Combining Interaction and Content for Feedback-Based Ranking

Emanuele Di Buccio[1], Massimo Melucci[1], and Dawei Song[2]

[1] University of Padua, Italy,
{dibuccio,melo}@dei.unipd.it
[2] The Robert Gordon University, UK,
d.song@rgu.ac.uk

Abstract. The paper is concerned with the design and the evaluation of the combination of user interaction and informative content features for implicit and pseudo feedback-based document re-ranking. The features are observed during the visit of the top-ranked documents returned in response to a query. Experiments on a TREC Web test collection have been carried out and the experimental results are illustrated. We report that the effectiveness of the combination of user interaction for implicit feedback depends on whether document re-ranking is on a single-user or a user-group basis. Moreover, the adoption of document re-ranking on a user-group basis can improve pseudo-relevance feedback by providing more effective document for expanding queries.

1 Introduction

Query-based search is the most widespread way to access information. Nevertheless, the intrinsic ambiguity of natural language, query brevity and the personal quality of information needs make query-based search insufficient to deal with every need. It should not come as a surprise that some research works investigate the use and the combination of other approaches to expanding queries by using relevance feedback.

The combination and the exploitation of post-search interaction features ("features" from now on) observed during the interaction between the user and an IR system is one of the most investigated approaches to expanding queries by using relevance feedback. Features are an inexpensive evidence and do not require user effort. They can be gathered by monitoring user behavior when interacting with the documents and may surrogate explicit user's relevance judgments. Hence, the relevance prediction power of features is the main issue.

The gap between what users perceive as relevant to the achievement of an information goal and what IR systems predict to be relevant suggests to exploit personalization at user-system interaction time [15]. However, post-search interaction features are often unavailable, insufficient or unnecessary, thus a broader definition can be useful or necessary. To this end, we consider the evidence gathered from groups of users with similar tasks or requests.

This paper is concerned with the combination of interaction and content-based feedback for document re-ranking. We exploit the features of the first documents visited by the user because the users steadily spend less time for searching, examine only the top-ranked results and therefore expect the relevant information in the top-ranked documents [5,6]. We propose a method for extracting behavioral patterns from the features and for representing both retrieved documents and user behavior. Users and groups are sources to distill features. User features are gathered during post-search navigation activity (e.g., when interacting with the results or the landing documents). Group features are distilled from the behavior of the group (e.g., the average dwell time spent on a page). The level of detail present in a set of features is named *granularity*. The specific representation adopted is based on the geometric framework originally proposed in [10]. The basic rationale is to model the behavior of the user when interacting with the first visited documents as a vector subspace. The behavioral patterns extracted from the features observed during the visit are basis vectors and the subspace is that spanned by a subset of these patterns. The subspace is adopted as a new dimension of the information need representation. Each document is represented as a vector of features. Then, documents are matched against the user behavior – the distance between the vector and the subspace is adopted to measure the degree to which a document satisfies the user behavior dimension. Documents are re-ranked according to this distance. In the paper behavior-based representations are adopted to re-rank documents uniquely using the user behavior dimension or to support query modification by extracting terms from the top documents re-ranked by user behavior.

In the paper the `P` label denotes features distilled from the individual user behavior, namely at personal granularity, while `G` will denote features at group granularity. In the latter case the value of a feature was obtained as the average computed over all the users that search for that topic other than the user under consideration. Since the adopted framework requires both a representation for the information need (the user behavior dimension) and one for the documents, and two are the possible feature granularities, that leads to four possible combinations `X/Y`, where `X` denotes the granularity of the user model and `Y` that for document representation — `X` or `Y` is either `P` or `G`.

2 Research Questions

This paper is mainly experimental and aims at addressing the following research questions:

1. When personal data is not available or insufficient, groups of users searching the same topic or performing the same task can be considered as another possible source for features. *What is the effect of the group data on document re-ranking when modeling user behavior and representing documents instead of personal data?* (Section 4.1)
2. Recent research activity has been investigating implicit indicators of relevance. The number of relevant documents affects relevance feedback. Simi-

larly, it affects implicit feedback. *What is the effect of the number of relevant documents among those used for user behavior dimension modeling on the effectiveness of document re-ranking?* (Section 4.2)

3. If the user behavior-based re-ranking were able to increase the number of good documents for feedback in the top-ranked, query expansion would benefit from re-ranking. Thus, the question is: *What is the effect of top-ranked document re-ordering by user behavior on query expansion?* (Section 4.3)
4. When considering the top-ranked document re-ranked by user behavior dimension as a source for query expansion, a further question is if user behavior-based query expansion is less sensitive to the number of relevant documents among the top-ranked than Pseudo Relevance Feedback (PRF). Thus, the research question is: *What is the effect of the number of relevant documents among the top-ranked on user behavior-based query expansion? Is it less sensitive than PRF?* (Section 4.4)

3 Evaluation

3.1 User study and test collection

The test collection adopted to address the research questions has been created through a user study. Fifteen volunteers have been recruited, particularly three undergraduate students, and twelve among PhD students or postdoctoral researchers. A set of topics was assigned to each user. The users have been asked to examine the top ten retrieved results in response to assigned topics and to assess their relevance with a four-graded scale. Explicit judgments have been gathered through a web application. Moreover, the application monitored the behavior of the user during the assessment, specifically collecting interaction features. The user study has resulted in an experimental dataset which contains content-based document features, interaction features, and explicit judgments of different users on the same document-topic pairs, for a set of topics. The remainder of this section provides a detailed description of the test collection and the experimental tool adopted in the user study, and the collected features.

