UNIVERSITY OF TAMPERE

This document has been downloaded from
Tampub – The Institutional Repository of University of Tampere

Authors:	Keskustalo Heikki, Järvelin Kalervo, Pirkola Ari
Name of article:	The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback
Name of work:	Advances in Information Retrieval : 28th European Conference on IR Research, ECIR 2006 London
Editors of work:	Lalmas Mounia et al.
Year of publication:	2006
ISBN:	3-540-33347-9
Publisher:	Springer
Pages:	191-204
Series name and number:	Lecture Notes in Computer Science 3936/2006
ISSN:	0302-9743
Discipline:	Natural sciences / Computer and information sciences
Language:	en
School/Other Unit:	School of Information Sciences

URL: http://www.springerlink.com/content/834rw5g327010314/fulltext.pdf
URN: http://urn.fi/urn:nbn:uta-3-869
DOI: http://dx.doi.org/10.1007/11735106_18

Additional infromation:
The original publication is available at www.springerlink.com.

All material supplied via TamPub is protected by copyright and other intellectual property rights, and duplication or sale of all part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorized user.

The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling

Heikki Keskustalo, Kalervo Järvelin, and Ari Pirkola

Department of Information Studies,
FIN-33014 University of Tampere, Finland

Abstract. Experiments on the effectiveness of relevance feedback with real users are time-consuming and expensive. This makes simulation for rapid testing desirable. We define a user model, which helps to quantify some interaction decisions involved in simulated relevance feedback. First, the relevance criterion defines the relevance threshold of the user to accept documents as relevant to his/her needs. Second, the browsing effort refers to the patience of the user to browse through the initial list of retrieved documents in order to give feedback. Third, the feedback effort refers to the effort and ability of the user to collect feedback documents. We use the model to construct several simulated relevance feedback scenarios in a laboratory setting. Using TREC data providing graded relevance assessments, we study the effect of the quality and quantity of the feedback documents on the effectiveness of the relevance feedback and compare this to the pseudo-relevance feedback. Our results indicate that one can compensate large amounts of relevant but low quality feedback by small amounts of highly relevant feedback.

1. Introduction

Selection of good search keys is crucial for successful text retrieval, yet users of information systems often find it difficult to find the best expressions for their information needs [3, 4]. On the other hand, although users may have difficulties in expressing exactly their information needs, they are often able to identify useful information when they see it. This fact leads to the notion of relevance feedback (RF). In RF, the users mark documents as relevant to their needs and present this information to the system. This information can be used for automatically modifying better queries [8, 9].

Actually, users of information systems might best be served by systems that retrieve especially highly relevant documents [6, 15]. The results of a user study [14] indicate that the users are also able to identify highly relevant documents. Moreover, the textual characteristics of the documents at various relevance levels differ: [13] showed that in highly relevant documents a larger number of aspects of the request was discussed, and a larger set of unique expressions was used. These observations lead to our research questions: How effective is RF when we consider various levels of relevance in evaluation phase? How is the quality and quantity of the feedback documents related to the effectiveness? From the point in view of creating RF interfaces, we should learn what kind of evidence we should try to collect from the searchers. In this paper, we shall explore these questions through user simulation in a laboratory setting. We use a test collection, a subset of TREC collection providing graded relevance assessments of documents for 41 topics [12]. The graded assessments are scaled from 0 (non-relevant) to 3 (highly relevant). We shall simulate the quantity of RF by the number of documents in the initial result marked as feedback by the user, and the quality by the relevance threshold set by the user. As an additional research question we ask whether the simulated relevance feedback may successfully compete with pseudo-relevance feedback (PRF), and if so, by what effort in terms of the amount and quality of the user feedback? We evaluate all cases using non-interpolated average precision (MAP) at three different relevance thresholds.

Our laboratory simulation provides a rapid means of exploring the limits of user feedback without laborious experiments with real users. For example, one may find out, as we will also report, what kind of user RF effort is most effective and how it compares with the PRF. One needs to verify these findings in real world situations. However, this may be done more efficiently when one has better insight into what to test.

The rest of the paper is organized as follows. In Section 2 we explain our experimental methodology – user modeling for simulations, the test collection, the retrieval system and the test runs. Section 3 presents our findings, Section 4 discusses the main result and Section 5 presents the conclusions.

2. Methods

2.1 User modeling for relevance feedback simulation

Pseudo relevance feedback is a highly parameterized process. For example, the number of documents used in the feedback, the methods for selecting and weighting the feedback keys, and the number of the feedback keys extracted may be varied.

