Improving Information Access by Relevance and Topical Feedback

Rianne Kaptein[1] and Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

Abstract. One of the main bottle-necks in providing more effective information access is the poverty of the query end. With an average query length of about two terms, users provide only a highly ambiguous statement of the, often complex, underlying information need. Implicit and explicit feedback can provide us with additional information that can help disambiguate the query and provide more focused search results. We investigate the effects of using different types of feedback. Retrieval results of pseudo-relevance, explicit relevance and topical feedback are compared. Although on average explicit relevance feedback in combination with pseudo relevance feedback works best, for individual queries results are unpredictable. There is a large potential for improvement if we can predict which type of feedback will perform best for a query. Since we are dealing with feedback potentially provided by users standard evaluation measures are not sufficient to evaluate feedback techniques, and the quality of user interaction should also be taken into account.

1 Introduction

Relevance feedback is a commonly used feedback technique to improve search results [5, 6]. Documents that are considered relevant, either because the documents are top-ranked, or because the user marked them as relevant, are exploited in a second iteration of the retrieval process. Another, not so common, feedback technique we discuss in this paper is topical feedback. Instead of using the (presumed) relevant documents, topical feedback uses topic categories considered relevant to the query.

A known drawback of using feedback methods is that while the results on average improve, there are large differences between individual queries. Pseudo-relevance feedback works best for queries where the initial retrieval results are already good. Explicit feedback and topical feedback produce more unpredictable results. While the risks of most feedback techniques can be mitigated by putting more weight on the original query (run) and less on the feedback, this also decreases the potential positive effects of the feedback. If we could predict the effect of using the different feedback techniques on individual queries we can improve retrieval performance and minimize the effort needed from the user. In this paper we analyze the effects of using pseudo-relevance feedback, explicit feedback and topical context by looking at different statistics and at the individual queries.

While explicit relevance feedback does produce the best results when looking at standard evaluation measures, for practical applications it might not be the most attractive option. Explicit relevance feedback can consist of marking one or more documents relevant, or by providing example relevant documents. Users can judge either top-ranked documents or some documents predicted to be the most informative for feedback purposes. Only when the supposed most informative documents are presented for judging to the user, the system is involved.

There are some other forms of feedback that are less static, i.e. the required input from the user depends on the query and the system supports the user by providing intelligent suggestions. For example, Google's spelling suggestions detect possible spelling mistakes; when your query is "relevence", on top of the result list Google asks: "Did you mean *relevance*". Or, when we want to use topical feedback, questions like "Do you want to focus on sports?" or "Are you looking for a person's home page?" can be asked. When these follow-up questions are relevant to the query and easy to answer these kinds of interaction might be more appealing to users than simply marking relevant documents.

Evaluation of feedback approaches is complicated because the interaction with the system is dynamic, and performance depends on the feedback of users. Standard TREC evaluation measures are static and do not have a natural way to incorporate feedback [4]. Instead, feedback documents can be removed from the result ranking, creating a so-called residual ranking, or the feedback documents can be frozen on their position in the initial ranking [1]. Since the standard evaluation measures on their own are not satisfactory, in this paper also we look at some other factors that influence the user's experience.

2 Feedback

2.1 Relevance Feedback

A widely used relevance feedback model was introduced by Lavrenko and Croft [3]. This so-called relevance model provides a formal method to determine the probability $P(w|R)$ of observing a word w in the documents relevant to a particular query. They are using the top-ranked documents retrieved by the query as implicit feedback, but the same model can be used when explicit relevance judgments are available. The method is a massive query expansion technique where the original query is completely replaced with a distribution over the entire vocabulary of the feedback documents. Their results show significant improvements in performance with increases in MAP from 10 to 30% on TREC datasets. However, the gain in performance tends to be foremost the improvement of topics that already did well in the initial run. From the user's point of view, returning worse results on weak queries might not weigh up to the benefit of returning better results on already well performing queries.

