Authors:	Järvelin Kalervo
Name of article:	Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments
Name of work:	CIKM'09 Proceedings of the 18th ACM Conference on Information and Knowledge Management
Editors of work:	Cheung D & al.
Year of publication:	2009
ISBN:	978-1-60558-512-3
Publisher:	ACM
Pages:	2053-2056
Discipline:	Natural sciences / Computer and information sciences
Language:	en

URN: http://urn.fi/urn:nbn:uta-3-875
DOI: http://dx.doi.org/10.1145/1645953.1646299

Preprint from: Järvelin, K. (2009). Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM CIKM'09), Hong Kong, Nov. 2-6, 2009, pp. 2053-2056. Full text at ACM DL http://portal.acm.org/dl.cfm

Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments

Kalervo Järvelin
University of Tampere, Finland
kalervo.jarvelin@uta.fi

ABSTRACT

Research on relevance feedback (RFB) in information retrieval (IR) has given mixed results. Success in RFB seems to depend on the searcher's willingness to provide feedback and ability to identify relevant documents or query keys. The paper is based on simulating many user scenarios regarding the amount and quality of RFB. In addition, we experiment with query-biased sentence extraction for query reformulation. The baselines are initial no-feedback queries and queries based on pseudo-relevance feedback. The core question is: under which conditions would RFB based on sentence extraction be successful? The answer depends on user's behavior, implementation of feedback query formulation, and the evaluation methods. A small amount of feedback from a short browsing window seems to improve the final ranking the most. Longer browsing allows more feedback and better queries but also consumes the available relevant documents.

Categories and Subject Descriptors

H.3.1 [**Content analysis and indexing**]: Linguistic processing
H.3.3 [**Information Search and Retrieval**]: Relevance feedback

General Terms

Measurement, Performance, Theory.

Keywords

User simulation, relevance feedback, summarization.

1. INTRODUCTION

It is common knowledge in information retrieval that real users of information retrieval (IR) systems often use simplistic initial queries, which are prone to fail due to vocabulary mismatch, ambiguity and/or lack of discrimination power. Real searchers' first query formulation often acts as an entry to the search system and is followed by browsing and query reformulations [5]. Relevance feedback (RFB) based on initial query results, and query expansion (QE) have been the main approaches to query reformulation. There are several reviews of the techniques, e.g., [1] [7].

In the present paper we focus on interactive RFB. In this method, users either point out relevant documents and the retrieval system infers the expansion keys for the feedback query, or the retrieval system presents a list of candidate expansion keys for the user to choose from. Knowledgeable experienced searchers may benefit more of RFB because they recognize relevant vocabulary and are better able to articulate their needs initially [8]. Users also seem more likely to identify highly relevant documents than marginal ones [12]. There are however two difficulties in providing feedback: capability and willingness [7].

Pseudo-relevance feedback (PRF) [7] avoids the challenges of RFB by assuming that the first documents of an initial search result are relevant without user's interaction. Evaluation results have been somewhat mixed while there is a dominating belief in the IR community in the potential of PRF. Long documents and non-relevant documents however introduce much noise in the PRF process causing query drift. To counteract this, one may use query-biased summaries [4] [10]. Lam-Adesina & Jones [4] employed query-biased document summarization on the initial result and extracted the expansion keys from the summaries. Their results show improvement in retrieval performance (MAP) using document summaries for term selection of up to 15% (to 0.275) compared to the baseline search without feedback (0.24). Further, the use of document summary expansion performed up to 11% better than using standard whole document term selection (from MAP 0.244 to 0.274). Best results using query-biased summaries were better than those for standard summaries, but overall there was little difference between them. Retrieval improvement was also discovered not to be dependent on the relevance of feedback documents. The study was based on using the Top-5 documents (summaries) as feedback and binary relevance.

