Evaluating the Effectiveness of Relevance Feedback Based on a User Simulation Model: Effects of a User Scenario on Cumulated Gain Value

Heikki Keskustalo[1] ● **Kalervo Järvelin** ● **Ari Pirkola**

Abstract We propose a method for performing evaluation of relevance feedback based on simulating real users. The user simulation applies a model defining the user's relevance threshold to accept individual documents as feedback in a graded relevance environment; user's patience to browse the initial list of retrieved documents; and his/her effort in providing the feedback. We evaluate the result by using cumulated gain-based evaluation together with freezing all documents seen by the user in order to simulate the point of view of a user who is browsing the documents during the retrieval process. We demonstrate the method by performing a simulation in the laboratory setting and present the "branching" curve sets characteristic for the presented evaluation method. Both the average and topic-by-topic results indicate that if the freezing approach is adopted, giving feedback of mixed quality makes sense for various usage scenarios even though the modeled users prefer finding especially the most relevant documents.

1. Introduction

We address in this paper the issue of evaluating simulated relevance feedback (RF) specifically from the user viewpoint. We describe our simulation and evaluation methodology and demonstrate it by performing RF simulations in the laboratory setting using empirical data containing graded relevance assessments.

The starting point for our simulations is the fact that selecting good search keys is difficult yet important (Efthimiadis 1996). Therefore, the first query formulation of the user often acts as an entry to the search system followed by subsequent phases of browsing and query reformulations (Marchionini et al. 1993). In such a context, relevance feedback based on the initial retrieved set of documents offers one solution to the query reformulation problem (Ruthven and Lalmas 2003). We focus here on automatic RF where the user provides document level feedback to the information retrieval system; additional keys are automatically extracted from the selected feedback documents and used for the query reformulation.

According to earlier studies, users of information retrieval systems might actually prefer finding especially highly relevant documents (Järvelin and Kekäläinen 2000; Voorhees 2001) and they are also able to identify documents belonging to various relevance levels (Vakkari and Sormunen 2004). Moreover, the textual characteristics of the documents belonging to various relevance levels differ (Sormunen et al. 2001). First, a larger number of aspects of the request are discussed in highly relevant documents, and secondly, a larger set of unique expressions is used in them. These observations lead us to ask how effective RF is when we assume a user preferring highly relevant documents, and when we modify the quality and quantity

---

[1] H. Keskustalo · K. Järvelin · A. Pirkola

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

Email: Heikki.Keskustalo@uta.fi

K. Järvelin

Email: Kalervo.Jarvelin@uta.fi

A. Pirkola

Email: Ari.Pirkola@uta.fi

of user feedback assuming differences in the final browsing length that the user tolerates, during the evaluation phase.

IR experiments regarding user simulations may be classified into four classes: (i) observing users in real situations (i.e., real users; no simulation), see, e.g., Spink and Saracevic (1998); (ii) observing users performing simulated tasks (e.g., Belkin et al. 1995); (iii) performing simulations in the lab without users (simulation; no users) (e.g., Keskustalo et al. 2006; White et al. 2004); and (iv) traditional lab research (no users and no simulation point of view regarding user attributes or user action). Studies on real users performing real or simulated RF tasks provide realistic and rich data but it is difficult to cover extensively numerous test cases. On the other hand, laboratory studies typically do not model users explicitly, though they can be seen as abstractions of user searching (e.g., Rocchio 1971). Our goal in this paper is to extend the lab model towards the user point of view and perform user simulations in the lab (without users) to explore explicitly the consequences of variation in user's feedback behavior (class (iii) above).

Throughout this paper we will utilize a test collection containing graded relevance assessments. We model users who interact with the search system via a simplified feedback process consisting of two phases. In the first phase the *initial query* is performed and the user selects some (possibly none) of the retrieved documents as feedback documents. Feedback terms are extracted from these documents automatically and added into the initial query to form the reformulated query (the *RF query*). In the second phase the RF query is performed and the final document list is evaluated.

We address two research questions. First, how should we evaluate the effectiveness of simulated user feedback considering the issue of graded relevance assessments? In order to perform a user-centered evaluation we need to answer this question first. Note that the user attitude towards relevance may be different during the feedback phase and the evaluation phase, for example, the user might purposefully want to accept low quality feedback in order to find documents of highest quality if such a strategy is known to be successful. We will address the evaluation question in Section 2.4 and propose a solution to the problem. Secondly, how successful are various relevance feedback strategies? In Section 3 we will analyze the effectiveness of several user feedback scenarios considering varying quality and quantity of feedback, varying levels of user patience, and different requirements for relevance.

We utilize the simple *user model* introduced in our earlier study (Keskustalo et al. 2006) for defining user scenarios. It models user's willingness to browse; willingness to provide feedback; and tolerance towards accepting documents as feedback. These dimensions are motivated in Section 2.2. The present paper differs essentially from our earlier study (Keskustalo et al. 2006) in several aspects. First, we now freeze all documents seen by the simulated user during browsing in order to imitate closely the user viewpoint. We will explain the importance of this aspect in Section 2.4. Our first study did not apply freezing because we compared the effectiveness of user-based RF with pseudo RF per se. Secondly, in the present study we use a directly user-oriented evaluation measure, i.e., cumulated gain (CG) (Järvelin and Kekäläinen 2000) whereas a system-oriented measure was used in our first study, i.e., non-interpolated average precision (MAP). Third, we measure the effectiveness up to the last rank position of interest (position 10, 20 or 100) for the modeled user, using thus relatively short document lists (to demonstrate the user orientation), while all the ranks up to 1000 were used in the first study (a system-oriented measure). Fourth, in the current paper we also present novel topic-by-topic results.

Our main motivation for developing simulation methods is to understand the consequences of a specific user feedback behavior. The hypotheses derived in this way may be verified with real users and utilized in user education and in systems design. For example, we might be interested in predicting what kind of feedback strategies (e.g., the role of patience or impatience towards giving feedback) are likely to succeed or fail assuming specific user requirements for documents (e.g., a user preferring especially the most relevant documents or a user liberally accepting many documents as relevant).

The rest of the paper is organized as follows. In Section 2 we explain our experimental methodology – the test collection; modeling some basic parameters of user feedback; the retrieval system; the construction of the feedback queries for simulating specific user types; and the evaluation and freezing methods used. Section 3 presents our findings, and Section 4 discusses the results and concludes the paper.

2. Methods
2.1 Test collection

We used the test collection from TREC 7 and TREC 8 ad hoc tracks in the experiment including 528,155 documents together with 41 topics with graded relevance assessments (Sormunen 2002). The relevance assessments were done using a four-point relevance scale: (0) non-relevant, (1) marginally relevant, (2) fairly relevant, and (3) highly relevant documents. In the creation of the collection with graded assessments, the original TREC assessed documents (all relevant and a sample of the non-relevant) were reassessed - altogether 6122 documents. In the recall base there are 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, there are on average 59 documents for each topic which are at least marginally relevant. The database index was constructed by lemmatizing the words using ENGTWOL morphological lemmatizer by Lingsoft Inc. The strengths of the collection include its size and the number of topics. Still the use of a single collection limits the generalizability of the findings discussed below. However, this is no limitation of the approach we propose: it may be followed when other large collections with graded assessments are available.