Human relevance feedback has similar characteristics when one considers, as we do, user feedback based on document level judgments. The number of top documents the user is willing to examine varies. The user has also many methods for selecting and weighting the feedback keys. The number of feedback documents that the user actually selects may vary. In addition, importantly, the user may tolerate irrelevance, require relevance, or ignore marginal relevance to different degrees in the feedback documents. This is a characteristic of human feedback that escapes automatic methods of PRF. It might also provide a qualitatively better basis for RF, which leads to outperforming automatic PRF if the user is willing to provide the effort.

Since users vary greatly, we developed a simple user model to grab the parameters above. We use three concepts for modeling:

- requirement of document relevance (stringent, regular or liberal): relevance threshold *R*
- willingness to browse (patient/impatient): window size *B*
- willingness to provide feedback (eager/reserved): feedback set size *F*

The *requirement of document relevance, R,* is an important dimension since many users may want to focus on highly relevant documents only [6, 15]. Users can also identify them while marginal documents easily escape the user's attention [14]. We model the relevance threshold dimension by possible values of graded relevance $R \in \{0,1,2,3\}$. In other words, $R = 3$ indicates that the user is capable and willing to recognize and accept only highly relevant documents for RF, whereas $R = 1$ indicates that the user liberally accepts even marginal documents for RF. As a special case, $R = 0$ models the case where all the documents considered are accepted, i.e., PRF (blind feedback).

The *willingness to browse, B,* models the user's capability and willingness to browse through the ranked retrieval result. The user's willingness to study retrieved sets is limited (futility point) [1]. We model the browsing dimension by the number of documents considered (window size *B*). For example, $B = 1$ indicates that the user is impatient and only willing to consider the first document and gives up after that, whereas $B = 30$ indicates a patient user willing to examine a long list of retrieval results. In the present study we shall only consider a limited set of values for *B*, i.e., $B \in \{1, 5, 10, 30\}$.

The *willingness to provide feedback, F* ($\leq B$) models the user's willingness to mark documents as relevant. We separate this dimension from the previous one since even if the user is willing to browse through a long list, she may give up after finding the first or first few relevant documents. In this paper we examine only positive RF. This dimension is essential since, as [2] argues, users may be reluctant to provide feedback, and on the other hand the amount of feedback may be critical to success. We model the willingness to provide feedback by the maximum number of documents the user is willing to mark as relevant *F*. As an example, $F = 1$ indicates that the user is reserved and only willing to consider the first relevant document encountered as feedback and gives up marking after that, whereas $F \geq 10$ indicates an eager user willing to provide lots of feedback. In the present study we shall only consider a limited set of values for *F*, i.e. $F \in \{1, 5, 10, 30\}$ while $F \leq B$.

User model is a triplet $M = <R, B, F>$ which defines a three-dimensional space of user characteristics. Each triplet with specified values is a point in the space modeling a distinct type of user (*a user scenario*) or RF interaction. It is obvious that some regions of the space are more interesting than others. However, in general, relations between the more distant areas are of interest, e.g., can one compensate low quality feedback by giving it in large amounts. Moreover, how much user's effort, and what kind, is needed to outperform pseudo-relevance feedback, i.e., which scenarios $< R, B, F >$ ($R > 0$) provide better effectiveness than $<0, B', F' >$? (In PRF R=0 and $B' = F'$.)

2.2 The test collection

In this study the reassessed TREC documents from [12] are used including altogether 41 topics from TREC 7 and TREC 8 ad hoc tracks. The non-binary relevance judgments were obtained by re-judging documents judged relevant by NIST assessors together with about 5% of irrelevant documents for each topic. The selection of topics was based on the size of recall bases, i.e., each topic should have more than 30 relevant documents but the size of the pool to be reassessed should not exceed 200 documents (for details, see [5, 12]). The relevance judgment in the reassessment process was based on topicality. The new assessments were done on a four-point scale:

- (0) Irrelevant document - the document does not contain any information about the topic.
- (1) Marginally relevant document - the document only points to the topic and does not contain more or other information than the topic description.
- (2) Fairly relevant document - the document contains more information than the topic description but the presentation is not exhaustive. In case of multi-faceted topic, only some of the sub-themes or viewpoints are covered.
- (3) Highly relevant document - the document discusses the themes of the topic exhaustively. In case of a multi-faceted topic, all or most sub-themes or viewpoints are covered.