In the present paper, we take another look at user behavior and IR evaluation. The novel features are based on graded relevance assessments in feedback and evaluation [3], and simulation of user behavior [3]. Binary relevance cannot reflect the possibility that documents may be relevant to different degrees [9]. Highly relevant documents may be more effective in RFB due to the richer relevant vocabulary they provide. While [3] employed both simulation and graded assessments, their work was based on full-document feedback. We shall employ query-biased document summaries. While [4] employed both query-biased and query-independent summaries, their experiment was based on PRF alone and binary relevance in evaluation. Highly relevant documents are effective in RFB [3] and users can readily recognize them in search results [12]. Apparently relevant query-biased summaries are also good indicators of document relevance [11]. Thus users could provide effective feedback and summaries would be effective sources of search keys.

We base our experiments on user simulation (like [3]) rather than tests with real users. This has several advantages, including cost-effectiveness and rapid testing without learning effects. The

Preprint from: Järvelin, K. (2009). Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM CIKM'09), Hong Kong, Nov. 2-6, 2009, pp. 2053-2056. Full text at ACM DL http://portal.acm.org/dl.cfm

informativeness and realism of user simulation can be enhanced by explicitly modeling those aspects of users and RFB that pertain to RFB effectiveness. We shall employ several RFB scenarios (as in [3]) to evaluate the effectiveness of a range of behaviors.

We will study, whether some amount of user effort in providing RFB would be justified based on improved results over initial query results or PRF results, which do not require additional user effort. The user's evaluation dilemma in RFB is that, in addition to the feedback effort, the documents seen in the feedback process need to be frozen to their ranks and only the unseen documents may be re-ranked for presentation. PRF allows re-ranking of the entire collection before presentation of any results to the user. We will show that PRF provides a hard challenge to RFB and that RFB is most promising when the user searches for highly relevant documents only and provides mixed quality RFB early, without excessive browsing.

We utilize the TREC 7-8 corpus with 41 topics for which graded relevance assessments are available [9]. The search engine is Lemur. We will have initial query results and PRF results as baselines to our simulated RFB experiments. To render our simulations empirically relevant, we focus on user's browsing depth shorter than or equal to 20 documents and the number of feedback documents less than or equal to 10. More is unlikely to happen.

2. STUDY DESIGN

2.1 Research Questions

Our overall research question concerns the interaction of RFB and query-biased sentence extraction in query reformulation effectiveness. More specifically:

- What is the effect of the quality of feedback on effectiveness? This is studied in terms of the relevance level of documents given as RFB.
- What is the effect of the quantity of user feedback on effectiveness? This is studied through the number of documents examined and the maximum number given as RFB.
- What are the effects of extraction and QE parameters? Here we study the effects of summary length, total feedback sentence number, the QE key count.

2.2 The Test Collection, Search Engine, Tools

We used the reassessed TREC 7-8 test collection including 41 topics [9]. The document database contains 528155 documents indexed under the retrieval system *Lemur*. The index was constructed by lemmatizing document words. The relevance assessments were done on a four-point scale: (0) irrelevant, (1) marginally relevant, (2) fairly relevant, and (3) highly relevant document. In the recall base there are on average 29 marginally relevant, 20 fairly relevant, and 10 highly relevant documents for each topic.

Expansion sentences and expansion keys were extracted from the feedback documents using the RATF weighting scheme [6]. The scheme computes *relative average term frequency* values for key words as follows:

$$RATF(k) = (cf_k / df_k) * 10^3 / (\ln(df_k + SP))^p$$

cf_k = collection frequency of the key k
df_k = document frequency of the key k
SP = a collection dependent scaling parameter
p = the power parameter

The scheme gives high values for the keys whose average term frequency (i.e., *cf/df*) is high and *df* low. The scaling parameter *SP* is used to down weight rare words. For *SP* and *p* we used the values of $SP = 3000$ and $p = 3$. These values are based on previous work using different topic sets but a similar database.

When scoring sentences, if a non-stop query word did not match any sentence word, an n-gram type of approximate string matching with a threshold was attempted [2].