2.2 User Modeling

Human relevance feedback given at the document level is a complex process. For example, the number of top documents the user is willing to browse in order to give feedback varies; the number of feedback documents the user is willing to collect may vary; and the user may also have many criteria for selecting the feedback documents. In the current paper we utilize the user model $M = <R, B, F>$ described in Keskustalo et al. (2006) in order to simulate users who want to improve their final result by providing relevance feedback and collect it up to some limit while using a specific relevance threshold in accepting documents as feedback. Our model consists of the following three attributes:

1. The requirement of document relevance (relevance feedback threshold R) to accept a document as feedback.
2. The willingness to browse the initial document list (browsing window size B).
3. The willingness to provide feedback (feedback set size F).

Each value combination $<R, B, F>$ ($F \leq B$) defines a unique *user scenario* which characterizes the user's feedback behavior.

The requirement of document relevance, R, is an important dimension since different users may have different thresholds for accepting feedback documents, e.g., a user may want to focus on highly relevant documents only, or liberally accept feedback documents from several relevance levels (Kekäläinen and Järvelin 2002; Voorhees 2001). According to Vakkari and Sormunen (2004) the users are also able to identify highly relevant documents, while marginal documents more easily escape their attention. We model the requirement of relevance to accept a document as feedback by the possible values of graded relevance ($R \in \{1, 2, 3\}$). The liberal threshold, R=1, indicates that the user correctly recognizes and accepts both marginally relevant documents (relevance level 1), fairly relevant documents (relevance level 2), and highly documents (relevance level 3) as feedback. The regular threshold, R=2, indicates that the user accepts both the fairly and highly relevant documents as feedback. Finally, the stringent threshold, R=3, indicates that the user only accepts highly relevant documents as feedback.

The willingness to browse, B, is used to model the user's willingness to search through the initial ranked list of documents. This dimension is motivated by the fact that the willingness of the user to study the retrieved lists is limited ("the futility point") (Blair 1984) and it varies. For example, patent searchers may sometimes require high recall (Kando, 2000). Such users may need to scan through long lists of retrieved documents, making modeling of high values of B and high final evaluation length reasonable. On the other hand, some searchers are more precision-oriented. For example, family practice physicians may have a limited amount of time allotted to retrieval during the patient visit (Price et al., 2007) making scanning through long lists impossible. Such usage situations can be modeled by smaller values of B and final evaluation lengths. We model the browsing length dimension by the maximum number of documents considered (window size B). For example, B=1 indicates a very impatient user who is willing to consider only the first document for relevance feedback. On the other hand, B=30 indicates a very patient user who

is willing to examine a long list of documents (at the maximum 30 documents). In the present study we will perform simulations with a limited set of values for B, i.e., B∈ {1, 5, 10, 30}.

The willingness to provide feedback, F (≤B), models the user's willingness to mark documents as relevant. (the unmarked documents are assumed to be non-relevant). The user will examine the initial result at most up to B documents but quits this examination and launches the RF query as soon as the maximum of F relevant feedback documents (of relevance level R or higher) have been identified. However, if the particular result has less than F documents of the required level, the user does not continue looking for feedback documents beyond the limit B. Dimension F is separate from B since even if the user is willing to browse through a long list he may give up marking documents relevant after finding one (or a few) relevant items. It is essential to consider because the quantity of the feedback (together with its quality) may be critical to success. Often users have been found to be reluctant to give feedback (Ruthven and Lalmas 2003; Jordan et al. 2006) motivating low F, whereas it is also interesting to study what happens if users do use lots of feedback (motivating higher values of F). For example, F=1 indicates a very reserved user who will give up giving feedback immediately after finding the first relevant document. On the other hand, F=30 indicates a very eager user who is willing to provide lots of feedback. We will use a limited set of values for F in our simulations, i.e., F∈ {1, 5, 10, 30}.

Any *user scenario* defines uniquely the set of feedback documents for a topic once we have available both the initially retrieved list of documents and the recall base information. Thus, once a user scenario has been specified, we can automatically recognize its corresponding feedback documents for the simulations. The expansion keys can be extracted automatically from these documents (using some extraction method) and added to the initial query in order to form the final (expanded) RF query related to a specific user scenario.

In the present paper, we will simulate several types of user behavior by utilizing the parameter space of our user model. First, we model the behavior of a *patient* user by the user scenario <R, B, F> = <1,30,30>. This kind of user selects feedback documents from the initial result list using a low relevance threshold (R=1), thus accepting marginally, fairly and highly relevant documents as feedback. The user is considered to be patient as he is prepared to browse 30 documents (B=30), and does not give up before finding all possible relevant documents within this browsing window size (F=30). We also experiment with user scenarios <2,30,30> and <3,30,30> which differ from the previous one by different threshold R (a regular threshold R=2, and a stringent threshold R=3) for accepting documents as feedback.

We model *moderately patient* users via user scenarios <R,10,10> and <R,10,5>, R∈ {1, 2, 3}, where a browsing window size 10 is used (B=10) and the user gives up after finding F relevant document (F=10 or F=5). *Impatient* users are modeled by user scenarios <R,1,1> and <R,5,1>, R∈ {1, 2, 3}. Impatient users tolerate only very small browsing window sizes (either B=1 or B=5), and they give up immediately after finding the very first relevant document (F=1). Slightly more patient users are modeled by user scenarios <R,5,5>, R∈ {1, 2, 3}. Here the modeled user provides feedback from all relevant document within the top five initial documents (F=5).

2.3 Retrieval System and Feedback Queries

The *InQuery* system based on Bayesian inference networks (Broglio et al. 1994) was used in the experiment. Each initial query was based on TREC topic wordings which were lemmatized, excluding the stop list words, having the structure `#sum(#syn(… key …), … #syn(… key …))` where the synonym (*syn*) operator treats all of its arguments as instances of one search key. These queries were used as baseline queries. The synonym structure was used due to the fact that some keys were ambiguous and produced two or more lemmas, in which case they all were included in one synonym set.

The expansion keys were extracted from the whole text of the feedback documents using the *RATF* weighting formula (Pirkola et al. 2002).

$$RATF(k) = (cf_k / df_k) * 10^3 / ln(df_k + SP)^p \qquad (1)$$

where
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 $p = 3$ (Keskustalo et al. 2006).

In the expansion key extraction, a word list containing the 50 best keys was extracted from each feedback document 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 and the 30 best expansion keys (keys shared by the greatest number of word lists) were selected as the expansion keys. The expansion keys were formed into a simple *sum* operation `#sum(… key …)` and the final RF query was produced by combining the two queries, with equal weights, within an outmost *sum* structure: the initial baseline query, and the expansion key structure. Non-relevant documents were not applied (negatively) when the final RF query was constructed.