Altogether 6122 documents were reassessed (Table 1). Almost all of the originally irrelevant documents were also assessed irrelevant in reassessment (93.9%). Of the TREC relevant documents about 76% were judged relevant at some level and 24% irrelevant. This seems to indicate that the re-assessors have been somewhat stricter than the original judges. Among the relevant documents one half were marginal, a third fairly relevant, and a sixth highly relevant [5].

Table 1. The distribution of relevance assessments in the test collection (41 topics)

Relevance Level	Total Number of Documents	%	% of Relevant	Avg number per Topic
Rel = 0	3719	62.1	..	..
Rel = 1	1197	18.6	49.8	29.2
Rel = 2	812	12.8	33.8	19.8
Rel = 3	394	6.6	16.4	9.6
Total	6122	100.0	100.0	58.6

In the recall base there were on the average 29 documents of relevance level 1 per each topic, 20 documents at relevance level 2, and 10 documents at relevance level 3 per topic. In other words, on the there were on the average 59 relevant documents of some relevance level per each topic (Table 1).

The document collection contained 528155 documents organized under the retrieval system *InQuery* (see below). The database index is constructed by lemmatizing the document words (using ENGTWOL morphological analyzer by Lingsoft, Inc.).

2.3 The retrieval system InQuery and the feedback key extraction

The *InQuery* system was chosen for the test, because it has a flexible query language and it has shown good performance in several tests (see, e.g., [4]). *InQuery* is based on Bayesian inference networks. All keys are attached with a *belief value*, which is approximated by the following *tf.idf* modification:

$$0.4+0.6*\left(\frac{tf_{ij}}{tf_{ij}+0.5+1.5*\left(\frac{dl_j}{adl}\right)}\right)*\left(\frac{\log\left(\frac{N+0.5}{df_i}\right)}{\log(N+1.0)}\right) \qquad (1)$$

where tf_{ij} = the frequency of the key i in the document j
dl_j = the length of document j (as the number of keys)
adl = average document length in the collection
N = collection size (as the number of documents)
df_i = number of documents containing key i.

The *InQuery* query language provides a large set of operators to specify relations between search keys. In the present paper we only need the typical probabilistic operator *#sum* and the synonym operator *#syn*. The probabilistic interpretations for these operators are given below:

$$P_{sum}(Q_1, Q_2, ..., Q_n) = (p_1 + p_2 + ... + p_n)/n \quad (2)$$

where P denotes probability, Q_i is either a key or an InQuery expression, and p_i, i = 1...n, is the belief value of Q_i.

The probability for operands connected by SYN operator is calculated by modifying the tf.idf function as follows:

$$0.4 + 0.6 * \left(\frac{\sum_{i \in S} tf_{ij}}{\sum_{i \in S} tf_{ij} + 0.5 + 1.5 * \left(\frac{dl_j}{adl} \right)} \right) * \left(\frac{\log\left(\frac{N + 0.5}{df_s} \right)}{\log(N + 1.0)} \right) \quad (3)$$

where tf_{ij} = the frequency of the key i in the document j
S = the set of search keys within the SYN operator
dl_j = the length of document j (as the number of keys)
adl = average document length in the collection
N = collection size (as the number of documents)
df_S = number of documents containing at least on key of the set S.

Our initial queries are based on the TREC topic wording, excluding the stop list words, and have the structure *#sum(#syn(key$_1$, key$_2$, …), #syn(…, key$_n$) „…)*. The synonym structures are due to lemmatizing topic words. Some of them are ambiguous, and for a given word all its interpreted lemmas are included in one synonym set. The extracted expansion keys form a sum structure *#sum(key$_{e1}$, key$_{e2}$, …)* and the revised feedback query has the structure *#sum(#sum(#syn(key$_1$, key$_2$, …), #syn(…, key$_n$) „…) #sum(key$_{e1}$, key$_{e2}$, …))*.

Expansion keys were extracted from the feedback documents using the RATF weighting scheme [7]. The scheme computes *relative average term frequency* values for the keys of documents, as follows:

$$RATF(k) = (cf_k / df_k) * 10^3 / \ln(df_k + SP)^p \quad (4)$$

cf_k = the collection frequency of the key k
df_k = the 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 a previous study using different topic sets but a corresponding database [7].

In the expansion key extraction, from each feedback document a word list containing the 50 best keys was extracted by the ranked order of their descending RATF values. When more than one document was given as feedback, the RATF key lists for each document were united followed by the extraction of 30 best keys (keys shared by the greatest number of word lists).