Initial queries were constructed by applying stopping and lemmatization on long (T+D) topic texts to construct bag-of-word queries. Feedback queries were constructed by appending the feedback keys to the initial query as a second bag-of-words.

2.3 Search Space

The study design consists of the following major variables:

- RFB document relevance threshold (R): 0, 1, 2, 3
- Max number of documents browsed (B): 5, 10, 20
- Max number of RFB documents (F): 1, 3, 5, 10
- Max number of sentences extracted per doc (SD): 2, 4, 6
- Max number of total extracted sentences (ST): 10, 15
- Max number of QE keys (E): 5, 10, 15, 20, 30, 40.

These variables yield some 1700 combinations or cells in the search space, each containing a run of 41 topics and associated evaluation results. Clearly, an exhaustive search would be very laborious. We have aimed to identify in preliminary experiments the best or empirically reasonable ranges for each variable.

2.4 Experimental Protocol

Figure 1 illustrates the overall experimental protocol. TREC topics are first turned to initial Lemur queries and executed, followed by feedback document selection. This is based on the simulated user's relevance requirement (R), amount of feedback (F) and max browsing depth (B). Document relevance data come from the recall base in our simulation. The basis of relevance assessment (seen summary vs. full document) was left open – see [11] for potential effects of this decision.

In feedback query construction, the feedback documents for each query are split into sentences, and the sentences are scored on the basis of the query and the word RATF scores. Word to word matches are facilitated by lemmatization and, in the case of Out-of-Vocabulary words (OOVs), by n-gram string matching. The sentences are ranked and the *SD* best ones are extracted for each document. After processing all feedback documents, the *ST* overall best sentences are identified for expansion key extraction. For each query's set of feedback sentences, their non-query, non-stop words are ranked by their RATF scores and the *E* overall best keys are identified as expansion keys for the query and added to the initial query. The new query is executed and both the original and feedback query results go to evaluation.

Preprint from: Järvelin, K. (2009). Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM CIKM'09), Hong Kong, Nov. 2-6, 2009, pp. 2053-2056. Full text at ACM DL http://portal.acm.org/dl.cfm

2.5 Evaluation and Statistics

We use standard evaluation metrics available in the TREC-eval package and report evaluation results for P@10 documents, and mean average precision MAP. We employ three RFB and evaluation levels, where *liberal* accepts all at least marginal documents as relevant, *fair* accepts all at least fairly relevant as relevant, and *strict* only highly relevant as relevant. Statistical testing is based on Friedman's test between selected RFB runs and their baselines.

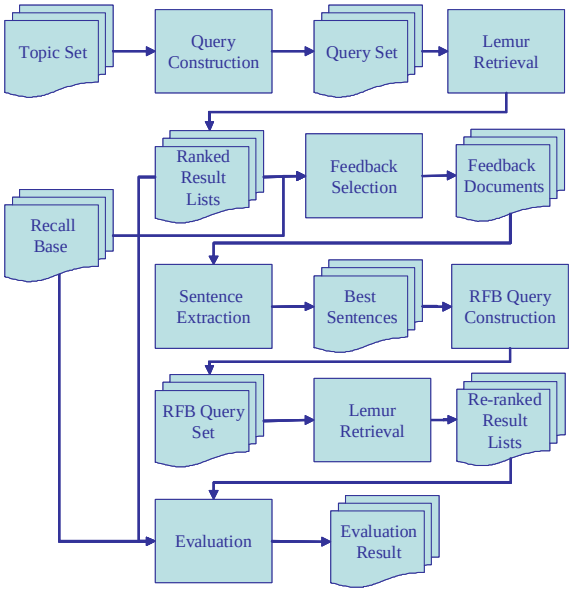

Figure 1. The feedback query construction process.

