Flexible Pseudo-Relevance Feedback for NTCIR-2

Tetsuya Sakai[∗⋆] Stephen E. Robertson[†‡] Stephen Walker[†]

∗Computer Laboratory, University of Cambridge, UK
⋆Knowledge Media Laboratory, Toshiba Corporate R&D Center, Japan
†Microsoft Research Ltd, Cambridge, UK
‡Department of Information Science, City University, London, UK
tetsuya.sakai@{cl.cam.ac.uk,toshiba.co.jp}

Abstract

The University of Cambridge/Microsoft/Toshiba team participated in the NTCIR-2 Japanese-English Cross-Language IR task only, using the Okapi Basic Search System at the University of Cambridge. The aim of our participation this year was to improve the reliability of pseudo-relevance feedback *(PRF) rather than to achieve a high cross-language performance. We therefore regarded the cross-language task as a monolingual one, after having translated the Japanese search requests into English using a machine translation system that was not tuned in any way for the NTCIR-2 task. Our* flexible pseudo-relevance feedback *(FPRF) strategies attempt to estimate the best PRF parameter values for each test request based on results with a set of training requests. Seven* automatic *runs were submitted using the DESCRIPTION fields: one FPRF run which varies the number of pseudo-relevant documents across requests, four FPRF runs which vary the number of expansion terms across requests, and traditional runs with and without PRF. This paper reports on our pre-submission and post-submission FPRF experiments.*

Keywords: *Pseudo-relevance feedback, local feedback, query expansion.*

1 Introduction

The University of Cambridge/Microsoft/Toshiba team participated in the NTCIR-2 Japanese-English Cross-Language IR task only. The aim of our participation this year was to improve the reliability of *pseudo-relevance feedback* (PRF) rather than to achieve a high cross-language performance. We therefore regarded the cross-language task as a monolingual one, after having translated the Japanese search requests into English using Toshiba's machine translation service called MT Avenue (`http://mtave.softpark.jplaza.com/MTave`) [2], which was not tuned in any way for the NTCIR-2 task. Our *flexible pseudo-relevance feedback* (FPRF) strategies attempt to estimate the best PRF parameter values for each test request based on results with a set of training requests [7]. Seven *automatic* runs were submitted using the DESCRIPTION fields: one FPRF run which varies the *number of pseudo-Relevant documents* (R) across requests, four FPRF runs which vary the *number of expansion Terms* (T) across requests, and traditional runs with and without PRF.

(Although NII also released a set of manual English translations of the Japanese search requests, we did not use them to generate any English monolingual baselines. This is not only because we are interested in comparing FPRF with traditional PRF rather than filling the cross-language gap, but also because we found that some of the translations were not quite accurate. These English translations should perhaps be treated with care, since previous work suggests that comparing a cross-language performance against a monolingual baseline is heavily dependent on the manual request translation process even if the translations *are* accurate [8].)

In all of our training experiments, we used the 60 NTCIR-1 cross-lingual requests with the *entire* NTCIR-2 English document collection that consists of 322,058 documents. Therefore, unlike a TREC *ad hoc* task, the relevance data used in the training phase are *incomplete* in that they only cover a subcollection that consists of 187,080 documents (i.e. the NTCIR-1 English document collection).

All of our experiments used a family of shell scripts that utilize the Okapi Basic Search System with the BM25 probabilistic retrieval formula [4][5]. Our basic PRF algorithm involves *term reweighting* based on the Robertson/Sparck Jones *relevance weight* and *query expansion* based on the *offer weight*, as described fully in [3][10]. The actual retrieval steps, also used in [6][8], are as follows:

1. Remove "Document will discuss" and the like

from the search request and generate an initial query using a stopword list and stemming.

2. Perform initial (or pilot) search.

3. Assume the top R documents in the initial ranked output relevant and extract T new search terms from them, using offer weights as term selection values.

4. Reweight the initial search terms with relevance weights. To this reweighted initial query, add the new search terms with *downweighted* relevance weights, using a downweighting factor $w(\leq 1)$.

5. Perform final search using the expanded query.

While traditional PRF uses the same values of R and T for *all* requests, FPRF attempts to estimate the best value of R and/or T for each given test request.

The remainder of this paper is organized as follows. **Section 2** describes our FPRF strategies in detail. **Section 3** describes our *pre-submission* experiments for generating the official runs, in which we mistakenly used noninterpolated average precision over *retrieved* relevant documents instead of that over *all* relevant documents (i.e. TREC average precision). Therefore, **Section 4** repeats some of our experiments based on TREC average precision using Chris Buckley's evaluation program. These new experiments also cover a few FPRF strategies that were not considered before submission. Finally, **Section 5** gives conclusions and directions for future research.