Test collection for the user study. In the user study we have adopted the Ad-hoc TREC 2001 Web Track Test Collection. The corpus in this test collection is the WT10g, which is constituted by 1,692,096 documents (2.7 GB compressed, 11 GB uncompressed). The test collection includes fifty Ad-hoc topics together with the corresponding relevance judgments[3].

We are interested in comparing the behavior of diverse users when assessing the same topic. However, fifty topics are too many to be judged for each user. Hence, we have considered only a subset of the Ad-hoc topics. When preparing the dataset for the user study, a document has been considered relevant if it is assessed relevant by the TREC assessors. The number of the top-ten retrieved documents that are assessed relevant has been considered as an indicator of *topic*

[3] TREC 2001 Web Track data at `http://trec.nist.gov/data/t10.web.html`

Table 1. TREC 2001 Ad-Hoc web track topics divided according to the number of relevant documents in the top 10 retrieved.

Difficulty	Number of relevant documents	Topics
High	1-2	506-517-518-543-546
Medium	3-5	501-502-504-536-550
Low	6-10	509-510-511-544-549

Table 2. Topic sets, each one constituted by three topics for each set in Table 1.

	Difficulty		
Topic Set	High (1-2)	Medium (3-5)	Low (6-10)
A	506-517-518	501-502-504	509-510-511
B	517-518-543	502-504-536	510-511-544
C	518-543-546	504-536-550	511-544-549

difficulty. The topics without relevant documents, namely 534, 542, 513, 516, and 531, have been removed. The remaining topics have been divided by difficulty in three sets according to the number of relevant documents. Five topics have been randomly selected from each difficulty level, thus obtaining the fifteen topics reported in Table 1. Three distinct sets of nine queries have been built, each set being composed of three topics for each set of Table 1, thus achieving the Latin squares reported in Table 2. We have decided to distribute the topics so that at least one topic from each set would be assessed by all the users and the average topic difficulty was uniform per user. One of the three sets of topics, namely A, B, or C, was assigned to each user.

The use of a test collection is crucial since it allows us to simulate a realistic scenario in which the system is unaware of the real user's information need and it has to exploit the description provided by the topic. The document corpus has been indexed by the Indri Search Engine[4]. English stop-words have been removed and the Porter stemmer has been adopted. The documents of the WT10g have been ranked by Indri (default parameters were used) and the top 10 documents have been considered for each topic.

Experimental tool. We have developed a web application to collect the information about the user interaction behavior. The first web page presented to the user provided the list of topic identifiers. Once the user selected an assigned topic, a new web page divided in three frames was presented. The upper right frame reports the topic descriptions, i.e. title, narrative, and description; the left frame reports the title of the top ten retrieved documents ranked by Indri. A user reads the document in the bottom right frame; the user could access the documents in any order. A drop down menu allows the user to select the relevance degree

[4] `http://www.lemurproject.org/indri/`

Table 3. Features adopted to model the *user behavior* dimension and to represent documents.

Feature	Description
Features observed from document/browser window	
query terms	number of topic terms displayed in the title of the corresponding result
ddepth	depth of the browser window when examining the document
dwidth	width of the browser window when examining the document
doc-length	length of the document (number of terms)
Features observed from the user behavior	
display-time	time the user spent on the page in his first visit
scroll-down	number of actions to scroll down the document performed both by page-down and mouse scroll
scroll-up	number of actions to scroll up the document performed both by page-up and mouse scroll
sdepth	maximum depth of the page achieved by scrolling down, starting from the ddepth value

of the document. We have adopted the four graded relevance scale – (0) non relevant, (1) marginally, (2) fairly, (3) highly relevant – proposed in [7].

Features. Each action concerning the selection of the topic, the selection of a result and the relevance assessments have been centrally stored; both information about the type of action and the timestamp have been collected. Other features have been stored locally in the browser cookies. The features gathered from the user study are reported in Table 3. They can be divided in two groups: the features concerning the results or the displayed document, specifically the way in which they have been presented, and the features concerning user behavior. Document length was considered together with the display-time because a large display-time on a short document can have a different meaning than a display-time on a long document. The dimensions of the browser window have been considered together with the scrolling actions because different styles of scrolling interactions observed for diverse users can be also due to the different size of the browser window when visiting the same document with regard to the same query.