2.4 Experimental Set-up

The overall experimental set-up consists of the following steps:

1. For each TREC topic (N=41) the title and description fields are processed and automatically formulated into the initial query.
2. Each initial query is run in the test collection and the initial result set (the top 50 documents for each topic) is retrieved.
3. By using the user scenario <*R, B, F*> together with the recall base, the set of feedback documents (defined uniquely by each user scenario) is extracted automatically from the initial result set.
4. The expansion keys are extracted from the set of feedback documents (among the relevant documents from the initial run). RATF weighting scheme is used here. The 30 best expansion keys are extracted and formed into a *#sum* –clause. This clause is combined with the initial query to form the feedback query.
5. Each feedback query is run in the test collection and the final result (document set) is retrieved.
6. Both the initial result and the final results are analyzed for their mean average precision (MAP), applying three different evaluation criteria: stringent, regular and liberal. The same evaluation criteria were also used for conceptualizing the feedback requirements as highly relevant (*Rel* =3), at least fairly relevant (*Rel* ≥ 2), or at least marginally relevant (*Rel* ≥ 1) documents, respectively.

The initial queries were formed automatically by excluding stop words, lemmatizing the content-bearing words and applying fuzzy matching from the database index in case of words which could not be lemmatized (two best matches were selected). Lemmatization leads to synonym sets of one or more components. These are combined by the *#sum* –operator into the initial query.

The evaluation of retrieval effects of RF methods when *real users* are involved has some special requirements. Admitting that users may be lazy to browse, an evaluation measure based on DCV (document cut-off values) or discounted cumulated gain [6] might be preferred over MAP. If only a small evaluation window is used, one may argue that only the unseen documents should be shown at the RF phase and the evaluation should not reward re-ranking of the feedback documents among the final document set [10]. This could be achieved by keeping the documents identified as relevant (within the browsing and feedback scope) "frozen" to their initial ranks. On the other hand, because there are no intermediate results seen by the user in the PRF, it is not possible to make an entirely fair comparison between the simulated user RF and the PRF case. In PRF, MAP or precision at 10 % recall are typical effectiveness measures.

In the present paper, we compare the user RF and PRF and want to find out how various user feedback scenarios are related to search effectiveness. We measure this by using MAP at various relevance thresholds. An inherent problem with using MAP this way is that the RF documents may be re-retrieved by the feedback query (although not necessarily) but with a better ranking, especially when the recall bases are small. On the other hand, we may think of a user situation where the user is collecting relevant documents at the end of the process and has not yet really read initial feedback documents. In this situation, re-retrieving the relevant documents is not problematic, as the user is simply interested in the quality of the *final* search result. In later studies, we shall apply DCV based measures and take specifically into account the role and effect of the feedback documents among the final result set considering the user view point differently.

3. Findings

3.1. Effect of User Scenario on the Amount of Feedback

The first obvious question is what is the relationship of user's relevance criteria and effort to the quantity and quality of relevance feedback available? In order to answer to that question, we first study the effect of selecting a specific user scenario to the number of feedback documents available (the third column in Tables 2-4), to the cases of no feedback (fourth column), to the maximum number of feedback documents available (fifth column), and to the actual window size used before the browsing window limit is reached (the last column). Table 2 presents the stringent user case, that is, the user accepts only highly relevant documents as feedback documents ($R = 3$) while we vary the values of the two effort thresholds (B and F).

Table 2. Stringent user (R = 3): the effect of user effort on the availability of RF. All feedback documents are highly relevant.

Browsing Effort B	Feedback Effort F	Average No of RF Docs Available	No of Topics with no RF Docs	Max No of RF Docs per Topic	Average Search Length
30	30	2.3	11	11	30.0
30	10	2.2	11	10	29.6
30	5	1.9	11	5	27.4
30	1	0.7	11	1	15.0
10	10	1.2	21	8	10.0
10	5	1.1	21	5	9.9
10	1	0.5	21	1	6.7
5	5	0.8	23	4	5.0
5	1	0.4	23	1	3.9
1	1	0.3	29	1	1.0

We can see that on the average, in case of stringent feedback threshold, the number of RF documents is very low even if both the browsing effort and the feedback effort thresholds are high (30). With a relatively small effort, e.g., B=10 and F=5, only 1.1 feedback documents on the average could be collected. Also, the number of topics with an empty feedback set increases as B decreases. Yet for some topics a high number of feedback documents can be found even with low effort thresholds. For example, if browsing and feedback effort are set 5, for some topic 4 highly relevant threshold documents can be found. As one might expect, there is a weak connection between the relative sizes of B and F and the average window size actually used. As it is difficult to find enough highly relevant documents to fill up the size of F, in many instances the average search length is actually close to B. For example, with B = 30 and F = 10, the average search length is 29.6 – nearly the whole window of B=30.

