documents that are retrieved), measuring accuracy and coverage. This evaluation technique enables system testing to tell us what (component and algorithm) works and what doesn't. Such experiments are usually repeatable.

This kind of system-oriented evaluation is essential for creating more effective and efficient systems for retrieving and organising information, although it has its drawbacks. For example, it treats the search as a one off session, the relevance judgement is usually binary and held constant and users are abstracted away from the search loop. In reality, relevance is dynamic, situational, cognitive, and context dependent [5], and users are often involved in an extended search process, especially now that accessing information through the web has become a routine part of life.

Actually since the beginning of the field, there have been active discussions and debates on addressing the issues of the evaluations of IR as an interactive system, of the including users in the evaluation and in the use of IR systems. It has been argued that the evaluation of IR systems includes not only the retrieval effectiveness, but also the utility, satisfaction and use [8] [10][11][12][13]. This is increasing recognized as we have seen that the effectiveness of a system may not be automatically transferred to the user performance [4] and it is increasingly hard to achieve further improvements in retrieval accuracy.

Since 1995 (TREC4), the TREC started to set up the interactive track. The goal of this track is "to investigate searching as an *interactive* task by examining the *process* as well as the outcome" [15]. Like the batch model evaluation, this track is also a common platform for research groups to evaluate their systems, and a forum for researchers in this area to discuss how to evaluate an IR system as an interactive process. Unlike the batch model evaluation, the interactive track includes the user. The user is given a scenario that simulates a real world search problem, and is asked to perform the search and made their own judgement on what information is relevant and what is not. The evaluation criteria display a mixed flavour of IR and HCI, it includes the user performance and accessibility, usability and user satisfaction.

Each year, the track defines a task and develops a set of topics. The topic is a brief statement of the task (or information need). A common test collection is stipulated and a common evaluation criteria and procedure is recommended. The criteria on user performance are usually task dependent, but overall, they are measures of the success a subject in completing a specified task. During the experiment, a set of instruments are used to collect subjects' subjective evaluation of the testing systems. The instruments include an entry questionnaire to collect subjects' demographic information and search experience, a pre-search questionnaire for each topic to get subjects' familiarity with the search topic, a post-search questionnaire for each topic to collect subjects' opinion on the search experience and the perceived task completeness, a post-system questionnaire for each system to collect subjects' use experience of that system, and finally, an exit questionnaire to let subjects compare the two systems using a given set of criteria.

The experiment design concerns three factors: system, topic and subject. The experimental design should be able to measure separately the effect of topics, subjects and systems as well as gather some information about the strength of expected interactions between system and topic, topic and subject, as well as subject and system. Thus a factorial design, Latin-square experiment design, is recommended by the TREC interactive track. Table 1 shows an example of such a design.

As we can see, this kind of design is comparative in nature. Indeed, in the past, the participating groups compared various feedback methods (e.g implicit versus. explicit [2]), answer organizations (e.g. ranked list versus clustering versus visualization [1][9][16]), document summaries (e.g. general summary versus. task-biased summary [17]), the transfer of system performance to the user performance [4] and many more (please refer: http://trec.nist.gov/pubs.html).

Table 1. Minimal Latin-square experimental design
(T: test system, C: control system B1 and B2: two blocks of topics)

Subjects	System, Topic	
1	T, B1	C, B2
2	C, B2	T, B1
3	T, B2	C, B1
4	C, B1	T, B2

2 Our Experience

We have been interested in the effect of different delivery methods on the user's performance of a certain task. Through the interactive track platform, we evaluated and compared an organization of clustering search result with a ranked list, and query-biased document summary with generic document summary on the question answering task, and a topic distillation oriented delivery and a general ranked list on the topic distillation task. In this paper, we will just discuss our participation in the question answering task.

The question answering task here is not simply to find a single fact. The task involves collecting multiple independent data items (called instances) from one or more resources and synthesizing them into an answer. Figure 1 shows an example of such a search topic as given to subjects.

The assessment of the user's performance measures the ability of a subject to identify documents that contain topic instances. It is measured by instance recall and instance precision, here instance recall equates to the proportion of the known topic instances contained in the documents identified by a subject, and instance precision the proportion of the documents identified by a subject that were deemed to contain topic instances. The assessment process, therefore, provides indirect or, more accurately, circumstantial evidence of the effectiveness of the interactive system's ability to help the subject develop an answer to the information need represented by the interactive topic.