At the end of the evaluation session, the file with the cookies stored by the browser where the interaction data have been stored, was returned by each participant. Two users did not assess all the documents in the result list for some topics. For this reason, only the user behavior of thirteen among the fifteen users have been considered in this work, for a total of 79 (user,topic) pairs and 790 entries where each entry refers to the visit of a specific user to a particular document with regard to a topic.

3.2 Re-ranking Methods

User behavior-based re-ranking. A user visited n documents among the ten displayed in the result page returned in response to a query, the latter list (provided by Indri) being the *baseline*. Thus, for each query q and for each user u who searched using that query, namely for each pair (q,u), the following steps have been performed:

1. *Selection of the combination of the source for features.* Either `P/P`, `P/G`, `G/P` or `G/G` has been selected.
2. *Collection of the features from the first $n_B = 3$ visited documents.* The collected features are prepared in a $n_B \times k$ matrix where k is the number of features collected from the n_B visited documents. The reason for adopting the top visited documents and not the top ranked documents is to simulate a scenario where the first data obtained from interaction are adopted for feedback; the first visited results observed during the user study differ from the top ranked results. When considering the `G` source the value of a feature is obtained as the average computed over all the users that search for that topic other than the user under consideration.
3. *Modeling user behavior dimension by using patterns.* We applied Principal Component Analysis (PCA) [12] on the $n_B \times k$ matrix. The result of PCA is an orthonormal basis such that the basis vectors are 1:1 correspondence to the patterns. Thus, a pattern $\mathbf{p}$ is an eigenvector associated with non-null eigenvalue.
4. *Representation of the top-ranked documents.* Each document is represented as a vector $\mathbf{y}$ of the features reported in Table 3. Feature values are distilled from the source `Y` of the combination `X/Y` selected at step 1.
5. *Re-ranking of the top-ranked documents.* Each feature of $\mathbf{y}$ is used to retrieve the documents associated to the feature and ranked on the top of the initial list by the system. For all patterns, $|\mathbf{y}'\mathbf{p}|^2$ is computed using a document-at-a-time-like algorithm. The best performing pattern is manually selected.
6. *Effectiveness measurement.* The NDCG@n (for different n's) is computed for the new result list obtained after document re-ranking. DCG is computed according to the alternative formulation reported in [2], namely

$$DCG(i) = \sum_i (2^{r(i)} - 1)/\log(i+1),$$

   where $r(i)$ is the relevance of the document at position i. The normalization factor is the DCG of the perfect ranking. The gains adopted to compute NDCG are those provided by the user u when assessing the query q.

User behavior-based re-ranking to support query expansion. Besides the impact on document re-ranking, the effectiveness of user behavior to support query expansion is investigated. It is supposed that a first stage prediction has been performed based on Indri. For each query the following steps are performed:

1. Consider the top $n_B = 3$ documents retrieved by the baseline.

2. Perform step 1–5 described in the previous section by the `G/G` combination for user behavior dimension-based re-ranking using the $n_B = 3$ documents considered in the previous step. Consider the top $n_F = 5$ documents re-ranked by user behavior dimension.
3. Re-ranking of the top $m = 50$ documents returned by the baseline by using the PRF algorithm of Indri, adaptation of relevance models proposed in [9], on the $n_F = 5$ considered documents with $k = 10$ expansion terms.
4. Computation of the NDCG@n (for different n's) for the new result list obtained from the feedback at step 3.

This user-behavior based query expansion (IRF) is compared with the PRF algorithm of Indri on the top $n_F = 5$ documents retrieved by the baseline.

At step 2, the combination of sources of features adopted is `G/G`, that is the tests are performed in a non personalized scenario. Interaction features of all the users who searched using the query are adopted for dimension and document modeling for user behavior-based re-ranking. The dimension is automatically obtained using the first eigenvector among those extracted.

At step 3, when considering IRF the strategy is not actually PRF since we are using the top re-ranked by user behavior dimension.

Differently from the previous re-ranking method (user dimension-based re-ranking without query expansion), the gains adopted in Step 4 for the computation of the NDCG@n are those provided by the TREC assessors, that is those in the qrels of the TREC 2001 Web Track Test Collection. The basic idea is that the TREC assessor is considered as a new user, not among those in the group, who will be supported using group evidence. In other words this experiment aims at investigating if the pattern extracted by PCA could be useful for a non-personalized re-ranking.

4 Experimental Results

4.1 Question 1: Effect of group data on document re-ranking

The first research question concerns with the impact of the selection of the source combination on document re-ranking (i.e., `P/P` vs. `G/-` and `-/G`). In particular, we are interested in understanding if using group data both for modeling the user behavior and for representing documents negatively affects document re-ranking in comparison with exploiting the data distilled from the individual. Some preliminary results for this research question only for $n_B = 3$ have been reported in [3]; $n_B < 3$ has not been considered because the number of patterns for the diverse combinations was usually one and this pattern was not effective. Here we extend the analysis with varying n_B.