3. EXPERIMENTAL RESULTS

3.1 Baseline Queries

Table 1 reports the baseline query performances for the T+D queries at the three evaluation levels. For the PRF queries, performance is shown for three feedback document counts (B = 5, 10, 20) and three QE levels (10, 20 and 30 keys). Bold face (for MAP) and underlining (for P@10) are used to indicate the best performance in each browsing lot. Gray background indicates the best overall performance for each query type.

Table 1. Baseline performance for evaluation levels: initial queries, and PRF queries by number of feedback documents and number of expansion keys extracted from summaries

Initial Query Baseline – Title+Description Queries				
	Metric	**Liberal**	**Fair**	**Strict**
	MAP	0.2542	0.2674	0.2653
	P@10	0.4829	0.3659	0.1895
PRF Baseline – Title+Description Queries				
	Metric	**Liberal**	**Fair**	**Strict**
B=5 E=10	MAP	0.2622	0.2567	0.2356
	P@10	0.4927	0.3756	0.1921
B=5 E=20	MAP	0.2769	0.2724	0.2354
	P@10	0.4878	0.3780	0.2026
B=5 E=30	MAP	**0.2861**	**0.2836**	**0.2581**
	P@10	0.4976	0.3976	0.2079
B=10 E=10	MAP	0.2543	0.2411	0.2312
	P@10	0.4659	0.3488	0.1711
B=10 E=20	MAP	0.2733	0.2662	0.2423
	P@10	0.4585	0.3537	0.1842
B=10 E=30	MAP	**0.2805**	**0.2745**	**0.2693**
	P@10	0.4829	0.3707	0.1868
B=20 E=10	MAP	0.2526	0.2392	0.2210
	P@10	0.4659	0.3463	0.1816
B=20 E=20	MAP	0.2730	0.2650	0.2543
	P@10	0.4805	0.3659	0.2053
B=20 E=30	MAP	**0.2767**	**0.2724**	**0.2624**
	P@10	0.4976	0.3732	0.1974

The greatest PRF improvements in MAP for T+D queries are from 3.2 % units (liberal) to 0.4% units (strict). The greatest PRF improvements in P@10 are from 3.2% units (fair) to 1.8% units (strict). Tighter evaluation tends to weaken PRF effectiveness, and shorter browsing with larger QE to improve it.

3.2 T+D Queries with Simulated RFB

Tables 2-3 present the results for RFB expanded T+D queries under liberal and strict feedback. Bolding, underlining and background shading are used as above. Note that we report P@10 even for browsing lengths B ≥ 10 despite freezing, because the effective browsing lengths (and thus effective freezing) can be less than 10, when F is less than 10.

When liberal RFB is used on T+D queries, the best performance is scattered (Table 2). Just giving the first satisfactory document from Top-5 as feedback puts one however within 0.4% units (MAP), or 0.8% units (P@10) from the best performance across all evaluation levels – marginally less considering user effort in browsing beyond Top-5. Comparing to best PRF, user effort improves effectiveness 0.4 - 2.5% units in MAP, and -1.0 to +4.4 % units in P@10, depending on evaluation level and user effort.

Table 2. Effectiveness of simulated RFB runs for evaluation levels by browsing length B, number of feedback documents F and number of expansion keys E extracted from max 10 overall best sentences. Liberal RFB - selected results

	Evaluation:	**Liberal**	**Fair**	**Strict**
B=5 F=1 E=30	MAP	0.2935	**0.3065**	0.2904
	P@10	0.5317	0.4146	0.2132
B=5 F=3 E=30	MAP	**0.2941**	0.2938	**0.2937**
	P@10	0.5293	0.4171	0.2211
B=5 F=5 E=30	MAP	0.2915	0.2907	0.2863
	P@10	0.5098	0.4049	0.2158
B=10 F=1 E=30	MAP	0.2939	**0.3069**	0.2906
	P@10	0.5390	0.4195	0.2184
B=10 F=3 E=30	MAP	**0.2942**	0.2952	**0.2930**
	P@10	0.5220	0.4171	0.2184
B=10 F=5 E=30	MAP	0.2905	0.2901	0.2821
	P@10	0.4805	0.3805	0.1974
B=20 F=1 E=30	MAP	**0.2957**	**0.3088**	0.2923
	P@10	0.5415	0.4195	0.2158
B=20 F=3 E=30	MAP	0.2949	0.2965	**0.2942**
	P@10	0.5244	0.4171	0.2158
B=20 F=5 E=30	MAP	0.2898	0.2891	0.2818
	P@10	0.4829	0.3805	0.1921