Table 3 presents the case of a regular user accepting both fairly and highly relevant documents as feedback.

Table 3. Regular user (R = 2): the effect of user effort on the availability of RF. All feedback documents are at least fairly relevant.

Browsing Effort B	Feedback Effort F	Average No of RF Docs Available	No of Topics with no RF Docs	Max No of RF Docs per Topic	Average Search Length
30	30	6.3	4	21	30.0
30	10	5.5	4	10	27.6
30	5	3.9	4	5	21.6
30	1	0.9	4	1	7.3
10	10	3.2	8	10	10.0
10	5	2.6	8	5	9.3
10	1	0.8	8	1	4.4
5	5	1.8	11	5	5.0
5	1	0.7	11	1	3.1
1	1	0.6	18	1	1.0

Compared to the previous table, in Table 3 the average number of RF document reaches clearly higher values. Also, the number of topics with an empty RF set is much smaller here. It is still difficult to find enough highly relevant documents to fill up the size F, so in many instances the average search length is also actually very close to B. By selecting a regular threshold instead of the stringent threshold, more feedback documents become available within a selected threshold, but the price of this is that their quality varies more than in case of using a stringent threshold.

Table 4 presents the case where the user accepts even the marginally relevant documents (relevance level 1) as feedback documents.

Table 4. Liberal user (R = 1): the effect of user effort on the availability of RF. All feedback documents are at least marginally relevant.

Browsing Effort B	Feedback Effort F	Average No of RF Docs Available	No of Topics with no RF Docs	Max No of RF Docs per Topic	Average Search Length
30	30	9.4	3	26	30.0
30	10	7.2	3	10	25.2
30	5	4.4	3	5	15.6
30	1	0.9	3	1	5.3
10	10	4.2	4	10	10.0
10	5	3.4	4	5	9.1
10	1	0.9	4	1	3.4
5	5	2.4	5	5	5.0
5	1	0.9	5	1	2.7
1	1	0.7	12	1	1.0

Now the number of feedback documents is rather high (almost 10) when both the browsing effort and the feedback effort thresholds are set high (30). However, in this case the user could expect that many of the feedback documents are actually of low quality. On the other hand, there are clearly more feedback documents available. The relationship between the quantity and the quality of the feedback cannot be solved by looking at the quantity of the feedback data available only. Therefore, next we proceed on testing what happens to the retrieval effectiveness when various user scenarios are used.

3.2. Effect of User Scenario on Feedback Effectiveness

In this section, we study the effect of the quality and the quantity of the relevance feedback to the effectiveness of RF. The results using the stringent relevance evaluation threshold are presented in Table 5.

Table 5. Average precision of user feedback scenarios. Stringent relevance threshold is used in evaluation - baseline MAP = 20.2 %

		MAP by Recognition of Relevance R, %					
Browsing effort B	Feedback effort F	R = 3	Diff. to baseline (% units)	R = 2	Diff. to baseline (% units)	R = 1	Diff. to baseline (% units)
30	30	**37.5**	+17.3	**27.1**	+6.9	**24.9**	+4.7
30	10	**37.5**	+17.3	**27.1**	+6.9	**24.9**	+4.7
30	5	**36.9**	+16.7	**27.5**	+7.3	**23.9**	+3.7
30	1	**31.7**	+11.5	**23.3**	+3.1	**22.6**	+2.4
10	10	**28.9**	+8.7	**24.7**	+4.5	**22.9**	+2.7
10	5	**28.6**	+8.4	**23.9**	+3.7	**23.5**	+3.3
10	1	**27.1**	+6.9	**22.7**	+2.5	**22.2**	+2.0
5	5	**25.9**	+5.7	**23.0**	+2.8	**22.9**	+2.7
5	1	**24.5**	+4.3	**22.7**	+2.5	**22.2**	+2.0
1	1	**20.8**	+0.6	**21.6**	+1.4	**22.0**	+1.8