Table 4 reports the mean and the median NDCG@10 over all the (topic,user) pairs for all the combinations when $3 \leq n_B \leq 10$ and the best performing pattern is considered for each pair. Looking at the mean and the median NDCG@10, the results show that `P/P` and `G/G` benefit from additional evidence, yet NDCG@10 increased monotonically with the number of documents used as evidence for

Table 4. Comparison among median NDCG@10 of the diverse source combinations when varying the number of documents n_B used to obtain the user behavior dimension. Values marked by asterisks are those for which the difference with the P/P case was significant (one asterisk denotes $p < 0.05$, two asterisks $p < 0.01$) according to the one-sided Wilcoxon signed ranked test based on the alternative hypothesis that P/P combination performed better than the group-based combinations.

	Personal	Group			Increment (%)		
n_B	P/P	P/G	G/P	G/G	$\Delta_{\mathtt{P/G}}$	$\Delta_{\mathtt{G/P}}$	$\Delta_{\mathtt{G/G}}$
3	0.817	0.832	0.799	0.825	1.748	-2.238	0.922
4	0.839	0.825	0.805*	0.827	-1.615	-4.056	-1.324
5	0.833	0.835	0.826	0.835	0.288	-0.817	0.288
6	0.839	0.843	0.833	0.839	0.524	-0.656	0.045
7	0.847	0.840	0.831	0.848	-0.798	-1.829	0.137
8	0.841	0.832	0.839**	0.835	-1.164	-0.333	-0.701
9	0.847	0.835	0.839**	0.838	-1.351	-0.985	-1.109
10	0.853	0.835	0.839**	0.832	-2.049	-1.686	-2.506

none of the combinations. The results obtained from the different combinations are comparable. The only significant difference is observed for the G/P case, whose poor performance could be due both to the fact the a non-personalized dimension is adopted and the comparison is performed between a dimension and a document representation obtained from diverse sources for interaction features. Because of its lack of effectiveness, in the remainder of this work this combination will be no longer considered.

Moreover, results in Table 4 shows that, even if the re-ranking effectiveness is comparable, the adoption of group data can negatively affect re-ranking. Indeed, even if the G/G combination seems to be promising for $n_B = 3$, when considering the results per topic and per user they show that also in this case the adoption of group data can affect effectiveness of re-ranking. This is the case, for instance, of the topics 536, 543 and 550 where all the three combinations involving group data performed worse than the P/P case.

The above remarks concerned with the comparison among the diverse combinations thus investigating the effect of using group data instead of personal data for personalized user behavior-based re-ranking. But no comparison was performed with the baseline B. Table 5 reports the median NDCG@10 for the baseline and the diverse combinations for different values of n_B. None of the combinations significantly outperformed the baseline ranking. Also when comparing results per topic and per user none of the combinations outperformed the baseline for all the topics or all the users.

The relatively small number of experimental records is definitely a limitation since small numbers make the detection of significant differences harder than the detection based on large datasets. But, the small size of the dataset allows us to note that the two sources of features (i.e. P and G) adopted seem to provide diverse contributions. For instance, for topics 518, 536, and 546 only one of the

Table 5. Comparison among median NDCG@10 of the baseline and the diverse source combinations when varying the number of documents n_B used to obtain the user behavior dimension.

	Baseline	Source combinations			Increment (%)		
n_B	B	P/P	P/G	G/G	$\Delta_{\texttt{P/P}-\texttt{B}}$	$\Delta_{\texttt{P/G}-\texttt{B}}$	$\Delta_{\texttt{G/G}-\texttt{B}}$
3	0.838	0.817	0.832	0.825	-2.462	-0.757	-1.563
4	0.838	0.839	0.825	0.827	0.053	-1.563	-1.272
5	0.838	0.833	0.835	0.835	-0.604	-0.317	-0.317
6	0.838	0.839	0.843	0.839	0.053	0.577	0.098
7	0.838	0.847	0.840	0.848	1.048	0.242	1.187
8	0.838	0.841	0.832	0.835	0.387	-0.781	-0.317
9	0.838	0.847	0.835	0.838	1.048	-0.317	-0.072
10	0.838	0.853	0.835	0.832	1.769	-0.317	-0.781

two combinations performs better than the baseline. This suggests to investigate combinations of the diverse feature granularities.

4.2 Question 2: Effect of the number of relevant documents on document re-ranking

The representation of the user behavior exploits the data gathered from the first visited documents by the users, extracts possible patterns (i.e. eigenvectors of the correlation matrix) from those data and uses the most effective pattern for re-ranking. If the visited documents are relevant, it is necessary to investigate whether the improvement in terms of effectiveness can mainly be due to the ability of the user to select relevant documents. To this end, we investigated the relationship between the number of relevant documents among the top $n_B = 3$ visited and NDCG@10 across the diverse combinations.