Overall, the process consisted of the following steps. For each topic (N=41):

1. The title and description fields were automatically formulated into the *initial query*.
2. The initial query was run in the test collection and the result list (a ranked list of documents) was retrieved.
3. The set of feedback documents was extracted from the result list by utilizing each user scenario <R, B, F> together with the recall base information.
4. A set of expansion keys was extracted from the feedback documents for each user scenario.
5. The *RF query* was constructed for each user scenario by combining its set of expansion keys to the initial query.
6. The RF queries were run in the test collection.
7. The freeze all approach was applied (see Section 2.4, Table 1) and the search result was evaluated using cumulated gain (CG) based evaluation.

2.4 Evaluation

The traditional system-oriented way to consider IR system performance is to measure precision over standard recall points. A typical measure used is the non-interpolated average precision (MAP). In relevance feedback the user assesses the relevance of some documents initially retrieved and they are utilized in subsequent query modification (RF query). When RF is evaluated, some freezing approach is typically applied during evaluation to prevent artificial improvement of the search results as a consequence of re-ranking relevant documents already seen (the "ranking effect" (Chang et al. 1971)).

In the traditional freezing approach (Salton 1989) the previously retrieved items identified as relevant are kept "frozen" in their original ranks. The previously retrieved non-relevant items are removed from the collection and their ranks are occupied by items that are newly retrieved in subsequent search iterations. This approach calculates the effect of feedback on the remainder of a search (finding more unseen relevant documents). First, because of freezing, the seen relevant documents are not re-retrieved with better ranks. Secondly, non-relevant documents are not re-retrieved either, which makes sense from the user point of view.

The traditional user-oriented way to consider performance is to measure precision at specific cut-off points. Also cumulated gain (CG) based evaluation methods (Järvelin and Kekäläinen 2000) can be applied as a user-oriented measure. In CG the degree of relevance of each document is used as a gained value measure for its ranked position in the result list. The gain is summed progressively from rank 1 to *N*. Both precision and CG based measures can be applied to results frozen with the traditional freezing method described above.

Generally speaking, the selection of the effectiveness measure depends on what we want to study. Measures based on CG are directly user-oriented in focusing on the *N* top-ranked documents. They allow us to define weighting schemes related to the relevance levels. The weights reflect the values the user gives to documents of different relevance levels (Järvelin and Kekäläinen 2000). For example, assuming our

four-point relevance scale, non-relevant documents (level 0) may be given a zero weight; marginally relevant documents (level 1) weight 1; regularly relevant documents (level 2) weight 10; and highly relevant documents (level 3) weight 100. Various weighting schemes may be utilized to reflect different user types modeled.

In the present paper our goal is to perform user-oriented evaluation of simulated RF where the feedback phase is an extension to the initial retrieval phase. We propose an evaluation procedure based on using CG measure using full freezing (Chang et al. 1971) and removal of documents which were used in RF during the evaluation of the feedback phase. In other words, we freeze all those documents in their original ranks - both relevant and non-relevant – seen by the user during the initial browsing phase. All the yet unseen documents (returned by RF) are simply placed into the following rank positions (see Table 1). Table 1 illustrates the effects of different approaches regarding freezing.

Table 1. Examples of ranked results formed using different freezing principles (relevance feedback window size = 5). Legend: First row: the original ranked result. Second row: the corresponding "raw" relevance feedback result. Third row: the corresponding "traditionally" frozen result. Fourth row: the result after *freeze all* approach (freezing all documents seen) is applied.

Original	d1:0	**d2:3**	d3:0	d4:0	d5:0	d6:2	**d7:3**	d8:0	d9:1	d10:0
Raw RF	**d2:3**	**d7:3**	d5:0	d6:2	d9:1	d10:0	d11:2	d1:0	d12:1	d3:0
Traditional freezing	**d7:3**	**d2:3**	d6:2	d9:1	d10:0	d11:2	d12:1	d13:3	d14:0	d15:1
Freeze all	d1:0	**d2:3**	d3:0	d4:0	d5:0	**d7:3**	d6:2	d9:1	d10:0	d11:2

The first row presents the result of the original retrieval, for example, *d1:0* denotes document *d1* having relevance value 0 (non-relevant) and *d2:3* denotes document *d2* having relevance value 3 (highly relevant).

The second row, raw RF result, simply reorders the documents including documents already seen by the user, and also introduces some new documents into the top-10. Our earlier study (Keskustalo et al. 2006) utilized original and raw RF results this way only (the first and the second row).

The third row, the traditional freezing result, is derived on the basis of the first two lists (rows). We assume that the user has studied an RF window containing 5 documents. The relevant documents seen by the user are frozen into their ranked positions (*d2* having relevance level 3, at rank 2); the remaining positions seen by the user are "freed"; and the corresponding documents are not included when we construct the frozen result. Therefore, documents *d1, d3, d4*, and *d5* are missing from the traditional freezing result.

The fourth row shows the situation when *freeze all* method is applied. We simply freeze all documents seen (the first five documents) into their ranked positions - also the non-relevant ones. The justification of the procedure lies in the fact that the user has already wasted effort in inspecting all the documents seen, despite their level of relevance. Thus, the *freeze all* procedure allows us to be faithful to the original user point of view to the extreme. The frozen documents (*d1-d5*) are removed from the raw RF result and we add the yet unseen re-ranked documents starting from rank 6 forwards (*d7, d6, d9, …*). Note that in Table 1 the highlighted documents *d2* and *d7* are positioned differently in different cases.

Next we proceed to looking at the findings. As mentioned above, we will use CG in evaluation which takes into account the ranked order of the documents in the result list, and which is summed progressively from rank one to rank *N*.

3. Findings

In this section we present results of relevance feedback simulations based on RF user modeling and compare them with the baseline result. We will use a sharp weighting scheme 0-1-10-100 in all test situations (Figures 1-9), that is, non-relevant documents are given weight 0, marginally relevant documents weight 1, fairly relevant documents 10, and highly relevant documents 100.

As multiple user scenarios were compared to each others, Friedman's test was used (Conover 1999; see also Kekäläinen 1999) to study the statistical significance between the scenarios (Table 3; Figs. 1-9). If significant differences were found overall between the scenarios, pair-wise comparisons were performed to find out which methods differ significantly from each other. The rank for each scenario for a topic was

calculated based on two comparison figures: (i) on the basis of the very last rank belonging to the scenario (called *final gain* case below), and (ii) as an average value from rank 1 up to the last rank (called *avg gain*).

We next present CG results grouped into cases (curve sets) in Figures 1-9. In all cases we present averaged results over the 41 topics. We will present nine figures showing the effect of RF patience to IR results when sharp weighting is used and when we vary the relevance feedback threshold. We show the numbers of actual feedback documents belonging to various relevance levels for each scenario (Table 2) and topic-by-topic results illustrating how many RF queries succeed or fail in each user scenario (Table 4).