> **Title:** Ferry Sinkings
> **Description:** Any report of a ferry sinking where 100 or more people lost their lives.
> **Narrative:** To be relevant, a document must identify a ferry that has sunk causing the death of 100 or more humans. It must identify the ferry by name of place where the sinking occurred.
> Detailed of the cause of the sinking would be helpful but are not necessary to be relevant. A reference to a ferry sinking without the number of deaths would not be relevant.
> **Aspects:** Please save at least one RELEVANT document that identifies EACH DIFFERENT ferry sinking of the sort described above. If one document discusses several such sinkings, the you need not save other documents that repeat those aspects, since your goal is to identify different sinkings of the sort described above.

Figure 1. An example topic

Investigation 1: Clustering structure versus ranked list

The topics of the above described task are structured. This, when combined with our goal of structuring and organizing information to form `answers', suggests the hypothesis: that *organizing information with regard to task structure is helpful to users.*

Intuitively, this makes sense. The goal of an interactive subject is to locate documents that pertain to as many different instances of the topic as possible. Given that there is no benefit[1] in locating documents that cover previously discovered topic instances, it would seem desirable to organize the candidate documents in such a way that documents addressing different instances of the query topic were separated into different groups. Ideally, the interactive user could then simply select a single representative document from each instance group. Further, these instance groups could help the user to organize the identified instances into a final answer.

How, therefore, should the candidate information be organized? The approach we chose to explore is clustering. The questions that we attempted to answer include: 1) Can users recognize good clusters? 2) Do users prefer a clustering approach? And 3) Are users more effective with a clustering approach?

We set up two experiments to answer these questions. In the first experiment, we recruited four subjects. Their task was to judge the relevance of a cluster to the topic based only on the description of cluster. As shown in Figure 1, the description of a cluster includes the ten highest-weighted terms from the cluster centroid, five most

[1] In fact, because Interactive Track experiments are conducted within a fixed time limit, it is counter-productive to locate or view documents that only address previously identified instances of a topic.

frequent word phrases from all documents in the cluster, and the title of the top three ranked documents in the cluster. This experiment shows that subjects were able to correctly determines from the cluster description which clusters were likely to contain relevant information, and which were not.

In the second experiment, we aimed to compare the clustering structure (as shown in Figure 3) with a ranked list on the question answering task [16]. This experiment was done in the TREC7 interactive track environment. We recruited sixteen subjects, each of them searched 4 topics on the cluster-based interface, and another 4 topics on the ranked list interface (similar to the design in Table 1). The result showed that overall there is no significant difference in terms of aspectual recall and aspectual precision between the cluster-based interface and the ranked list, although a fairly clear preference for the clustering structure was shown in subjects' comments as captured in questionnaires.

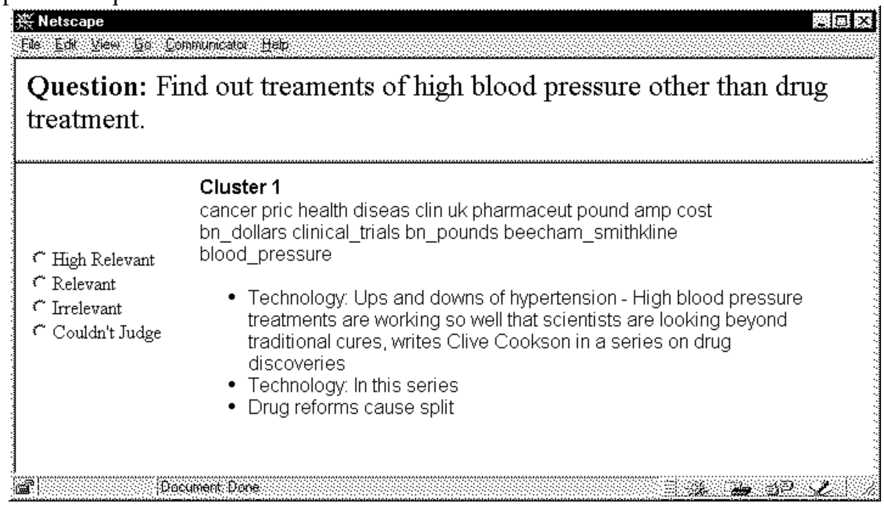

Figure 2. The interface for judging the relevance of clusters

Observations from Investigation 1:

- "Obvious" improvements often do not work. This has often been shown in information retrieval experiments, but extends to work involving users as well. This experiment is an example – there are others we have seen no performance gain.
- There may be a substantial difference between user performance and user preference. We have several times observed that users like a system that provides no performance gain. It is thus important to measure both, to understand the implications of a system variation.