In Figure 1 the results are depicted. The NDCG@10's measured for the baseline (`Indri`) when considering all the users and all the topics is plotted against the number of relevant documents in the top three visited — the regression lines are reported for providing an idea of the trend. The least steep lines refer to `P/P`, `P/G` and `G/G`. For the diverse combinations the dependence with the number of relevant documents among the top three visited is still linear, but the slope decreases and the intercept increases.

The mean and the median NDCG@10 was higher than the baseline when only one relevant document was present among those used for feedback, but this increment decreased when increasing the number of relevant documents; the same results have been observed for $n_B \in \{4, 5\}$. The main limitation is the robustness of the adopted approach. Indeed, when observing the variance of NDCG@10 values, in the event of one relevant document the variance is smaller than those obtained for the baseline; differently, when the number of relevant documents increases, the baseline has smaller variance, thus suggesting that even

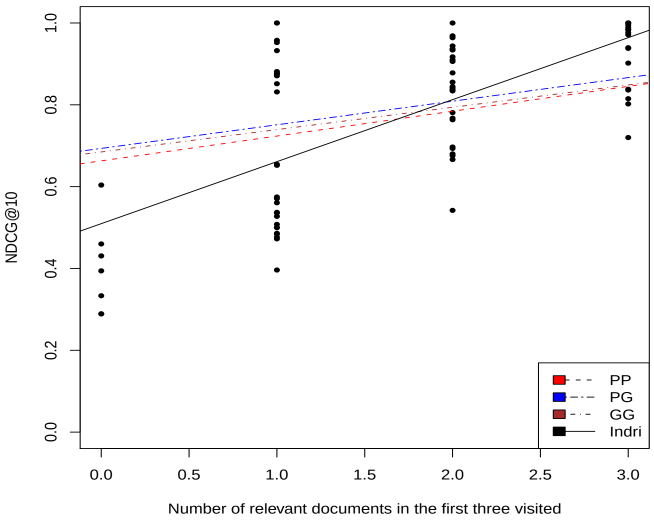

Fig. 1. Comparison among the regression line of the baseline(`Indri`) and those of the diverse combinations (`X/Y`'s).

if the user behavior based re-ranking can provide some improvement, the latter is less robust than the baseline.

4.3 Question 3: Effect of user behavior-based document re-ranking on query expansion

The results reported above showed that the improvement in terms of retrieval effectiveness is not consistent throughout all the topics or all the users, but the effectiveness of the top ten document re-ranking seems to be not strictly dependent from the relevance of the document used to model the user behavior. For this reason we investigated if the top n_F documents re-ranked by the user behavior are a more effective evidence for query expansion than the top n_F retrieved by the baseline. The basic idea is to investigate if user behavior-based re-ranking is able to bring at high rank positions good sources for query expansion, thus improving the effectiveness respect to PRF where the top ranked documents retrieved by the baseline are supposed to be good sources for feedback.

Table 6 reports the mean and the median NDCG@n's computed over the different values of the parameters (k, n_F) adopted, that is for diverse number of expansion terms and feedback documents; the number of documents used to model the user behavior dimension was fixed to $n_B = 3$, since it provided good results also for group-based source combinations. The results show that query expansion can benefit from user behavior based re-ranking, even if the improve-

Table 6. Median and Mean NDCG@n, with $n \in \{10, 20, 30, 50\}$, computed over all the values of the parameters k and n_F, where k denotes the number of expansion terms and varies in $\{5, 10, 15, 20, 25, 30\}$, and n_F denotes the number of feedback documents and varies in $\{1, 2, 3, 4, 5\}$. Results refer to the case where $n_B = 3$.

	NDCG@10			Δ (%)		NDCG@20			Δ (%)	
	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$
median	0.324	0.319	0.341	-1.31	5.44	0.295	0.286	0.311	-3.15	5.23
mean	0.329	0.318	0.350	-3.45	6.38	0.295	0.290	0.312	-1.69	5.60
	NDCG@30			Δ (%)		NDCG@50			Δ (%)	
	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$
median	0.249	0.281	0.293	12.86	17.78	0.208	0.207	0.220	-0.48	5.80
mean	0.288	0.285	0.303	-0.95	5.05	0.228	0.217	0.225	-5.10	-1.35

ments are modest: except for NDCG@30, the improvement is approximately 5%. Table 7 reports the NDCG@n's for different values of k, namely the number of expansion terms, and different n_F's, namely the number of feedback documents. The results show that the adopted approach, Implicit Relevance Feedback (IRF), benefits from a small number of feedback documents, $n_F = \{1, 2\}$, and an increment of the number of terms used for query expansion, i.e. $k = \{10, 20\}$. For most of the parameters pairs (k, n_F) IRF can improve PRF. But this specific approach should be improved since it is not robust. Let us consider, for instance, the case for $k = 10$ and $n_F = 2$, where PRF did not improve the baseline ($\Delta_{\texttt{PRF-B}} = 0.24\%$) differently from IRF ($\Delta_{\texttt{IRF-B}} = 8.67\%$), and the difference in terms of NDCG@10 was greater than 5%. Table 8 reports the results for each topic and shows that also in this case, IRF is not able to outperform PRF for all the topics.