3.1 An Impatient User

Figure 1 shows the baseline CG result together with three RF scenarios representing impatient users using small browsing window sizes (at most 5) during the RF phase. A low RF threshold is used in the scenarios, that is, the user accepts relevant documents from all three relevance levels as feedback. Note that our evaluation model allows us simulating this (perhaps counter-intuitive) behavior of a user where a low relevance threshold is used purposefully during the relevance feedback although sharp weighting is employed during the final evaluation. Below, only a small fraction of the list of ranked documents (ranks 1-10) are displayed because only these values are meaningful for an impatient user – he only wants to look at the top documents.
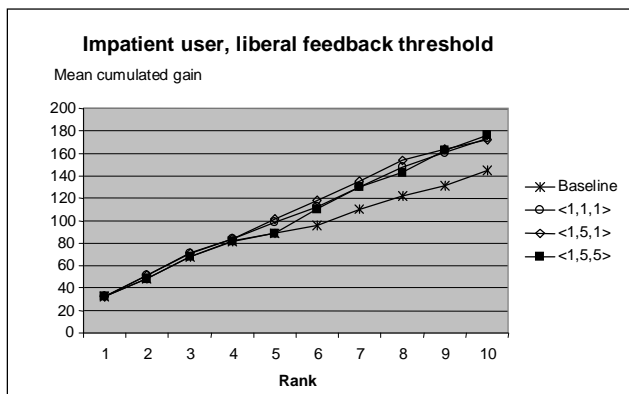

Figure 1. Cumulated gain results (averaged over 41 topics) for a user keeping a liberal relevance feedback threshold but having little tolerance for browsing and giving feedback. Effectiveness (CG) of four scenarios are presented: *baseline* together with three feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>*): *<1,1,1>, <1,5,1>,* and *<1,5,5>*.

…set size>): Tässä kaksoispiste kun taas muissa figure-captioneissa ei ole.

We can see the baseline curve (the lowest curve) together with three simulated cases where the user employs liberal relevance feedback. Relevance feedback always improves the retrieval results as compared to the baseline results. At the maximum rank 10 the baseline scenario reaches the cumulated gain value of 145 while the best feedback scenario <1,5,5> reaches value 176, i.e., an improvement of 21 %. Friedman's test corroborates that significant pair-wise differences exist between the baseline and any of the three RF scenarios both in the *final gain* and in the *avg gain* cases (Table 3). Additionally, in the *avg gain* case, statistically significant difference exists between the methods <1,5,1> and <1,5,5>. Note that because of freezing all the documents seen by the user the performance curve of scenario <1,5,5> can deviate from the baseline curve at rank 6 for the first time. These kinds of "branching" curves are characteristic for the "freeze all" based simulation results. Freezing slightly lowers the *avg gain* values for the scenario <1,5,5>. In contrast to this, the curve of scenario <1,5,1> is able to deviate as soon as the first relevant document has been identified. Note that also the differences between separate query expansion methods could be illustrated by utilizing the freezing together with CG in order to create similar type of branching curve sets.

Figure 2 differs from Figure 1 in that this time the simulated user raises the threshold in accepting documents as relevance feedback. In other words, the user has set a regular relevance feedback threshold and thus only documents from relevance levels 2 and 3 (fairly and highly relevant documents) are accepted. We can see here a similar trend as in Figure 1: relevance feedback significantly improves the retrieval results compared to the baseline results (significant pairwise differences exist between the baseline and any of the three RF scenarios both in the *final gain* and the *avg gain* methods, see Table 3).
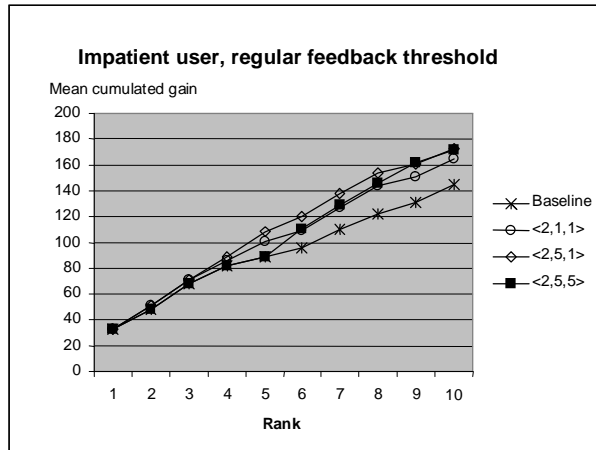
Figure 2. Cumulated gain results for an impatient user keeping a regular relevance feedback threshold. Baseline scenario and feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>*) *<2,1,1>, <2,5,1>*, and *<2,5,5>*.

However, the absolute gain value measured at the maximum rank 10 decreases compared to the situation in Figure 1. Note that two opposing trends simultaneously affect the outcome. First, the quality of the feedback increases, but secondly, there is less such feedback available. The numbers of actual feedback documents belonging to various relevance levels for each scenario are presented in Table 2.

Table 2. Number of actual feedback documents belonging to various relevance levels (averaged over the 41 topics) for different feedback scenarios.

Scenario	R=1	R=2	R=3	Scenario	R=2	R=3	Scenario	R=3
<1,1,1>	0.15	0.27	0.29	<2,1,1>	0.27	0.29	<3,1,1>	0.29
<1,5,1>	0.22	0.37	0.29	<2,5,1>	0.44	0.29	<3,5,1>	0.44
<1,5,5>	0.63	1.02	0.78	<2,5,5>	1.02	0.78	<3,5,5>	0.78
<1,10,5>	0.93	1.49	1.00	<2,10,5>	1.59	1.00	<3,10,5>	1.12
<1,10,10>	1.02	1.95	1.24	<2,10,10>	1.95	1.24	<3,10,10>	1.24
<1,30,30>	3.05	4.07	2.27	<2,30,30>	4.07	2.27	<3,30,30>	2.27

Table 2 shows that when an impatient user raises the relevance feedback threshold (as in scenario <2,5,1>) the improvement in the amount of high quality feedback is very small compared to the scenario with a liberal threshold (<1,5,1>). There are slightly more feedback documents available when a liberal RF threshold is used (scenario <1,5,1>) (0.22+0.37+0.29 documents), albeit of mixed quality, than in case of a using a higher threshold (scenario <2,5,1>) (0.44+0.29 feedback documents).

Note that the set of feedback documents selected by scenario <2,5,5> is a proper subset of feedback documents selected by scenario <1,5,5> because in both scenarios $B=F=5$. This guarantees that the browsing window is browsed to the end.
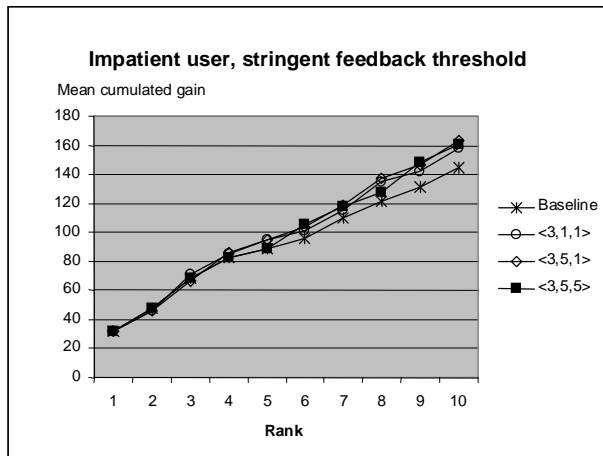
Figure 3. Cumulated gain results for an impatient user keeping a stringent relevance feedback threshold. Baseline scenario and feedback scenarios (<*R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size*>) <*3,1,1*>, <*3,5,1*>, and <*3,5,5*>.