Explicit feedback does require an effort from the user. One or more documents have to be marked either relevant or non-relevant. A known relevant document is far from a panacea. The search results after feedback will be biased towards documents similar to the documents marked as relevant. If the relevant document does not cover all aspects of the topic, documents covering other aspects of the topic will be ranked too low. Explicit relevance feedback can also exploit non-relevant documents. While the combination of relevant and non-relevant documents can be beneficial, if only non-relevant documents are available performance will marginally improve at best [7], and the user's effort of judging the documents may be ineffective.

2.2 Topical Feedback

Another method for feedback that we take into consideration is topical feedback. Topical feedback uses topic categories that are considered relevant to the query. Topic categories can be chosen from specialized or general topic directories such as DMOZ, Yahoo! Directory or Wikipedia. The categories can be assigned explicitly by the user, or derived implicitly by applying text categorization techniques to either the query or the top-ranked documents. Another possibility is to show some suggested categories that depend on the query, or to show questions like "Do you want to focus on the twentieth century?" We will use DMOZ as our topic directory, and look at feedback in the form of a DMOZ category that is relevant to the query. We assume that all web sites in the chosen DMOZ category, and all of its direct subcategories are relevant to the query. The feedback model is built from the text on these web sites. From a user study we conducted, we can conclude DMOZ categories are suitable to categorize query topics, and users think it is easy to categorize query topics.

Advantages of topical feedback are that the sites in the DMOZ directory are of high quality and selected by human editors, thus providing us with potentially good feedback documents. A disadvantage of using a topic directory is that not for every query there is an applicable topic category. The DMOZ directory is very general however, and if there is no topic category that applies to the query, there is usually a higher level category under which the query can be placed. Effectively communicating the category to the user is essential, and the topical feedback will by design generate clear intelligible labels (in constrast with, for example, clustering techniques [2]).

Pseudo-relevance feedback can easily be combined with explicit relevance feedback and topical feedback. In our case, after the initial run we first apply explicit relevance or topical feedback, and then apply pseudo-relevance feedback in a third iteration.

3 Experiments

In order to explore the effects and performance of relevance and topical feedback we have conducted experiments on the TREC 2008 Relevance Feedback Track data. First, to compare pseudo-relevance feedback with explicit relevance feedback, we applied both of them and their combination. As explicit relevance feedback, we use the first document of the initial retrieval results that is judged

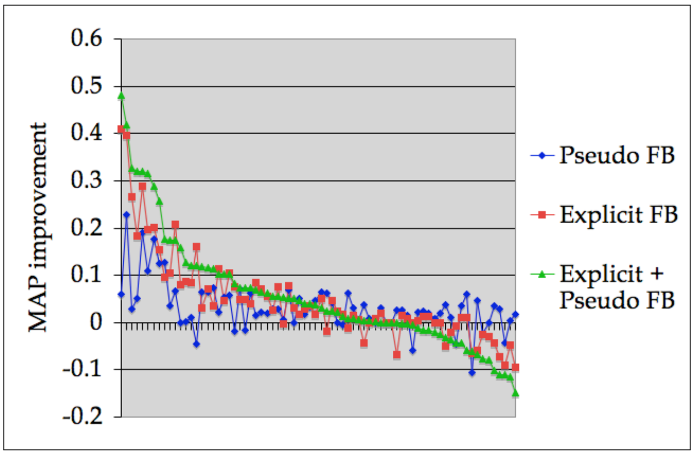

Fig. 1. Absolute difference in Map per query.

as relevant. For evaluation we remove the known relevant document from the results.

Figure 1 gives absolute improvement in MAP per query ordered by MAP improvement of explicit relevance feedback combined with pseudo-relevance feedback. On average, explicit feedback combined with pseudo-relevance feedback performs best with a MAP of 0.3300. Pseudo-relevance feedback achieves a MAP of 0.3044, explicit feedback a MAP of 0.3198. Looking at precision at 10, similar performance improvements can be seen. Comparing relevance feedback to the combination with pseudo-feedback the effects of the combination are larger, on the positive and the negative side.