When strict RFB is used, the best performance is nearly always obtained by pointing just one highly relevant document in Top-20 (Table 3). However, by identifying the one highly relevant document in Top-5, if it exists, one is within 0.3-0.8 % units in MAP,

Preprint from: Järvelin, K. (2009). Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM CIKM'09), Hong Kong, Nov. 2-6, 2009, pp. 2053-2056. Full text at ACM DL http://portal.acm.org/dl.cfm

or 0.7-2.2 % units in P@10, from the best performance. Comparing to PRF, the best user effort improves effectiveness 0.5 - 3.5 % units in MAP, and -1.0 to +3.2 % units in P@10, depending on the evaluation level, hardly worth the minor effort.

3.3 Comparative Findings

Table 4 compares simulated RFB T+D queries to the two T+D query baselines. As representatives of simulated RFB T+D queries we have chosen those performing best with the least user effort, i.e. little browsing and little feedback. We note that, overall, simulated RFB improves over both baselines. In comparison to initial T+D queries, the RFB improvements have no clear tendency in MAP or P@10 when evaluation becomes tighter. In comparison to PRF queries, RFB queries improve MAP -0.2 to 3.0 % units, and P@10 0.2 to 3.2 % units. When evaluation becomes tighter, the difference in MAP tends to grow while the difference inP@10 tends to diminish.

Table 3. Effectiveness of simulated RFB runs for evaluation levels by browsing length B, number of feedback documents F and number of expansion keys E extracted from max 10 overall best sentences. Strict RFB - selected results

	Evaluation:	Liberal	Fair	Strict
B=5 F=1 E=30	MAP	**0.2845**	**0.3020**	**0.2997**
	P@10	0.5073	0.4000	0.2105
B=5 F=3 E=20	MAP	0.2830	0.2923	0.2957
	P@10	0.4976	0.3927	0.2105
B=5 F=5 E=30	MAP	0.2841	0.2916	0.2918
	P@10	0.4951	0.3976	0.2158
B=10 F=1 E=30	MAP	0.2870	**0.3023**	**0.3015**
	P@10	0.5073	0.4024	0.2211
B=10 F=3 E=20	MAP	**0.2873**	0.2915	0.2909
	P@10	0.4878	0.3878	0.2158
B=10 F=5 E=30	MAP	0.2858	0.2868	0.2838
	P@10	0.4634	0.3634	0.1947
B=20 F=1 E=30	MAP	**0.2919**	**0.3049**	**0.3044**
	P@10	0.5293	0.4073	0.2184
B=20 F=3 E=30	MAP	0.2896	0.2915	0.2913
	P@10	0.5122	0.3927	0.2184
B=20 F=5 E=30	MAP	0.2862	0.2864	0.2816
	P@10	0.4854	0.3683	0.1921

Table 4. T+D queries - simulated RFB difference to initial queries and PRF queries in MAP and P@10 at three evaluation levels. For each evaluation level, the 1st data column is the difference to initial queries and the 2nd to PRF queries

Simulated RF – Title+Description Queries – Liberal RFB				
	Evaluation:	Liberal	Fair	Strict
B=5 F=3 E=30	MAP	4.0 0.8	2.6 1.0	2.8 2.4
	P@10	4.6 3.2	5.1 2.0	3.2 1.3
Simulated RF – Title+Description Queries – Strict RFB				
	Evaluation:	**Liberal**	**Fair**	**Strict**
B=5 F=1 E=30	MAP	3.0 -0.2	3.5 1.8	3.4 3.0
	P@10	2.4 1.0	3.4 0.2	2.1 0.3