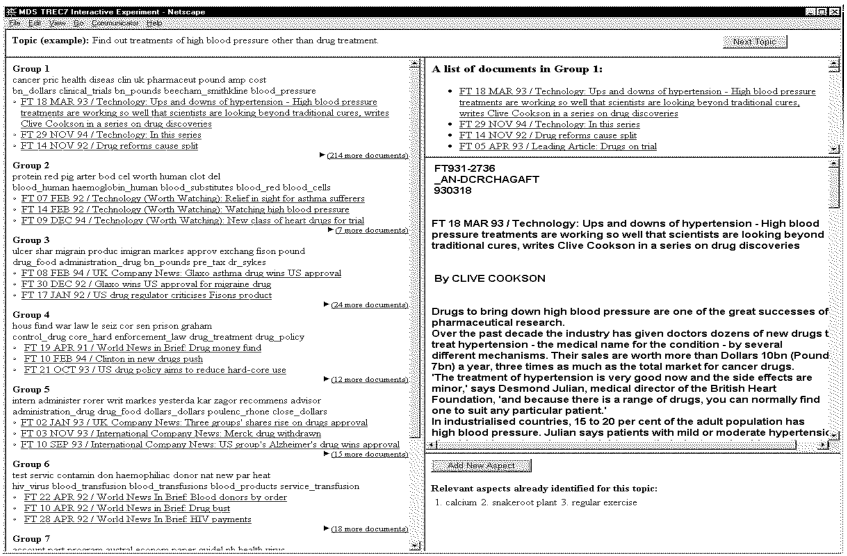

Figure 3. The cluster-based interface

Investigation 2: Task-biased summary versus generic document summary

We analyzed the logged data from the above second experiment, we did find that while the subjects could use the structured delivery format to locate the group of relevant documents, the subjects often either failed to identify a relevant document from the document surrogate or were unable to locate the answer component present in a relevant document. Thus we hypothesized that one of the causes for potential gains from the structured delivery not being realized is that in our test systems the tools used to differentiate the answer containing documents from non-answer containing documents were inadequate for the task of question answering. There is a lack of proper document preview mechanism for subjects to identify the answer-carrying documents.

In the previous experiment, the preview of a retrieved document is represented by its title. The title usually tells the main theme of the document, but very often an answer component (or an instance) exists only within a small chunk of the document, and this small chunk may not necessarily be related to the main theme of the document. For example, for the question "which was the last dynasty of China: Qing or Ming?", the titles of the first two documents are: "Claim Record Sale for Porcelain Ming Vase" and "Chinese Dish for Calligraphy Brushes Brings Record Price". The themes of the two documents are "Ming Vase" and "Chinese Dish" respectively, but there are two sentences in each document that mention the time range for Ming Dynasty and Qing Dynasty. By reading only the title, the subjects may miss a chance to find the answer easily and quickly, even the answer components are located in the top ranked documents.

So in our next experiment, we moved on to provide a user a task-biased document preview/summary. The research question we investigated in the experiment is: given a same list of retrieved documents, will the variation in document summaries improve user's performance on question answering task?

We conducted two experiments that compared two types of candidate lists in two experimental systems [17]. One system uses the document title and the first 20 words of a document as the document's summary, while the other system uses the document title and the best three "answer indicative sentences" extracted from the documents as the document's summary (as shown in Figure 4). The first experiment was done in the TREC9 interactive track environment (the documents in the collection are news articles). A second confirming experiment repeated the first experiment, but with different test collection (with web document collection) and subjects. The purpose of the second experiment was to confirm the strong results from the first experiment and to test whether the methodology could be generalized to web data.

Both experiments showed that subjects took less effort (issued fewer queries and read fewer documents), but found significantly more answers from the interface with task-biased document summary than the interface with the general document summary. Subjects also preferred the task-biased document summary. This result indicates that it makes difference by constructing a delivery interface that takes into account the nature of the task.

The set of experiments indicate that different search tasks may require different delivery methods. For the task of finding relevant documents, it has been found that under some circumstances the clustering of retrieved documents was better than a ranked list [1]. However, for the task of question answering, we didn't find this delivery method performed better than a ranked list. While the second experiment indicates that a relatively simple document summary can significantly improve the searcher's performance in question answering task.