The user behavior-based re-ranking is able to provide an improvement respect to PRF, increasing the number of good sources for feedback at high rank positions and supporting feedback when small evidence is adopted as input, e.g. one or two documents. As shown in Table 7, for 39/45 cases IRF performed as good as or better than PRF, and for 26/45 the increment in terms of NDCG was higher than 5%. But an analysis of the effectiveness per topic shows that a more robust approach is required since for some topics PRF performed better than IRF.

4.4 Question 4: Effect of the number of relevant documents used for dimension modeling on query expansion

The objective of the research question discussed in Section 4.3 was to investigate the capability of the user behavior dimension to increase the number of good sources, namely documents, for query expansion at high rank position, thus increasing the effectiveness of Pseudo Relevance Feedback. The results showed that PRF can benefit from a preliminary user behavior based re-ranking. In order to

Table 7. Comparison among the NDCG@n's of the baseline (B), PRF and IRF for different values of n, n_F and k. The results in bold type are those for which the increment respect to the baseline B is higher than 5%. The results marked by an asterisk are those for which the increment of IRF respect to PRF in terms of NDCG@n, $\Delta_{\texttt{IRF-PRF}}$, is greater than 5%; those marked by two asterisks are those for which $\Delta_{\texttt{IRF-PRF}} > 10\%$.

		NDCG@10			NDCG@20			NDCG@30		
k	n_F	B	PRF	IRF	B	PRF	IRF	B	PRF	IRF
5	1	0.329	0.341	0.343	0.295	0.301	0.308	0.288	0.290	0.296
	2	0.329	**0.348**	0.313	0.295	0.309	0.287	0.288	0.296	0.282
	3	0.329	0.302	0.320*	0.295	0.287	0.293	0.288	0.283	0.289
	4	0.329	0.340	0.331	0.295	0.305	0.297	0.288	0.295	0.293
	5	0.329	0.299	0.312	0.295	0.278	0.292*	0.288	0.277	0.284
10	1	0.329	**0.351**	**0.370***	0.295	**0.313**	**0.326**	0.288	0.298	**0.309**
	2	0.329	0.330	**0.357****	0.295	0.291	**0.317***	0.288	0.285	**0.305***
	3	0.329	0.292	0.342**	0.295	0.283	0.303*	0.288	0.280	0.300*
	4	0.329	0.286	0.338*	0.295	0.269	0.301**	0.288	0.273	0.300*
	5	0.329	0.311	0.324	0.295	0.283	0.293	0.288	0.278	0.295*
20	1	0.329	**0.371**	**0.378**	0.295	**0.313**	**0.331***	0.288	**0.305**	**0.319**
	2	0.329	0.328	**0.376****	0.295	0.292	**0.324****	0.288	0.286	**0.311***
	3	0.329	0.280	0.341**	0.295	0.277	**0.315****	0.288	0.278	0.299*
	4	0.329	0.299	0.334**	0.295	0.274	0.309**	0.288	0.280	0.298*
	5	0.329	0.288	0.343*	0.295	0.273	0.299*	0.288	0.277	0.298*

Table 8. NDCG@$\{10, 20, 30\}$'s per topic for IRF and PRF when $k = 10$ and $n_F = 2$.

	NDCG@10			NDCG@20			NDCG@30		
Topic	PRF	IRF	Δ(%)	PRF	IRF	Δ(%)	PRF	IRF	Δ(%)
501	0.386	0.432	12.02	0.400	0.497	24.22	0.408	0.505	23.82
502	0.300	0.454	51.62	0.219	0.318	45.68	0.199	0.276	38.62
504	0.393	0.333	-15.10	0.316	0.289	-8.52	0.372	0.308	-17.11
506	0.125	0.108	-13.84	0.125	0.108	-13.84	0.125	0.108	-13.84
509	0.518	0.518	0.00	0.478	0.478	0.00	0.466	0.466	0.00
510	0.697	0.697	0.00	0.450	0.450	0.00	0.393	0.393	0.00
511	0.310	0.323	4.36	0.337	0.383	13.57	0.370	0.395	6.62
517	0.140	0.155	10.81	0.154	0.121	-21.40	0.157	0.139	-11.43
518	0.000	0.000	-	0.102	0.070	-31.96	0.101	0.100	-1.19
536	0.355	0.426	20.12	0.269	0.323	20.13	0.291	0.345	18.68
543	0.064	0.078	23.27	0.041	0.051	23.11	0.037	0.045	23.29
544	0.673	0.700	4.10	0.670	0.737	9.97	0.633	0.675	6.67
546	0.169	0.240	42.03	0.188	0.266	41.48	0.190	0.242	27.37
550	0.489	0.539	10.25	0.328	0.348	6.04	0.252	0.267	6.03
all	0.330	0.357	8.40	0.291	0.317	8.86	0.285	0.305	6.80

Table 9. Median NDCG@$\{10, 20, 30\}$'s for different numbers of relevant documents among the top three documents of the baseline, when considering $n_F = 3$. In the event of PRF, this number corresponds to the number of relevant documents among those used for feedback. In the event of IRF, this number corresponds to the number of relevant documents among those used for modeling the user behavior dimension.

