A Methodology for Simulated Experiments in Interactive Search

Nikolaos Nanas
Centre for Research and Technology - Thessaly
Greece
n.nanas@cereteth.gr

Udo Kruschwitz, M-Dyaa Albakour, Maria Fasli
University of Essex
Colchester, United Kingdom

Dawei Song, Yunhyong Kim, Ulises Cerviño
Robert Gordon University
Aberdeen, United Kingdom

Anne De Roeck
Open University
Milton Keynes, United Kingdom

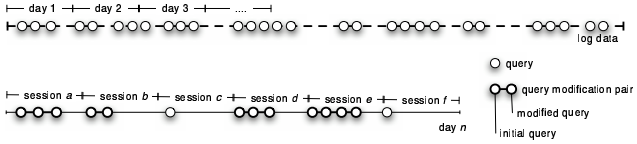
1. INTRODUCTION

Interactive information retrieval has received much attention in recent years, e.g. [7]. Furthermore, increased activity in developing interactive features in search systems used across existing popular Web search engines suggests that interactive systems are being recognised as a promising next step in assisting information search. One of the most challenging problems with interactive systems however remains evaluation.

We describe the general specifications of a methodology for conducting controlled and reproducible experiments in the context of interactive search. It was developed in the AutoAdapt project[1] focusing on search in intranets, but the methodology is more generic than that and can be applied to interactive Web search as well. The goal of this methodology is to evaluate the ability of different algorithms to produce domain models that provide accurate suggestions for query modifications. The AutoAdapt project investigates the application of automatically constructed adaptive domain models for providing suggestions for query modifications to the users of an intranet search engine. This goes beyond static models such as the one employed to guide users who search the Web site of the University of Essex[2] which is based on a domain model that has been built in advance using the documents' markup structure [6].

Over a period of more than two years we have collected a substantial query log corpus (more than 1 million queries) that records all queries and query modifications submitted to the University of Essex search engine. These logs include information about the searching session id, date, the queries and their modifications. Query modifications derive from the user selecting one of the modified queries suggested by the static domain model, from the suggestions proposed by the system which have been extracted from the top-matching snippets, or the user defining a new query in the provided text box. In any case, we are interested in the queries that a user has submitted to the system after the initial query within a session or an information-seeking dialogue (a "search mission").

Figure 1: Experimental Setting



2. SIMULATED QUERY RECOMMENDATION EXPERIMENTS

Here we propose a methodology for performing simulated query recommendation experiments based on log data of the type outlined above. The methodology can be used to perform both "static" and "dynamic" experiments. In particular, we treat the log data as a collection of query modification pairs (initial query – modified query) for building a domain model, but also for evaluating its ability to recommend accurate query modifications. Any log file that records the user queries along with a time stamp and a session id can be used. The log data are traversed in chronological order and in daily batches (see fig. 1). Within each day, subsequent queries submitted within the same searching session are treated as a query modification pair. For instance in the example of figure 1, there are eight query modification pairs within the six sessions of day n.

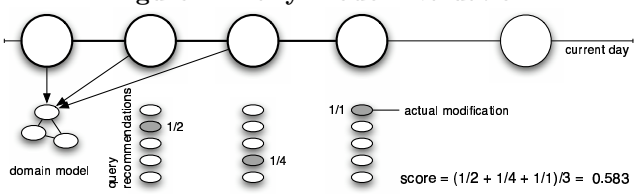
In the static experiments, we start with an existing domain model that remains unchanged during the evaluation process. The model's evaluation is performed on a daily basis as depicted in figure 2. It only takes place for days with at least one query modification pair. For example, let us assume that during the current day, three query modifications have been submitted (fig. 2). For each query modification pair, the domain model is provided with the initial query and returns a ranked list of recommended query modifications. We take the rank of the actual modified query (i.e., the one in the log data) in this list, as an indication of the domain model's accuracy. The assumption here is that an accurate domain model should be able to propose the most appropriate query modification at the top of the list of rec-

[1]http://autoadaptproject.org
[2]http://www.essex.ac.uk

SIGIR Workshop on the Simulation of Interaction, July 23, 2010, Geneva.

ommended modifications. This is based on the observation that users are much more likely to click on the top results of a ranked list than to select something further down [4], and it seems reasonable to assume that such a preference is valid not just for ranked lists of search results but for lists of query modification suggestions as well. The underlying principle of a graded scoring is inherited from DCG [3].