The baseline MAP figure 20.2 % corresponds to the search result for the 41 initial queries measured by stringent criteria. The differences with respect to the baseline are percentage units, not percentages. In case of every user feedback scenario the changes were positive with respect to baseline MAP. It seems that on the average, the searcher can expect good feedback results even pointing only one feedback document as

long as it is highly relevant. This is shown in Table 5 as an improvement of +11.5 % units in case of a user scenario *M* = <3, 30, 1>. Notice, however, that if the relevance threshold for the feedback document is lower, such an improvement does not take place, even though we know that there are many feedback documents available (Tables 3 - 4). The improvement in average precision is only +3.1 % units in case of user scenario <2, 30, 1> (the 6th column) even though the relevance feedback document may be highly relevant occasionally. This fact is probably due to the differences in the terminological properties of the documents at various relevance levels. Interestingly, the user strategy of a hard working user (B = F = 30) who collects lots of feedback documents using a liberal RF threshold (R = 1, 7th and 8th columns) is not as successful (+4.7 %-units). It seems to be essential that the user keeps the RF threshold high. As we can see from Table 5, the scenarios <3, 30, 30>, <3, 30, 10>, and <3, 30, 5> give by far the best of all results (improvements of +16.7 to +17.3 % units), while the scenarios <2, 30, 30>, <2, 30, 10>, <2, 30, 5> fall behind (improvements of +6.9 to +7.3 % units). Of course, for the scenario <3, 30, 30> there are seldom RF documents available even close to F=30 in the window B=30. In conclusion, considering the relevance feedback quality, the quality of the input matters.

The results of the feedback effectiveness using the regular relevance evaluation threshold are presented in Table 6.

Table 6. Average precision of user feedback scenarios. Regular relevance threshold is used in evaluation - baseline MAP = 22.7

MAP by Recognition of Relevance R, %							
Browsing effort B	Feedback effort F	R = 3	Diff. to baseline (% units)	R = 2	Diff. to baseline (% units)	R = 1	Diff. to baseline (% units)
30	30	**30.8**	+8.1	**34.7**	+12.0	**32.0**	+9.3
30	10	**30.8**	+8.1	**34.8**	+12.1	**31.9**	+9.2
30	5	**30.9**	+8.2	**33.7**	+11.0	**30.4**	+7.7
30	1	**27.6**	+4.9	**27.7**	+5.0	**26.6**	+3.9
10	10	**26.7**	+4.0	**30.8**	+8.1	**30.0**	+7.3
10	5	**26.4**	+3.7	**30.4**	+7.7	**29.2**	+6.5
10	1	**25.0**	+2.3	**27.0**	+4.3	**26.2**	+3.5
5	5	**24.9**	+2.2	**27.6**	+4.9	**27.4**	+4.7
5	1	**24.2**	+1.5	**26.5**	+3.8	**26.1**	+3.4
1	1	**23.7**	+1.0	**24.9**	+2.2	**25.3**	+2.6

In Table 6 the baseline MAP of 22.7 % corresponds to the search result for the 41 initial queries measured by regular criteria. Also here, in every user feedback scenario the changes were positive with respect to baseline. An essential trend compared to the previous table seems to be that here the differences are smaller between the user scenarios having different threshold for accepting feedback documents. The scenarios <2, 30, 30>, <2, 30, 10>, and <2, 30, 5> give the best results (improvements of +11.0 % units to +12.1 % units).

Table 7 presents the effectiveness figures when a liberal evaluation threshold is used. The baseline MAP figure 20.7 % corresponds to the search result for the 41 initial queries measured by liberal criteria. The trend noticed previously in Tables 5 and 6 is accentuated here: now the difference is very small between the user scenarios having different threshold for accepting feedback documents. If the final result set is evaluated by using a liberal threshold (Table 7), the results do not grow better by using a high threshold in selecting the RF documents. The situation is completely different if the final result set is evaluated by using a stringent threshold (Table 5) – in that case the user clearly should keep also high threshold in selecting the feedback documents.

Table 7. Average precision of user feedback scenarios. Liberal relevance threshold is used in evaluation - baseline MAP = 20.7 %

		MAP by Recognition of Relevance R, %					
Browsing effort B	Feedback effort F	R = 3	Diff. to baseline (% units)	R = 2	Diff. to baseline (% units)	R = 1	Diff. to baseline (% units)
30	30	**26.5**	+5.8	**29.5**	+8.8	**30.2**	+9.5
30	10	**26.5**	+5.8	**29.4**	+8.7	**30.1**	+9.4
30	5	**26.6**	+5.9	**28.6**	+7.9	**28.7**	+8.0
30	1	**24.4**	+3.7	**24.2**	+3.5	**24.0**	+3.3
10	10	**23.9**	+3.2	**26.7**	+6.0	**27.5**	+6.8
10	5	**23.7**	+3.0	**26.4**	+5.7	**26.9**	+6.2
10	1	**22.6**	+1.9	**23.6**	+2.9	**23.7**	+3.0
5	5	**22.8**	+2.1	**25.0**	+4.3	**26.0**	+5.3
5	1	**22.1**	+1.4	**23.3**	+2.6	**23.5**	+2.8
1	1	**21.6**	+0.9	**22.5**	+1.8	**22.9**	+2.2

