Eliciting Information for Adaptive Retrieval

Giridhar Kumaran and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive, Amherst MA 01003 USA

1 Introduction

The task of developing interaction strategies [1] involves determining what additional information will be useful in the context of the query, and the method to obtain this information. In the quest to obtain as much data from the user as possible it is important to keep usability in mind. The most complex interaction mechanisms, however effective, can discourage a user due to high cognitive load. This motivates us to focus on developing a suite of very effective interaction strategies that do not demand much effort, cognitive and physical, from the user. User responses are aimed to be simple too - usually yes/no decisions or selecting from a very small set of options. While this explorative study did not involve an actual user study, each of the techniques described have the potential to be more effective in an interactive setting.

2 Simple Techniques

We designed a few interaction strategies to handle a subset of failures described in a study of why search engines fail [3].

1. *Spelling mismatch due to typographical errors and cultural differences.* To address this problem, we used string edit distance as a simple type of spelling correction, and treated the variants found as synonyms. The user could be asked to verify if the identified variant was truly one.
   Is oestrogen a reasonable variant spelling of estrogen?
2. *Recognizing phrases in the query using punctuation.*Apostrophes, hyphens and double quotes which are usually discarded while indexing indicate the possibility that the associated terms form a phrase. For example, in response to the query *Find documents that discuss issues associated with* **so-called "orphan drugs"**, a user could be asked
   Is it correct that you see so called as a phrase related to the query?
   Is it correct that you see orphan drugs as a phrase related to the query?
3. *Identifying patterns in top-ranked documents* Similar patterns of terms, either as phrase or within certain term windows, occur frequently in similar documents. Questions posed to the user could be of the form
   Would you expect to see leaning and pisa nearby, with terms such as tower and of between them?

3 Interesting Directions and Challenges

Experiments with the three questions described in the previous section with **simulated** interaction[1] on the TREC 2004 and 2005 Robust track data sets have validated their utility[2] . We are currently looking at several additional interaction strategies, mostly motivated by the availability of data annotations from the Automatic Content Extraction program.

1. *Entity context.* It is useful to have a mechanism to further clarify the context a term or entity is used in. For example, users can define context by reporting if the term 'Bonaire' should be part of an address,(*Bonaire, Netherlands Antilles*) or an organization (*Bonaire Democratic Party*).
2. *Person named entities.* The user can be asked to choose the entities related to the query found in the top-ranked results from an initial run. A very short biography from a source like Wikipedia can help the user make the decision.
3. *Top-ranked sentences.* The negative feedback obtained by asking the user to mark non-relevant sentences from the top-ranked ones could be used to clear the results list or reformulate the query.
4. *Targeting named entities.* Specifying the type of named entities the user is interested in can help disambiguate and focus a query.
5. *Query expansion/relaxation.* Providing users with pictorial feedback in the form of an online pie chart showing the percentage of the corpus affected by addition or removal of query terms could potentially guide the user in determining the best set of terms to use in a query.

Each of the above interaction strategies are *light-weight*, but in unison could defeat our goal of minimal interaction. Determining a set of appropriate strategies on a per-query basis is a challenge, with implicit feedback playing a major role. Adapting for different environments - the web, TREC-style querying or templated querying - is a challenge too. In addition to using IR metrics like precision and recall for evaluating result quality, we plan to develop or adapt measures from other areas to measure aspects like cognitive load and usability.

Acknowledgments This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

[1] Kumaran, G., Allan, J.: Simple Questions to Improve PseudoRelevance Feedback Results. ACM SIGIR 2006 Proceedings (2006) 661–662

[2] Harman D., Buckley C.: The NRRC reliable information access (RIA) workshop ACM SIGIR 2004 Proceedings (2004) 528–529

[1] Although our experiments have sidestepped the actual questions, we envision each of the techniques being used interactively

[2] The techniques apply to a small subset of the queries, which were improved on precision