Figure 3 differs from Figures 1 and 2 in that this time the user accepts only highly relevant documents as feedback (R=3, i.e., stringent feedback threshold). All the RF methods significantly improve the retrieval results compared to the baseline (in *final gain* based comparison). Yet the highest gain at the rank 10 does not improve compared to the scenarios based on more liberal feedback thresholds (Figs. 1-2). When we look at the results, it is clear that demanding and being able to find slightly more high quality RF documents does not really improve the situation compared to accepting mixed level feedback (Fig. 1). For an *impatient user* it is difficult to find enough high quality feedback even if he would like to. This is likely to hold in real life as well.

Our general conclusion based on Figures 1-3 is that for an impatient user it makes sense to give mixed level feedback. This method accepts also highly relevant feedback and generally more feedback becomes available. If the user is very picky, it may happen that no feedback is available in small browsing windows. In that case, no improvement can be made regarding the baseline query.
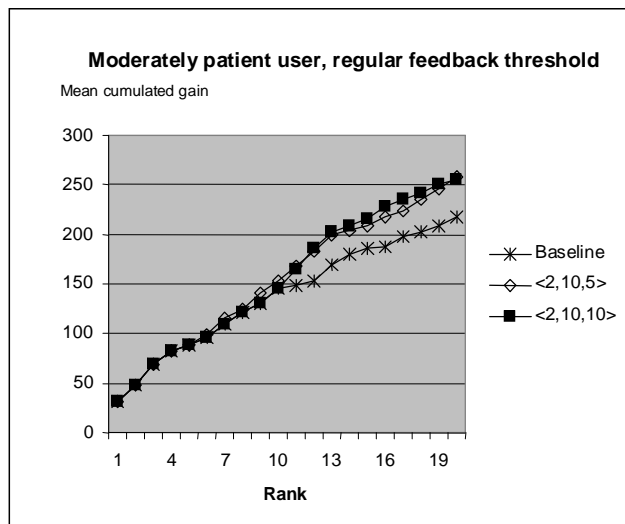
3.2 Moderately Patient User

Next we move from modeling a very impatient user into modeling a more patient user. The ranks inspected now are raised up to 20 because a more patient user is simulated as behaving in this way. This kind of user may give more feedback (he tolerates a longer browsing period during the RF phase) and he is also prepared to look further in the result list (document rank 20 as the maximum rank) during the evaluation phase.

Figure 4 shows the baseline CG result together with two RF scenarios using moderate window sizes (at most 10) during the RF phase. A low RF threshold is used, that is, the user accepts relevant documents from all three relevance levels.
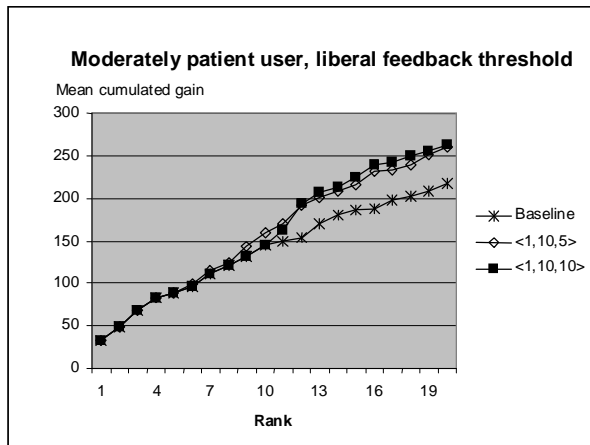
Figure 4. Cumulated gain results for a moderately patient user keeping a liberal relevance feedback threshold. Baseline scenario and feedback scenarios (<*R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size*>) <*1,10,5*> and <*1,10,10*>.

Both RF methods significantly improve the retrieval results compared to the baseline. At the maximum rank 20 the baseline scenario reaches the cumulated gain value of about 217 while relatively small feedback efforts (the scenarios <1,10,10> and <1,10,5>) yield the values of 263 (improvement of 21 %) and 260 (improvement of 20 %). Smaller user effort would suggest <1,10,5> as the optimal behavior.
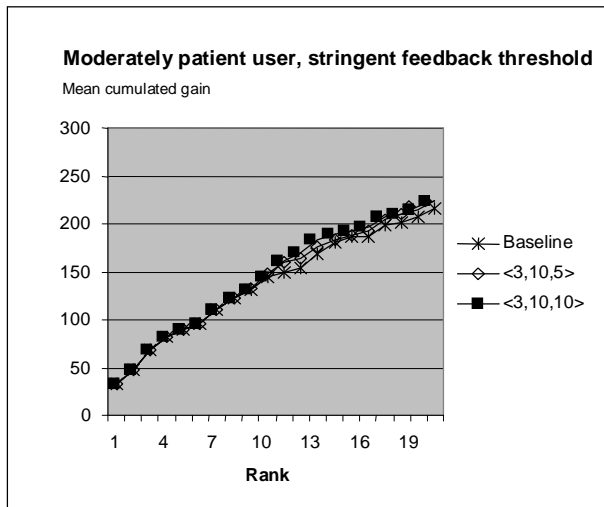

Figure 5. Cumulated gain results for a moderately patient user with a regular relevance feedback threshold. Baseline scenario and feedback scenarios (<*R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size*>) <*2,10,5*>, and <*2,10,10*>.

In Figure 5 the RF threshold is raised to regular level (R=2), i.e., the user accepts both fairly relevant and highly relevant documents as feedback. Also here the feedback always significantly improves the retrieval results as compared to the baseline results. At rank 20 the baseline scenario reaches the CG value of 217, while the scenarios <2,10,10> and <2,10,5> reach the CG values of 255 and 258, respectively. Yet the maximum gains at rank 20 are lower compared to the maximum gains of the corresponding scenarios with a liberal threshold (Fig. 4).

In Figure 6 a stringent RF threshold is used, i.e., the user accepts only highly relevant RF documents. This time no significant differences exist between any scenarios (using either *final gain* or *avg gain* based comparison, Table 3). Also the level of improvement at rank 20 is inconsequential. We can see from Table 2 that the number of highly relevant feedback documents available for the scenarios <3,10,5> and <3,10,10> is small (1.12 and 1.24, correspondingly). For the scenario <3,10,5> it is difficult to find enough

highly relevant documents to fill up feedback set size F=5 before the whole browsing window of size 10 runs out. Therefore, the results start improving late (around rank 10) because of freezing.
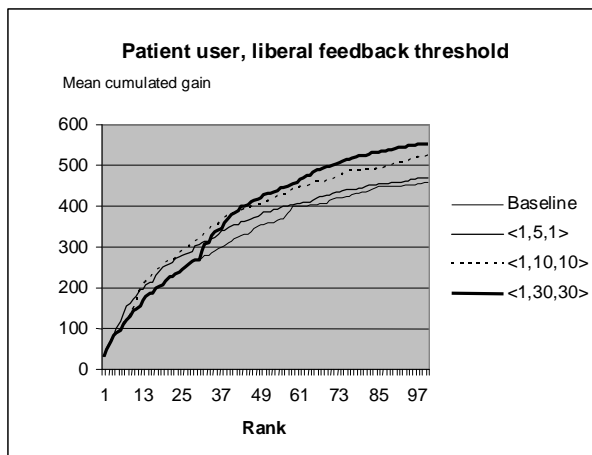

Figure 6. Cumulated gain results for a moderately patient user with a stringent relevance feedback threshold. Baseline scenario and feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>*) *<3,10,5>* and *<3,10,10>*.