3.3. Comparison to PRF

We also tested the effectiveness of PRF by extracting terms from the top B documents ($B \in \{1, 5, 10, 30\}$) and added them to the initial query as in RF (see Section 2.4). These results are presented in Table 8.

Table 8. Average precision of PRF scenarios evaluated by stringent, regular and liberal relevance thresholds.

PRF Set Size	PRF MAP (%) Stringent	Diff. to baseline (% units)	PRF MAP (%) Regular	Diff. to baseline (% units)	PRF MAP (%) Liberal	Diff. to baseline (% units)
30	19.8	-0.4	25.1	+2.4	24.2	+3.5
10	19.5	-0.7	25.8	+3.1	24.5	+3.8
5	21.2	+1.0	25.8	+3.1	24.1	+3.4
1	22.0	+1.8	25.3	+2.6	22.8	+2.1

In Table 8, columns 2 to 3, we can see that our PRF method hardly improves the baseline results when stringent relevance threshold is used in evaluation. The improvements are small (at best only +1.8 % units) compared to the great improvements gained in the best user RF scenarios (+17.3 % units at scenario <3,30,10>) (Table 5).

The columns 4 to 5 show that PRF improves the baseline results slightly when the regular relevance threshold is used in evaluation. The best improvement is +3.1 % units compared to the baseline when the top 5 documents are used in pseudo-relevance feedback. This improvement is modest compared to the improvement in the best user RF scenarios (+12.1 % units using scenario <2,30,10>) (Table 6).

In columns 6 to 7, we can see that the same trend continues also when the liberal relevance threshold is used in evaluation. The improvements here are closer to the improvements of the user RF scenarios evaluated at the liberal relevance threshold, although the very best user RF scenario improvement of 9.5 % units is gained using scenario <1,30,30>. Here we can see that as the quality of the user RF sinks, it approaches PRF, and the effects become similar.

4. Discussion

Our original research questions were as follows:

1. How effective is RF when we consider various levels of relevance in evaluation?
2. How is the quality and quantity of the feedback documents related to the effectiveness?
3. Can the simulated relevance feedback successfully compete with pseudo-relevance feedback (PRF), and if so, by what effort in terms of the amount and quality of the user feedback?

For the first and the second research questions, our results indicate that RF can be effective at all three evaluation levels. When the stringent evaluation criterion for the final results is used (Table 5), if the user keeps also the feedback threshold high, as in scenario <3, 30, 30> the MAP of RF run improves from 20.2 % (baseline) to 37.5 %. However, if the user lowers the feedback threshold (user scenario <1, 30, 30>) the MAP of the RF run improves only from 20.2 % to 24.9 %. Also, the case of a single "pearl" feedback document (user scenario <3, 30, 1>) outperformed the case of several "mixed" documents (user scenario <1, 30, 5>); MAP values are 31.7 % and 23.9 %, respectively. Thus it seems that one cannot compensate even a small amount of high quality feedback by giving lots of low quality feedback if the stringent criterion is applied in the evaluation phase.

On the other hand, if the final evaluation criterion is liberal, the opposite happens (Table 7). For example, the RF scenario <3, 30, 30> performs worse (MAP = 26.5 %) than the scenario <1, 30, 30> (MAP = 30.2 %).

For the third research question, our PRF method improved the search results evaluated by any relevance level, but it was not very competitive with the best RF user scenarios when the stringent evaluation criterion was used. However, if the liberal evaluation criterion was used, PRF was close to the best RF user scenarios.

5. Conclusions

In real usage situations, the users of information systems would often be best served by enabling them to find the very best documents instead of collecting also marginally relevant documents. As the users are also able to identify highly relevant documents, it is natural to consider developing relevance feedback methods concentrating on finding especially the highly relevant documents. In this paper, we explore the effects of the quality and quantity of the relevance feedback documents to the effectiveness of the feedback measured at various relevance levels.