Observations from Investigation 2:

- In both of these experiments, it appeared intuitively "obvious" that the test system (the clustering structure and the task-biased document summary) offered advantages over the base system (the ranked list). Only in the second case did the evidence back up the obvious.
- In this experiment we found a significant difference in performance. The raw average difference in performance was high – there was a 33% performance difference between the test system and the experimental system. This did give a statistically significant difference, but a smaller experiment, or an experiment where the difference was less extreme might have showed a difference that was not statistically significant. Using an ANOVA analysis below we can see that while the effect of the system is significant, so is the effect of the topic, and to a lesser extent, the user (our user population was chosen to be homogeneous.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
System	1	1.205128	1.205128	16.54529	0.0001086
User	15	1.870922	0.124728	1.71240	0.0642876
Topic	7	5.234581	0.747797	10.26656	0.0000000
System:User	15	0.820926	0.054728	0.75137	0.7255804
System:Topic	7	0.518039	0.074006	1.01603	0.4262646
Residuals	82	5.972726	0.072838		

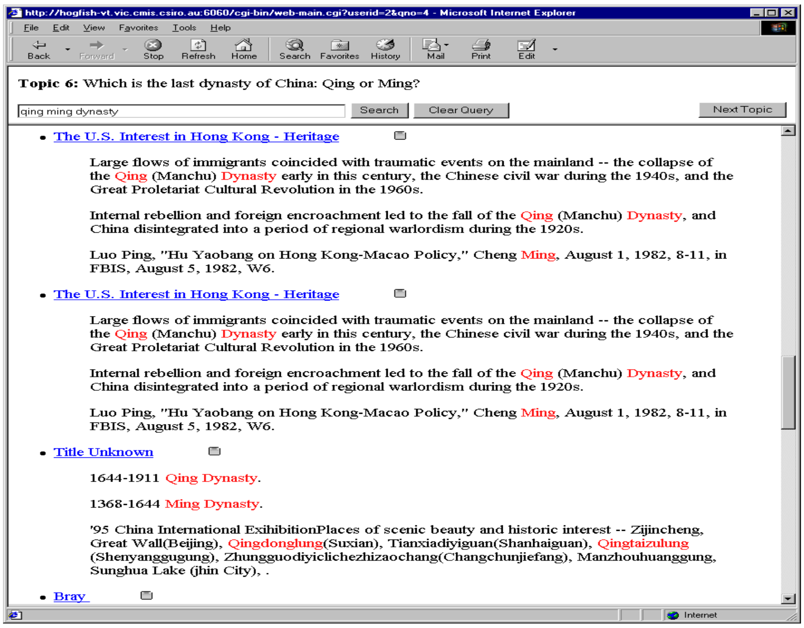

Figure 4. The interface for showing the answer indicative sentences.

3 Lesson Learned

From these experiments, and our observations of other research teams who have contributed to the TREC interactive track, we can make some comments. Firstly from a positive perspective:

- ➢ The TREC leverage the effort to build the evaluation platform – there is significant value to research teams working on the same task using the same data. What is learned by one group can be checked with other research teams' outcomes, so it is possible to identify potential trends, but also possible outliers.
- ➢ As we know the user involved evaluation is costly, each time only a couple of hypothesis can be tested. A common evaluation platform provides an opportunity for two or more research groups to co-ordinate their experiments in a way so that more hypotheses can be tested, as demonstrated in our current study with Rutger's team [18].
- ➢ A carefully constructed study using balanced experimental conditions such as the Latin square design can deal effectively with some of the possible interaction effects – observe the results for the interaction effects in the ANOVA analysis above showing little contribution.
- ➢ Effective delivery works in the right context – knowing the detail of the task can allow information to be organised and delivered more effectively than the standard list format of "normal" information retrieval systems

On the other hand:

- Even we had a common evaluation platform, it is hard to make direct comparisons of different systems across sites – there is often just too much variation, and in any case participating research groups were often exploring different hypotheses.
- It is hard to get a statistically significant difference, as the task variation and user variation often overwhelmed potential system variation leading to a lack of experimental power. It is well worth attempting to determine the variation of performance in a small experiment to determine whether there is any hope of achieving a result of statistical significance.
- Repeatability of experiments is difficult and expensive and may well not produce identical/consistent results, particularly given user variation.