Relevant	NDCG@10			NDCG@20			NDCG@30		
in Top 3	PRF	IRF	Δ(%)	PRF	IRF	Δ(%)	PRF	IRF	Δ(%)
0	0.074	0.069	-6.09	0.067	0.045	-33.13	0.075	0.076	1.27
1	0.221	0.252	13.95	0.251	0.288	14.84	0.240	0.256	6.72
2	0.342	0.454	32.71	0.367	0.387	5.28	0.341	0.406	19.00
3	0.514	0.514	0.11	0.475	0.475	-0.09	0.472	0.497	5.24

gain more insights into the user behavior dimension capability to support query expansion, we investigated the effect of the number of relevant documents among the top n_B. The objective is to understand if, also when there is a small number of relevant documents among the top n_B, actually those adopted to model the dimension, user behavior-based re-ranking is able to improve the effectiveness of the system in ranking highly relevant documents at high rank positions.

Table 9 reports the NDCG@n's for different values of n and different numbers of relevant documents among the top three documents of the baseline, when considering $n_F = 3$. In the event of PRF, this number corresponds to the number of relevant documents among those used for feedback. In the event of IRF, this number corresponds to the number of relevant documents among those used for modeling the user behavior dimension. When there are no relevant documents among the top 3 of the baseline the effectiveness of feedback is low and PRF performs better; this results suggests that when no relevant documents are adopted for dimension modeling the effectiveness of the model is negatively affected. Differently when only one or two relevant documents are present in the top 3 used for pseudo-feedback (PRF) or dimension modeling (IRF), IRF outperforms PRF thus suggesting that is able to improve the number of good sources for content-based feedback in the top 3. When the number of relevant documents is three, namely all the feedback documents are relevant, the two approaches perform equally.

Tables 10 reports the median NDCG@n's for the two feedback strategies when compared to the baseline. IRF was able to provide a positive contribution when one or two relevant documents are present in the top 3, but both the feedback techniques hurt the initial ranking when the top three documents are relevant.

5 Related Work

A review on past works investigating implicit indicators and feedback techniques is reported in [8]. In that work individual and group granularities referred to two

Table 10. Median NDCG@$\{10, 20\}$'s (Table 10a) and NDCG@30's (Table 10b) for different numbers of relevant documents among the top 3 of the baseline. Results are reported for the baseline and the two feedback strategies IRF and PRF.

Relevant	NDCG@10					NDCG@20				
in Top 3	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$
0	0.069	0.074	0.069	6.48	-6.09	0.045	0.067	0.045	49.55	0.00
1	0.166	0.221	0.252	33.04	13.95	0.187	0.251	0.288	33.91	53.78
2	0.437	0.342	0.454	-21.78	32.71	0.382	0.367	0.387	-3.82	1.26
3	0.625	0.514	0.514	-17.88	0.11	0.596	0.475	0.475	-20.28	-20.35

(a)

Relevant	NDCG@30				
in Top 3	B	PRF	IRF	$\Delta_{\texttt{PRF-B}}$	$\Delta_{\texttt{IRF-B}}$
0	0.040	0.075	0.076	87.56	89.95
1	0.162	0.240	0.256	48.65	58.65
2	0.381	0.341	0.406	-10.46	6.56
3	0.591	0.472	0.497	-20.12	-15.93

(b)

distinct dimensions for classification: individual's and group granularity levels refer to explicit judgments that the implicit feedback strategy should predict.

Collaborative filtering, for instance, exploits group ratings gathered by similar users to predict the user interests. An application to web search that involves interaction data at group granularity is [14] where the author investigated the predicting effectiveness of click-through data gathered from users in a community, e.g. a interest-specific web portal. In [13] diverse grouping criteria are investigated for tag recommendation in a social network scenario. Users are grouped according to explicit connections with other members or according to the subscription to interest groups. Tag occurrence and co-occurrence information at personal and group levels are adopted to estimate the probability for ranking tags to suggest. Even if the work on collaborative filtering and recommendation is related, this paper concentrates on IRF and PRF as well as on the "tension" between users with the same topic in mind. The work reported in [16] is also related but it is focused on different criteria for group creation and the proposed *groupization* algorithm consists in aggregating personalized scores.