First we developed a simple user model which makes it possible to quantify three interaction decisions involved in relevance feedback: (1) the relevance criterion (threshold to accept documents used as the feedback), (2) the browsing effort, and (3) the feedback effort of the user. We measured the effectiveness of the final retrieved set after the RF by simulating the user behavior in a laboratory setting based on various user scenarios (three different relevance thresholds, ten different combinations of browsing and feedback efforts) and compared these RF methods to the pseudo-relevance feedback.

The best RF scenarios clearly outperformed all PRF scenarios, although PRF also improved the initial retrieval. When the stringent threshold was used in evaluation, the best user scenario clearly outperformed PRF, but instead, when a liberal evaluation threshold was used, the performance of the user scenarios in RF was close to the PRF results. This hints to the possibility that using binary relevance with a low relevance threshold hides meaningful variation caused by documents which actually belong to various relevance levels, as both marginally, regularly and highly relevant documents are seen as similar.

Acknowledgements

The InQuery search engine was provided by the Center for Intelligent Information Retrieval at the University of Massachusetts. ENGTWOL (Morphological Transducer Lexicon Description of English): Copyright © 1989-1992 Atro Voutilainen and Juha Heikkilä. TWOL-R (Run-Time Two-Level Program): Copyright © Kimmo Koskenniemi and Lingsoft plc. 1983-1993.

This research was funded by the Academy of Finland under Project Numbers 177033 and 1209960. The authors thank the anonymous referees and members of the research groups FIRE and IRiX for useful suggestions.

References

1. Blair, D. C., (1984) The Data-Document Distinction in Information Retrieval, *Communications of the ACM*, 4, Vol. 27, 1984, 369-374.
2. Dennis, S., McArthur, R. & Bruza, P.D. (1998). Searching the World Wide Web made easy? The cognitive load imposed by query refinement mechanisms. In: Proceedings of the 3rd Australian Document Computing Conference, Sydney, Australia. Sydney: University of Sydney, Department of Computer Science, TR-518: 65-71.
3. Efthimiadis, E.N. (1996). Query expansion. In: Williams, M.E. (Ed.), *Annual Review of Information Science and Technology, vol. 31 (ARIST 31)*. Medford, NJ: Learned Information for the American Society for Information Science: 121-187.
4. Kekäläinen, J. (1999). *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Tampere, Finland: University of Tampere, Department of Information Studies, Ph.D. Thesis, Acta Universitatis Tamperensis 678. [Available at: http://www.info.uta.fi/tutkimus/fire/archive/QCES.pdf . Cited Oct. 31 2005.]
5. Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. *Information Processing & Management, 41*(5): 1019-1033.
6. Kekäläinen, J. & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology, 53*(13): 1120-1129.
7. Pirkola, A., Leppänen, E. & Järvelin, K. (2002). The RATF Formula (Kwok's Formula): Exploiting average term frequency in cross-language retrieval. *Information Research*, 7(2). Available at: http://InformationR.net/ir/7-2/infres72.html
8. Ruthven, I., Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, 18*(2): 95-145.
9. Ruthven, I., Lalmas, M. & van Rijsbergen, K. (2003). Incorporating user search behaviour into relevance feedback. *Journal of the American Society for Information Science and Technology, 54*(6): 529-549.
10. Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis And Retrieval of Information by Computer*. Reading, MA: Addison-Wesley.
11. Sihvonen, A. & Vakkari, P. (2004). Subject knowledge, thesaurus-assisted query expansion and search success. In: RIAO 2004. *Coupling Approaches, Coupling Media And Coupling Languages for Information Retrieval, Proceedings of RIAO 2004 conference*. Paris: C.I.D: 393-404.
12. Sormunen, E. (2002). Liberal relevance criteria of TREC – Counting on negligible documents? In: Beaulieu, M. & Baeza-Yates, R. & Myaeng, S.H. & Järvelin, K. (Eds.) *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 25)*, Tampere, Finland, August 11-15, 2002. New York, NY: ACM Press: 320-330.
13. Sormunen, E., Kekäläinen, J., Koivisto, J., Järvelin, K. (2001). Document Text Characteristics Affect the Ranking of the Most Relevant Documents by Expanded Structured Queries. *Journal of Documentation, 57*(3):358-374.
14. Vakkari, P., Sormunen, E. (2004) The Influence of Relevance Levels on the Effectiveness of Interactive Information Retrieval. *Journal of the American Society for Information Science and Technology*, 55(11): 963-969.
15. Voorhees, E., (2001) Evaluation by Highly Relevant Documents, Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 24), New Orleans, Lousiana, USA, September 9-13, 2001. New York, NY: ACM Press: 74-82.