Interactive RFB performance is uninteresting unless clearly better than PRF performance. Taking at least 2 % units' difference as the criterion, Table 4 shows in grey background the four interesting cases. When MAP is the metric, the greatest benefits of RFB turn out when the evaluation criterion is strict. By Friedman's test, strict RFB is significantly better than either of the baselines ($p<0.5\%$), and liberal RFB is almost so ($p=6.2\%$). Regarding P@10, RFB appears most beneficial when evaluation is liberal but not significantly so (Friedman's test, $p>10\%$).

4. CONCLUSION

In conclusion, RFB systematically improved performance over both baselines. RFB is most effective when just one high-quality feedback document, or a few of mixed quality are indicated in the very top ranks of the initial result, and evaluation is by strict relevance. MAP shows favorable results for RFB if evaluation is by strict relevance. P@10 is improved when RFB and evaluation are liberal (i.e. more marginal documents are allowed in the top ranks). However, RFB requires user feedback effort, and long queries for best performance, whereas PRF is fully automatic and not far behind in average performance. RFB may remain little used in practice unless it becomes clearly easier to employ and more effective compared to its alternatives.

5. ACKNOWLEDGEMENTS

This research was supported by the Academy of Finland grants #120996 and #124131. E. Airio, H. Keskustalo and T. Talvensaari contributed to test system realization.

6. REFERENCES

[1] Efthimiadis, E. N. 1996. Query expansion. In Annual Review of Information Science and Technology, vol. 31. Information Today, Medford, NJ, 121-187.

[2] Järvelin, A. & Järvelin, A. & Järvelin, K. 2007. s-grams: Defining Generalized n-grams for Information Retrieval. Inform. Process. Manag. 43, 4, 1005-1019.

[3] Keskustalo, H., Järvelin, K., and Pirkola, A. 2008. Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value. Inform. Retrieval. 11, 5, 209-228.

[4] Lam-Adesina, A. M. and Jones, G. J. F. 2001. Applying Summarization Techniques for Term Selection in Relevance Feedback. In Proc. of the 24th Annual ACM Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 1-9.

[5] Marchionini, G., Dwiggens, S., Katz, A., and Lin, X. 1993. Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. Libr. Inform. Sci. Res. 15, 1, 35–70.

[6] Pirkola, A., Leppänen, E., and Järvelin, K. 2002. The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. Inform. Res. 7, 2. Online: http://informationr.net/ir/7-2/infres72.html

[7] Ruthven, I. and Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. Knowl Eng Rev 18, 2, 95-145.

[8] Sihvonen, A. and Vakkari, P. 2004. Subject knowledge improves interactive query expansion assisted by a thesaurus. J. Doc. 60, 6, 673-690.

[9] Sormunen, E. 2002. Liberal Relevance Criteria of TREC - Counting on Negligible Documents? In Proc. of the 25th Annual International ACM SIGIR Conference on Research and

Preprint from: Järvelin, K. (2009). Interactive Relevance Feedback with Graded Relevance and Sentence Extraction: Simulated User Experiments. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (ACM CIKM'09), Hong Kong, Nov. 2-6, 2009, pp. 2053-2056. Full text at ACM DL http://portal.acm.org/dl.cfm

Development in Information Retrieval. ACM Press, New York, NY, 320-330.

[10] Tombros, A., and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In Proc. of the 21st Annual ACM Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 2-10.

[11] Turpin, A. & al. 2009. Including Summaries in System Evaluation. In Proc. of the 32nd Annual ACM Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 508-515.

[12] Vakkari, P. and Sormunen, E. 2004. The influence of relevance levels on the effectiveness of interactive IR. J. Am. Soc. Inf. Sci. Tech. 55, 11, 963-969.