2 Flexible Pseudo-Relevance Feedback

2.1 Direct Mapping and Categorization

Although PRF is widely used in laboratory experiments and in TREC [1], traditional PRF is probably too unreliable for practical application since it is known to hurt performance for approximately one-third of a given set of search requests even when improving the *average* performance [7].

To enhance the reliability of PRF, Sakai *et al.* [7] recently proposed some methods for estimating the best PRF parameter values for each given test request: Firstly, the PRF parameters are optimized for *each* training request. Secondly, the *distance* $\mathbf{d}$ between the test request and each training request is calculated. Then, in the case of *direct mapping* of requests, the optimal parameter values for the training request that minimizes $\mathbf{d}$ are re-used for the test request. That is, the PRF parameters for a test request are selected based on a "similar request that the system has seen before". In contrast, in the case of *categorization* of requests, the training requests are first categorized into several *groups*, so that each group contains requests with the same optimal PRF parameter values. Then, the distance $\mathbf{D}$ between a given test request and each group is derived from individual $\mathbf{d}$'s, for example, by taking the average. Finally, the optimal parameter values for the group that minimizes $\mathbf{D}$ are used for the test request. Relying on the request-to-group distance ($\mathbf{D}$) instead of the request-to-request distance ($\mathbf{d}$) may possibly prevent "outliers" in the set of training requests from doing harm to the test-to-training mapping. In this paper, these approaches for per-request estimation of the best PRF parameter values will collectively be referred to as *flexible pseudo-relevance feedback* (FPRF).

2.2 Distance Measures for FPRF

The critical factor of FPRF (i.e. direct mapping and categorization) is how to measure the request-to-request distance $\mathbf{d}$. This section discusses some possibilities for measuring it, most of which were introduced in [7]. All related equations can be found in the **Appendix**.

We considered three types of distance measures: **Section 2.2.1**, **2.2.2** and **2.2.3** describe those that rely on the document scores in the initial ranked ouput, the offer weights of candidate expansion terms, and the weights of the initial search terms, respectively. Of these distance measures, those based on the offer weight (**Section 2.2.2**) or the relevance weight (**HIRW**, **LIRW** and **IRWC** in **Section 2.2.3**) can only be used for estimating the best value of T, but not R, since these weights can only be obtained after determining the set of pseudo-relevant documents [3][10].

Of the following, **ICWC**, **IRWC** and the distance measures based on the relevance probability (**RPC**, **ARP**, **HRP**) were examined in our post-submission experiments only.

2.2.1 Distance Measures based on Initial Document Scores

DSC, NDSC A *document score curve* (DSC) can be obtained by plotting the initial document scores against document rank. If the DSCs of two requests are similar, this may possibly imply that a PRF parameter value that is effective for one may also be effective for the other. Moreover, it is possible to normalize each DSC before comparison so that its shape will affect the similarity, but not its height. Equations 2 and 3 define distances based on the (**N**ormalized) **D**ocument **S**core **C**urve.

RDSC Another possible way to normalize a DSC is to re-use a linear formula proposed by Robertson and Walker for the TREC-8 filtering track, which is based on logistic regression [4]. Their normalization method aims at making the document scores comparable across requests based on the average score of initially top ranked documents, the theoretical maximum document score, and the query length. Equation 8 defines a distance based on the **R**obertson/Walker normalized **D**ocument **S**core **C**urve. The **Appendix** also describes some variations of this distance measure, such as **RDSC**$'$ and **RDSC**$''$.

RPC Yet another possible way to make document scores comparable across requests is to re-use a method proposed by Robertson and Walker for the TREC-7/TREC-9 filtering tracks [5], which attempts to convert the scores into actual probabilities of relevance. This method first computes calibrated document scores using the average score of initially top ranked documents and a pair of constants, and then converts them to estimated relevance probabilities. Equation 11 defines a distance based on the Robertson/Walker **R**elevance **P**robability **C**urve.

ADS, ANDS, ARDS, ARP In contrast to comparing document scores at each rank, it is possible to take the average score of some top ranked documents first and then make comparisons across requests. Equations 12–15 define distances based on the **A**verage (**N**ormalized) **D**ocument **S**core, the **A**verage **R**obertson/Walker normalized **D**ocument **S**core and the **A**verage **R**elevance **P**robability, respectively.