In regard to implicit indicators granularities, in [4] the authors investigated combination of implicit indicators by Bayesian modeling techniques. Two feature granularities have been considered: result-level features which referred to individual pages visited by the user for a specific query, and session-level features whose value, when the features are not session specific, was obtained as the average value of result-level features computed over all the queries in the session. In [1] group granularity interaction feature values are adopted together with derived features to learn user models. The value of a derived feature was obtained subtracting the feature background distribution value from the observed value: the assumption is that the relevance component of a feature can be obtained by considering its deviation from its background distribution. The value for a feature at group granularity was obtained as the average value computed across all the users and search sessions for each query-URL pair.

Also in this paper group granularity features are obtained as the average computed over all the users for a specific topic, but not including the user the IR system aims at supporting. Differently from [1] this paper is focused on the capability of indicators to support personalization and our approach is based on a different hypothesis: relevance information can be extracted from the correlation among the observed indicators. The approach adopted is that proposed in [11] where PCA was used to extract behavioral patterns, then used for document reranking; re-ranked documents are then adopted as source for query expansion. Differently from [11] this work investigates the impact of source combinations on the re-ranking effectiveness, and the effectiveness of group behavior document re-ranking to support query expansion in a non-personalized search task.

6 Conclusion

The results reported in this paper show that the contributions of the diverse sources are comparable, thus making personalized IRF feasible despite the data sparsity observed when the interaction features are collected on a per-user basis. Another finding was that the diverse source combinations `X/Y`'s provide complementary contributions with respect to the baseline for some topics and users, thus suggesting to investigate source combinations when both of them are available. Moreover, the effectiveness of query expansion based on the highest ranking results re-ranked by group behavior was investigated in a non-personalized search task. The results also show that the highest ranking results re-ranked by group behavior are comparable with the highest ranking results in the baseline list when used for query expansion, thus suggesting to investigate combinations both of content and user behavior as evidence to support query expansion at a larger scale than the study of this paper to see whether they can effectively complement each other or whether they instead tend to cancel each other out.

Acknowledgements

The work reported in this paper has been supported by the PROMISE network of excellence (contract n. 258191) project and by the QONTEXT project (grant agreement n. 247590), as a part of the 7th Framework Program of the European commission.

References

1. Agichtein, E., Brill, E., Dumais, S., Ragno, R.: Learning user interaction models for predicting web search result preferences. In: Proceedings of SIGIR '06. pp. 3–10. ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1148170.1148175
2. Croft, B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company, USA, 1st edn. (2009)
3. Di Buccio, E., Melucci, M., Song, D.: Exploring Combinations of Sources for Interaction Features for Document Re-ranking. In: Proceedings of HCIR 2010. pp. 63–66. New Brunswick, NJ, USA (2010)
4. Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM Transactions on Information Systems 23, 147–168 (April 2005), http://doi.acm.org/10.1145/1059981.1059982
5. Jansen, B.J., Spink, A.: An analysis of web searching by european alltheweb.com users. Information Processing and Management 41, 361–381 (March 2005), http://portal.acm.org/citation.cfm?id=1055766.1055778
6. Jansen, B.J., Spink, A.: How are we searching the world wide web?: a comparison of nine search engine transaction logs. Information Processing and Management 42, 248–263 (January 2006), http://dx.doi.org/10.1016/j.ipm.2004.10.007
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20, 422–446 (October 2002), http://doi.acm.org/10.1145/582415.582418
8. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37, 18–28 (September 2003), http://doi.acm.org/10.1145/959258.959260
9. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of SIGIR '01. pp. 120–127. ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/383952.383972
10. Melucci, M.: A basis for information retrieval in context. ACM Transaction on Information Systems 26, 14:1–14:41 (June 2008), http://doi.acm.org/10.1145/1361684.1361687
11. Melucci, M., White, R.W.: Utilizing a geometry of context for enhanced implicit feedback. In: Proceedings of CIKM '07. pp. 273–282. ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1321440.1321480
12. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2, 559–572 (1901)
13. Rae, A., Sigurbjörnsson, B., van Zwol, R.: Improving tag recommendation using social networks. In: Proceedings of RIAO '10. pp. 92–99. Paris, France, France (2010), http://portal.acm.org/citation.cfm?id=1937055.1937077
14. Smyth, B.: A community-based approach to personalizing web search. Computer 40, 42–50 (August 2007), http://portal.acm.org/citation.cfm?id=1300755.1301819
15. Teevan, J., Dumais, S.T., Horvitz, E.: Potential for personalization. ACM Transactions on Computer-Human Interaction 17, 4:1–4:31 (April 2010), http://doi.acm.org/10.1145/1721831.1721835
16. Teevan, J., Morris, M.R., Bush, S.: Discovering and using groups to improve personalized search. In: Proceedings of WSDM '09. pp. 15–24. ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1498759.1498786