Figure 2: Daily Model Evaluation



So for the total of three query modifications in the current day, we can calculate the model's accuracy score as $(1/r_1 + 1/r_2 + 1/r_3)/3$, where r_1 to r_3 are the ranks of the actual query modifications in the list of modifications recommended by the model in each of the three cases. In the figure's example the models score would be $1/2+1/4+1/1 = 0.583$. More generally, given a day d with Q query modification pairs, the model's accuracy score S_d for that day is given by equation 1 below.

$$S_d = (\sum_{i=1}^{Q} \frac{1}{r_i})/Q \quad (1)$$

Note that in the special case where the actual query modification is not included in the list of recommended modifications then $1/r$ is set to zero. The above evaluation process results in an accuracy score for each logged day for which at least a query modification pair exists. So overall, the process produces a series of scores for each domain model being evaluated. These scores allow the comparison between different domain models. A model M_1 can therefore be considered superior over a model M_2 if a statistically significant improvement can be measured over the given period.

In the case of dynamic experiments, the experimental process is similar. We start with an initially empty domain model, or an existing domain model. Like before, the model is evaluated at the end of each daily batch of query modifications, but unlike the static experiments it uses the daily data for updating its structure. This is essentially a continuous learning problem, where the domain model has to continuously learn from (adapt to) temporal query modification data. Again, we treat a model as superior over another (possibly static one) if an improvement can be observed that is significant.

3. DISCUSSION

The proposed methodology addresses one major weakness of interactive information retrieval, in that it does not involve users and is purely technical. However, the methodology cannot replace user experiments. One reason is that we cannot assume that the selection of a query suggestion will actually be successful in the sense that it leads to the right documents or narrows down the search as expected. A number of other issues remain. For example, we do not try to identify which query modifications within a session are actually related. We consider the entire session in this context. This implies that even subsequent queries that are not related are treated as a query modification pair, thus adding noise to the data. Automatically identifying the boundaries of sessions is a difficult task [2]. One of the reasons is that a session can easily consist of a number of *search goals* and *search missions* [5]. However, we assume that this noise does not affect the evaluation methodology because: a) it is common for all evaluated models and b) no model can predict an arbitrary, unrelated query modification from the initial query. In other words, all evaluated models will perform equally bad for such noisy query modification pairs. But note, that the fairly simplistic fashion of constructing query pairs can easily be replaced by a more sophisticated method without affecting the general methodology proposed in this paper.

Another issue is the question of how the presentation of query suggestions might influence the users' behaviour and how different ways of presenting such query modifications may affect their perceived usefulness.

4. NEXT STEPS

Our plan is to initially use the described methodology to evaluate a number of adaptive algorithms using the log data we have collected. We have already started conducting experiments, following this methodology, for static domain models as well as an adaptive model we have developed and which has been shown to be effective in learning term associations in a user study [1].

5. ACKNOWLEDGEMENTS

AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1.

6. REFERENCES

[1] S. Dignum, U. Kruschwitz, M. Fasli, Y. Kim, D. Song, U. Cervino, and A. De Roeck. Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs. In *Proceedings of WI'10*, Toronto, 2010. Forthcoming.

[2] A. Göker and D. He. Analysing web search logs to determine session boundaries for user-oriented learning. In *Proceedings of AH '00*, pages 319–322. Springer, 2000.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[4] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, pages 154–161, Salvador, Brazil, 2005.

[5] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceeding of CIKM*, pages 699–708, 2008.

[6] U. Kruschwitz. *Intelligent Document Retrieval: Exploiting Markup Structure*, volume 17 of *The Information Retrieval Series*. Springer, 2005.

[7] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 42:43–92, 2008.