Another factor in explicit feedback is the user's relevance judgment. The TREC style relevance judgments are on a three-way scale of "not relevant," "relevant," and "highly relevant." However, when looking at our TREC Terabyte topics average MAP improvement of feedback based on highly relevant documents is not higher than MAP improvement of feedback based on relevant documents. So, a document that is judged as highly relevant is not automatically a better document for feedback.

Secondly, we apply topical feedback, obtained by a user study, to 25 of the topics allowing us to explore the effects using the different kinds of feedback. In the user study test persons assign topical categories to query topics by selecting categories from a list and by searching in the DMOZ directory. The list contains categories that are produced by automatic topic categorization using the query, top-ranked documents, and a category title match with the query. Some additional questions about confidence and fit of the category also have to be answered. Each query is categorized by at least two test persons. We select one category that applies best according to the test persons for each query, and use this category as topical feedback.

Table 1. Number of queries for which a feedback method gives the best results.

Model	Baseline		Relevance FB		Topical FB	
Additional blind FB	no	yes	no	yes	no	yes
# Topics with best MAP	1	5	3	8	6	2
# Topics with best P10	4	7	9	12	4	10

We can now compare the results of implicit and explicit relevance feedback and topical feedback. One of the first things to be noticed is that there is a lot of variation in what kind of feedback works. As can be seen in Table 1, each of the retrieval techniques works best for some of the queries. In case multiple retrieval techniques have the same best P10, they are all counted as best. It is hard to predict however which kind of feedback will work best on a particular query. If we would be able to perfectly predict which feedback should be used, MAP would be 0.3917—an improvement of 42.3% over the baseline! This almost doubles the improvement that is achieved with the best single feedback technique.

We do find indicators to predict whether topical feedback technique will improve over the baseline results or not. It turns out the user provided factors “confidence” and the “fit of the category” (based on the user study) do not have a strong correlation to performance improvement. The factors “fraction of query terms in category title” and “fraction of query terms in top ranked terms” do have a strong correlation with performance improvements. When the weight of the feedback is adjusted according to the query terms in the category title or the top-ranked terms, we see an improvement in the results. For pseudo-relevance feedback and explicit feedback there is no such correlation between the fraction of query terms in top ranked terms of the feedback model and the performance improvement. Since the feedback is based on top ranked documents, the query terms always occur frequently in these documents.

There is also a positive side to the fact that the fit of the category does not correlate much to performance improvement. Sometimes categories that are clearly broader than the query, do lead to improvements. The queries “handwriting recognition” and “Hidden Markov Model HMM” both improve considerably when the topical model of category “Computers-Artificial Intelligence-Machine Learning” is applied. So it seems categories on more general levels than the specific queries are useful and one topical model can be beneficial to multiple queries.

4 Conclusion

In this paper we have analyzed the effects of different types of feedback. We found there is not one type of feedback that works best for all or the majority of queries. There are not only large differences between the performance of one type of feedback on different queries, but also between the performance of different types of feedback on the same query. If we would be able to use the full potential of feedback by correctly choosing the most beneficial type of feedback for each query, large performance improvements can be achieved. While user studies are needed to see what types of feedback users prefer, we do not expect users to be able to choose the best type of feedback considering standard evaluation measures, this should be a task for the system.

Acknowledgments Rianne Kaptein was supported by the Netherlands Organization for Scientific Research (NWO) grant # 612.066.513. Jaap Kamps was supported by NWO under grants # 612.066.302, 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

Bibliography

[1] Y. K. Chang, C. Cirillo, and J. Razon. Evaluation of feedback retrieval using modified freezing, residual collection and test and control groups. In G. Salton, editor, *The SMART retrieval system - experiments in automatic document processing*, pages 355–370, 1971.

[2] M. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1996.

[3] V. Lavrenko and W. B. Croft. Relevance-based language models. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2001.

[4] S. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36:95–108, 2000.

[5] J. Rocchio, Jr. Relevance feedback in information retrieval. In G. Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs NJ, 1971.

[6] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297, 1990.

[7] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2008.