It is important to measure both user performance on the task, and user preference. We have observed that these measures often do not correlate (in a short evaluation period and laboratory controlled setting). Yet both are important, if the target user population likes a system refinement, but it does not lead to productivity gains, it is questionable, but equally if performance improves, but the users do not like the approach, then there may well be a problem in uptake. (Maybe a longitudinal study is needed to observe the correlation between preference and the performance.)

4 Conclusions

We have described something of the history of evaluation in information retrieval from the perspective of the TREC interactive track. We then described some of the experiments that we have conducted, and discussed some of the lessons learned.

There are two key issues that we have often observed in interactive information retrieval. The first issue is that human preference is often not correlated with human performance. Consequently, evaluation that relies solely on either form of evaluation is not reliable. The second issue is that genuine improvements are very difficult to validate, as system variation tends to be dominated by task variation and user performance variation. Consequently, the statistical power of these experiments, often already very expensive to conduct because of user participation, can be quite low. Thus we argue for staged experiments where only very "obvious" system performance gains are explored.

In the end, simple performance measures have proved less helpful than deeper analysis of just how people interact with their information systems.

References

1. J. Allan, A. Leuski, R. Swan and D. Byrd. Evaluating combinations of ranked list and visualizations of inter-document similarity. *Information Processing and Management*, Vol. 37, Number 3, pp. 435-458, 2001.

2. N. J. Belkin, C. Cool, D. Kelly, S. J. Lin, S. Y. Park, J. Perez-Carballo and C. Sikora. Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing and Management*, Vol. 37, Number 3, pp. 403-434, 2001.
3. M. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. *In Proceeding of the 19th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 76-84, 1996.
4. W., Hersh, A. Turpin, S. Price, D. Kraemer, B. Chan, L. Sacherek, D. Olson, Do batch and user evaluations give the same results? In *Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.17-24, 2000.
5. P. Ingwersen. Information Retrieval Interactions. Taylor Graham, 1992.
6. K. Sparck-Jones. The cranfield tests. In K. Sparck-djones (editor), Information Retrieval Experiment, pp. 256-284. Butterworths, 1981
7. K. Sparck-Jones. Further relections on TREC. *Information Processing and management*, vol. 36, pp.37-85, 2000.
8. F. W. Lancaster. Information Retrieval Systems: Characteristics, Testing and Evaluation. New York: John Wiley & Sons, 1979.
9. R. Osdin, L. Ounis and R. W. White. Using hierarchical clustering and summarisation approaches for web retrieval: Glasgow at the TREC 2002 interactive track. In Proceedings of TREC 2002.
10. S. E. Robertson and M. Beaulieu. Research and evaluation in information retrieval. *Journal of Documentation,* Vol. 53, Number 1, pp.51-57, 1997.
11. G. Salton. The smart environment for retrieval system evaluation – advantages and problem areas. In Karen Sparck Jones (editor), *Information Retrieval Experiment*, pp. 316-329. Butterworths, 1981.
12. L. T. Su. The relevance of recall and precision in user evaluation. *JASIS*, vol.45, Number 3, pp.207-217, 1994.
13. J. Tague-Sutcliffe. Measuring the informativeness of a retrieval process. In *Proceddings of the 15th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.23-36, 1992.
14. E. M. Voorhees and D. Harman. Overview of the fifth text retrieval conference (TREC-5). In *Proceedings of the fifth text retrieval conference,* pp.1-28, Gainthburg MD,USA, 1997.
15. P. Over. The TREC interactive track: an annotated bibliography. *Information Processing and Management*, Vol. 37, Number 3, pp.369-381, 2001.
16. M. Wu, M. Fuller and R. Wilkinson. Using clustering and classification approaches in interactive retrieval. *Information Processing and Management*, Vol. 37, Number 3, pp. 459-484, 2001.
17. M. Wu, M. Fuller and R. Wilkinson. Search performance in question answering. In *Proceedings of the 24st ACM SIGIR International Conference on Research and Development in Information Retrieval.* pp.375-381, 2001
18. M. Wu, G. Muresan, A. McLean, M. Tang, R. Wilkinson, Y. Li, H. Lee and N. Belkin. Human versus Machine in the Topic Distillation Task. To appear in *Proceedings of the 27st ACM SIGIR International Conference on Research and Development in Information Retrieval.* July 2004.