Our general conclusion for the *moderately patient* user (Figs. 4-6) is the same as for the impatient user: it makes sense to give moderate amounts of mixed level feedback.

3.3 Patient User

Last, we move on to the *patient* user cases. The last rank inspected by the simulated user is increased to 100 now, because we simulate a very patient (highly motivated) user. This kind of user may possibly give more feedback - he tolerates a long browsing period during the RF phase - and he also is motivated to examine very long result lists (document rank 100 as the maximum rank).

Figure 7 shows the baseline CG result together with three RF scenarios using small and large browsing window sizes (30 at the maximum) during the RF phase.
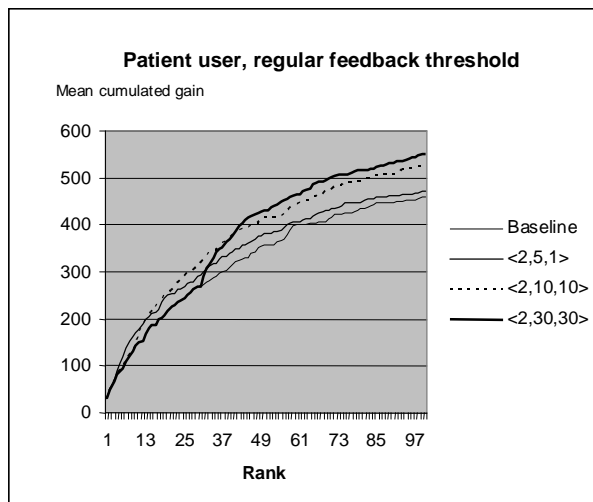

Figure 7. Cumulated gain results for a patient user keeping a liberal relevance feedback threshold. Baseline scenario and feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>*) *<1,5,1>*, *<1,10,10>*, and *<1,30,30>*.

A low RF threshold is used in the scenarios, that is, the user accepts relevant documents from all three relevance levels. We concentrate on ranks 1-100 because these ranks are meaningful for a very patient user. Especially, high rank values (such as 100) are of interest for the very patient user who is demanding valuable results for his high effort.

We can see that relevance feedback always improves the retrieval results as compared to the baseline results. Friedman's test corroborates that there are pair-wise differences between several scenarios in *final gain* and *avg gain* cases (Table 3).

At the maximum rank 100 the baseline scenario reaches the CG value of 460 while those feedback scenarios differing significantly from it (final gain at last rank) reach considerably higher values. The scenario <1,10,10> reaches the cumulated gain value of 525, i.e., an improvement of 14 % compared to the baseline. The scenario <1,30,30> reaches the CG value of 553, i.e., an improvement of 20 %. A good result is rapidly gained by providing lots of feedback despite the depressing effect of the long freezing zone. The scenario <1,5,1> also differs significantly from scenarios <1,10,10> and <1,30,30> using *final gain* based comparison. Our general conclusion on the basis of Figure 7 is that high feedback effort pays off well in the long run, while low effort (the scenario <1,5,1>) does not.

Figure 8 presents the results when the regular RF threshold is used (fairly relevant and highly relevant RF documents are accepted).

Figure 8. Cumulated gain results for a patient user using a regular relevance feedback threshold. Baseline scenario and feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>*) *<2,5,1>, <2,10,10>*, and *<2,30,30>*.

Also here the relevance feedback significantly improves the retrieval results as compared to the baseline results. At the maximum rank 100 the baseline scenario reaches the cumulated gain value of about 460 while two feedback scenarios differ significantly from it (using *final gain* evaluation) and reach considerably higher values. The scenario <2,10,10> reaches the value of 526, and the scenario <2,30,30> the value of 550. The best result is gained using high initial effort (<2,30,30>), while low feedback effort (<2,5,1>) does not pay off.

Last, Figure 9 shows the very patient user cases when a stringent RF threshold is used (only highly relevant documents are accepted as feedback).
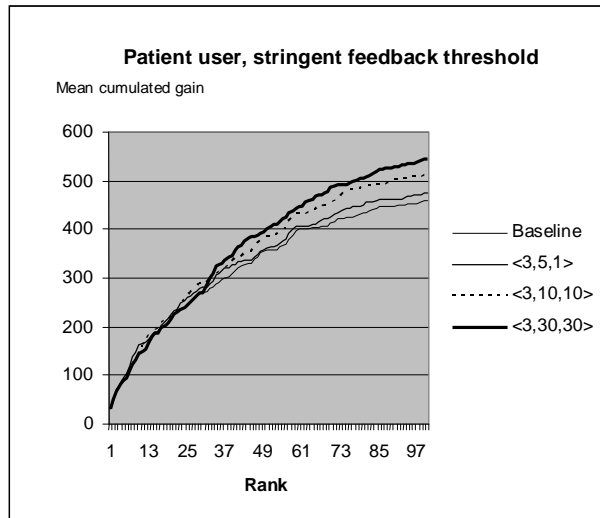
Figure 9. Cumulated gain results for a patient user keeping a stringent relevance feedback threshold. Baseline scenario and feedback scenarios (*<R = relevance feedback threshold, B = maximum browsing depth, F = maximum feedback set size>) <3,5,1>, <3,10,10>,* and *<3,30,30>*.

In all cases the RF significantly improves the final gain as compared to the baseline results. At the maximum rank 100 the baseline scenario reaches the cumulated gain value of 460; the scenario <3,10,10> reaches a value of 512; and the scenario <3,30,30> simulating a long initial browsing phase reaches the highest value of the scenarios compared – CG value of 544 (Table 3).

A summary of the gain results for all scenarios is presented in Table 3.

Table 3. Average CG over ranks (avg gain), and the final CG observed at the last rank (final gain). Results are averaged over 41 topics. Statistically significant differences to the baseline are marked (p ≤ 0.05 (*), p ≤ 0.01 (**), p ≤ 0.001 (***)). Three sets of evaluation scopes: at ranks 1-10, 1-20, and 1-100.

	avg gain			**final gain**		
	R=1	R=2	R=3	R=1	R=2	R=3
	@ ranks 1-10			@ last rank 10		
baseline	92	92	92	145	145	145
<R,1,1>	106***	103***	98	174***	164**	158**
<R,5,1>	108***	110***	99	172***	172***	163**
<R,5,5>	104*	104 **	98	176**	171**	160*
	@ ranks 1-20			@ last rank 20		
baseline	139	139	139	217	217	217
<R,10,5>	158**	155**	143	260***	258***	224
<R,10,10>	159**	156**	144	263***	255***	223
	@ ranks 1-100			@ last rank 100		
baseline	326	326	326	460	460	460
<R,5,1>	348**	347**	337	468	471	475 *
<R,10,10>	373***	375***	355	525***	526***	512**
<R,30,30>	379***	379***	368*	553***	550***	544***

Typically, the feedback scenarios perform significantly better than the baseline scenario. Yet when a high relevance feedback threshold is combined with a relatively long freezing scope and a relatively low last rank figure (as in the *final gain* of scenarios <3,10,5> and <3,10,10> measured at rank 20) statistically significant differences were not observed (Table 3).

In addition to the significant differences to the baseline presented in Table 3, significant pairwise differences were also noted between feedback scenarios themselves. The following differences were significant: using the *final* gain at rank 100: (<1,5,1>,<1,10,10>)***, (<1,5,1>,<1,30,30>)***, (<2,5,1>, <2,10,10>)**, (<2,5,1>, <2,30,30>)***, (<3,5,1>, <3,30,30>) ***, and (<3,10,10>, <3,30,30>)**; using *average gain* over ranks 1-10: (<1,5,5>,<1,5,1>)*, and over ranks 1-100: (<2,5,1>,<2,10,10>)*, and (<2,5,1>, <2,30,30>)*. The gain values of these scenarios are presented in Table 3.

As a general conclusion, more feedback is better in the long run, and mixed quality feedback always makes sense: the scenario <1,30,30> outperforms the scenario <1,5,1> despite the depressing effect of freezing at the first 30 ranks (Fig. 7), but it also fares well with scenarios <2,30,30> and <3,30,30> (Table 3, last row of the *final gain* columns).

Secondly, as the user's patience grows (as longer lists are evaluated) the relative position of stringent feedback improves. In case of a very patient user the stringent feedback fares almost as well as liberal. Yet it is currently an open question whether using the stringent RF threshold could outperform the liberal RF threshold, e.g., if we increase the browsing window size but use finer-grained values for the feedback set size, and simultaneously increase the last rank value used for measuring the final CG.

Table 4 compares the results of all RF scenarios to the baseline case topic-by-topic (41 topics altogether).

Table 4. Number of individual topics where the RF query performs either notably better, about equally as well as, or notably worse than the baseline query (N=41) measured by the final CG at last rank (final gain). Legend: notably better = greater than 105 % of the baseline result; about equal = from 95 % to 105 % of the baseline result; and notably worse = less than 95 % of the baseline result.

	final gain								
	notably better (> 105 %)			about equal (95 to 105 %)			notably worse (< 95 %)		
	R=1	R=2	R=3	R=1	R=2	R=3	R=1	R=2	R=3
	@ last rank 10								
<R,1,1>	11	8	4	30	33	37	0	0	0
<R,5,1>	13	11	5	25	28	36	3	2	0
<R,5,5>	11	11	4	28	27	37	2	3	0
	@ last rank 20								
<R,10,5>	19	16	8	17	21	29	5	4	4
<R,10,10>	19	16	8	17	21	29	5	4	4
	@ last rank 100								
<R,5,1>	9	8	6	26	29	33	6	4	2
<R,10,10>	16	15	9	21	24	30	4	2	2
<R,30,30>	19	18	14	20	21	25	2	2	2

For most scenarios only few individual topics perform worse than the baseline. Even modest feedback effort pays off if the evaluation is performed at low last rank, e.g., for the scenario <1,1,1> 11 topics perform notably better than the baseline at rank 10, while no topics perform notably worse. The situation is different when the evaluation rank is raised. For example, at rank 100, for the scenario <1, 5, 1> almost the same number of topics fail (6 topics) as outperform the baseline (9 topics). However, high original feedback effort pays off well. For example, for the scenario <1, 30, 30> only 2 topics failed notably, while for 19 topics the baseline was outperformed notably. The number of topics performing notably better than the baseline could not be raised by using a higher relevance feedback threshold – 18 and 14 topics performed notably better, respectively, for scenarios <2, 30, 30> and <3, 30, 30>. High relevance feedback threshold (R=3) leads to improved results when more feedback is given and a higher evaluation rank is used. Topic-by-topic results support the conclusion that mixed quality feedback generally makes sense. Especially if relatively small browsing window sizes are used to collect feedback documents, it may be difficult to find high quality feedback documents.

4. Discussion and Conclusions

Users of retrieval systems often still need long scanning of results in order to find the searched objects. One solution to this problem is to browse the initial result partially and give graded relevance feedback. The subsequent results would then have the remaining desired objects ranked better. In this paper we have performed a user model-based simulation of relevance feedback using graded relevance assessments, the freeze all method, and cumulated gain-based evaluation.

We used a simple user model allowing us to study the effects of users employing various thresholds to accept documents as feedback and manifesting various levels of patience in both browsing the initial result and giving feedback to the system. In the experimental part we investigated the browsing window size during the initial feedback up to rank 30, and the last rank inspected during the final evaluation up to rank 100. It makes sense for the user to give feedback on an initial relatively short browsing window in order to enrich the subsequent result up to his final examination length (and, consequently, our evaluation length) as much as possible.

Graded assessments allow experimentation with various weighting schemes. We modeled a user with a sharp weighting scheme 0-1-10-100, that is, a user giving high value (100) for the highly relevant documents compared to the documents of regular relevance (value 10) or documents of marginal relevance (value 1); non-relevant documents were given zero weight. We also modeled users with varying levels of patience in browsing the final result list. The method allows experimentation by varying attribute values, e.g., using flat weighting schemes (0-1-1-1) instead, or using additional value combinations <R, B, F> for modeling users. For example, we can see in Table 2 that it is not easy to find many highly relevant feedback documents even if the browsing window size is 30. Therefore, it would be interesting to run additional tests using finer-grained values of B (B∈{1, 2, 3, 4, 5}). We plan to do this in the future.

Cumulated gain-based evaluation and freezing all documents seen by the user allow our evaluation approach to put special emphasis on the user viewpoint. Cumulated gain is directly user-oriented in focusing on the *N* top-ranked documents. Our freezing method on the other hand is faithful to the original user point of view to the extreme – we freeze all seen documents into their ranked positions – even the non-relevant ones – because the user has wasted effort in inspecting them.

Our simulation showed several interesting results. First, despite freezing the ranks of the top documents seen by the simulated user, relevance feedback significantly improves the retrieval results measured by final cumulated gain at the last rank (the rank 10 in Figs. 1-3; the rank 20 in Figs. 4-6, and the rank 100 in Figs. 7-9) inspected for all user scenarios except for one case: when a moderately patient user sets unrealistic demands for the combined quality and quantity of the feedback documents (Fig. 6), significant improvements were not found. For impatient users, keeping low feedback threshold is generally most successful. It guarantees the maximum amount of feedback even if of mixed quality. For impatient users a high feedback threshold is discouraged as it may be difficult to find feedback at all.

As the second result, for very patient users it significantly pays off to give lots of feedback. When 30 documents are inspected for feedback, the results start to improve late due to long freezing zone (starting from rank 31), but the improvement curve is steep. The final cumulated gain result of scenario <1,30,30> at the last rank (rank 100) is both significantly and materially better than the result of both the baseline and the <1,5,1> scenarios. When the feedback threshold is raised at highly relevant documents, we notice significant difference also between scenarios <3,30,30> and <3,10,10>.

Interestingly, the final gain values at the last rank between scenarios <1,30,30> and <3,30,30> are close to each other (553 and 544, respectively) (Table 3). Thus, while the highly relevant feedback documents seem to be in a decisive role in order to improve the ranked result for a very patient user (the scenario <3,30,30>), it actually still makes sense to provide the maximum amount of mixed quality feedback, that is, to use a low relevance threshold in feedback (Figs. 7-9). However, we observed that the relative performance of stringent feedback improved as the result list in evaluation became longer. This suggests that for the laboratory IR type of long lists (up to 1000 documents) stringent RF might outperform liberal RF. This may also explain why our earlier study (Keskustalo et al. 2006) pointed to the profitability of stringent feedback threshold. In that study, the top 1000 documents were evaluated by using a system-oriented effectiveness measure (MAP) and freezing was not used because the effectiveness of the pseudo RF and the user model based RF were compared thus focusing on the quality of the final result only.

In the present paper, we have demonstrated a method to simulate and evaluate relevance feedback behavior from the user point of view. This entails user model-based simulation evaluated by using relatively short result lists and graded relevance assessments together with the "freeze all" approach regarding the initial results. We assumed a perfect user capable of making always correct relevance judgments during the RF phase. However, it is possible to extend the approach and model, e.g., errors made by the user in judging relevance, and their effects on performance. We will attempt next to extend the presented method for studying the performance of various RF techniques (Billerbeck 2005; Ruthven and Lalmas 2003), e.g., incremental RF (Aalbergsberg 1992).

References

Aalbersberg IJ (1992) Incremental relevance feedback. In: Belkin NJ, Ingwersen P and Mark Pejtersen A, eds., Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, pp. 11-22.

Belkin NJ, Cool C, Koenemann J, Ng KB, Park S (1995) Using Relevance Feedback and Ranking in Interactive Searching. TREC 1995. http://citeseer.ist.psu.edu/ belkin96using.html (visited August 14th, 2007).

Blair DC (1984) The Data-Document Distinction in Information Retrieval. Communications of the ACM, 4, Vol. 27, pp. 369-374.

Billerbeck B (2005) Efficient Query Expansion. Doctoral thesis. School of Computer Science and Information Technology, Portfolio of Science, Engineering and Technology, RMIT University. Melbourne, Victoria, Australia, 2005. http://goanna.cs.rmit.edu.au/~bodob/pubs/Bil05.pdf (visited August 14th, 2007).

Broglio J, Callan JP, Croft WB (1994) INQUERY system overview. In: Proceedings of the TIPSTER text program (Phase I), pp. 47-67.

Chang YK, Cirillo C, Razon J (1971) Evaluation of feedback retrieval using modified freezing, residual collection, and test and control groups. In: Salton G (Ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, Prencice-Hall, London, 1971. pp. 355-370.

Conover WJ (1999) Practical Nonparametric Statistics, 3rd edition. New York: John Wiley & Sons. 584 p.

Efthimiadis EN (1996) Query expansion. In: Williams ME, 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. http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html (visited August 14th, 2007).

Jordan C, Watters C, Gao Q (2006) Using Controlled Query Generation To Evaluate Blind Relevance Feedback Algorithms. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL'06), pp. 286-295. http://delivery.acm.org/

Järvelin K, Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: Belkin NJ, Ingwersen P, Leong M-K, eds., Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 41-48.

Kando N (2000) What Shall We Evaluate? - Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and patent Attorneys. Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval, Athens, Greece, July 28, 2000. http://research.nii.ac.jp/ntcir/ sigir2000ws/sigirprws-kando.pdf

Kekäläinen J (1999) The Effects of Query Compexity, Expansion and Structure on Retrieval Performance in Probalistic Text Retrieval. Doctoral thesis. Tampere, Finland: University of Tampere, Department of Information Studies. Acta Universitatis Tamperensis 678. 170 p.

Kekäläinen J, Järvelin K (2002) Using graded relevance assessments in IR evaluation. Journal of the American Society for Information Science and Technology, 53(13), pp. 1120-1129.

Keskustalo H, Järvelin K, Pirkola A (2006) The Effects of Relevance Feedback Quality and Quantity in Interactive Relevance Feedback: A Simulation Based on User Modeling. In: lalmas M, MacFarlane A, Rüger S, Tombros A, Tsirika T, Yavlinsky A, eds., Proceedings of the 28th European Conference on IR Research (ECIR), London, UK, pp. 191-204.

Marchionini G, Dwiggins S, Katz A, Lin X (1993) Information seeking in full-text end-user-oriented search systems: The roles of domain and search expertise. Library & Information Science Research 15(1), pp. 35-70.

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). http://informationr.net/ ir/7-2/paper127.html (visited August 15th, 2007).

Price SL, Nielsen ML, Delcambre LML, Vedsted P (2007) Semantic Components Enhance Retrieval of Domain-Specific Documents. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, Lisbon, Portugal, pp. 429-438.

Rocchio JJ, Jr (1971) Relevance feedback in information retrieval. In: Salton G (Ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, Prencice-Hall, London, 1971. pp. 313-323.

Ruthven I, Lalmas M (2003) A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review, 18(2): pp. 95-145.

Salton G (1989) Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Reading, MA: Addison-Wesley. 530 p.

Sormunen E (2002) Liberal relevance criteria of TREC – Counting on negligible documents? In: Beaulieu M, Baeza-Yates R, Myaeng SH, Järvelin K, eds., Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 320 – 330.

Sormunen E, Kekäläinen J, Koivisto J, Järvelin K (2001) Document Text Characteristics Affect Ranking of the Most Relevant Documents by Expanded Structured Queries. Journal of Documentation, 57(3), pp. 358-374.

Spink A, Saracevic T (1998) Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. Journal of the American Society for Information Science, 48(8): pp. 741-761.

Vakkari P, Sormunen E (2004) The Influence of Relevance levels on the Effectiveness of Interactive Information Retrieval. Journal for the American Society for Information Science and Technology 55(11), pp. 963-969.

Voorhees EM (2001) Evaluation by Highly Relevant Documents. In: Croft WB, Harper DJ, Kraft DH, Zobel J, eds., Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, USA, pp. 74-82.

White RW, Jose JM, van Rijsbergen CJ, Ruthven I (2004) A Simulated Study of Implicit Feedback Models. In: McDonald S, Tait J, eds., Proceedings of the 26th European Conference on IR Research (ECIR), Sunderland, UK, pp. 311-326.