HDS, HNDS, HRDS, HRP Rather than examining document scores at several ranks as in the above approaches, examining the highest document score may possibly be sufficient for determining the request-to-request distance, since it is possible to hypothesize that this value is correlated with *search request complexity* and with how successful PRF would be if applied [7]. Equations 16–19 define distance measures based on the **H**ighest (**N**ormalized) **D**ocument **S**core, the **H**ighest **R**obertson/Walker normalized **D**ocument **S**core and the **H**ighest **R**elevance **P**robability, respectively.

2.2.2 Distance Measures based on Offer Weights

OWC, NOWC Once the set of pseudo-relevant documents (and therefore the value of R) has been determined, a *term selection value curve* can be obtained by plotting the term selection values of candidate expansion terms in decreasing order. In our current probabilistic framework, this is the *offer weight curve* (OWC). If the OWCs of two requests are similar, this may possibly imply that the value of T that is effective for one may also be effective for the other. As with DSCs, normalization can be applied to OWCs. Equations 21 and 22 define distances based on the (**N**ormalized) **O**ffer **W**eight **C**urve.

AOW, ANOW, HOW, HNOW Equations 23–26 define distances based on the **A**verage (**N**ormalized) **O**ffer **W**eight and the **H**ighest (**N**ormalized) **O**ffer **W**eight, respectively.

2.2.3 Distance Measures based on Weights of Initial Search Terms

HICW, LICW It is possible to hypothesize that the *collection frequency weight* [3][10] of an initial search term is associated with search request complexity, and with how successful PRF would be if applied. For example, if an initial query contains a term with a high collection frequency weight, the search may be quite specific and easy [7]. Equations 27 and 28 define distances based on the **H**ighest/**L**owest **I**nitial **C**ollection frequency **W**eights, respectively.

HIRW, LIRW Once the set of pseudo-relevant documents (and therefore the value of R) has been determined, *term reweighting* replaces the collection frequency weights of the initial search terms with *relevance weights* (See **Section 1**). These can also be used for measuring the request-to-request distance, but only for estimating the best value of T. Equations 29 and 30 define distances based on the **H**ighest/**L**owest **R**elevance **W**eights, respectively.

ICWC, IRWC Instead of focusing on the highest/lowest weights of the initial search terms, it is possible to compare the entire initial queries by regarding them as lists of weights in decreasing order, so that they can be regarded as similar if their query length and their weights are similar. Equations 31 and 32 define distances based on the **I**nitial **C**ollection frequency **W**eight **C**urve and the **I**nitial **R**elevance **W**eight **C**urve, respectively.

3 Pre-Submission Experiments

Section 3.1 describes our pre-submission experiments based on average precision over *retrieved* relevant documents, which we used by mistake. This affected our optimization processes, selection of the more promising FPRF strategies, and therefore our official results shown in **Section 3.2**. Therefore, **Section 4** repeats some of our FPRF experiments using standard TREC average precision. Another difference between our pre-submission and post-submission experiments is that, for each NTCIR-1 request, the former treated the *A-relevant* and *B-relevant* documents as relevant while the latter treated only the *A-relevant* documents as relevant. Throughout this paper, our evaluation with the NTCIR-2 test requests treat *S-relevant* (i.e. highly relevant) and *A-relevant* documents as relevant unless otherwise stated.

3.1 Training with the NTCIR-1 Requests

3.1.1 Tuning the General Parameters

The BM25 parameters k_1 and b [4][5], and the PRF parameters R, T and w were optimized using the 60 NTCIR-1 requests with the incomplete relevance data (See **Section 1**). The best run without PRF used $k_1 = 1.5, b = 1.00$, while the best run with traditional PRF used $k_1 = 1.5, b = 0.75$ for the initial search, and $k_1 = 1.5, b = 1.00, R = 10, T = 30, w = 0.3$ for the final search. These are denoted by NOPRF and PRF, respectively.

Based on the above results, we considered two specific *problems* of FPRF:

Problem-R For each test request, select the best value of R from $\{0, 5, 10, 15\}$ when the values of all other parameters are fixed.

Problem-T For each test request, select the best value of T from $\{0, 10, 30, 50\}$ when the values of all other parameters are fixed.

When $R = 0$ (or $T = 0$) is selected for a particular request, that is, when it is estimated that it is better not to apply PRF at all, FPRF behaves like NOPRF. Otherwise, it behaves like PRF except for the value of R (or T).

Variations of the above problems, such as *binary decisions* (e.g. selecting the value of T from $\{0, 30\}$) were not successful in our preliminary experiments. Previous work suggests that varying R and T at the same time is very difficult [7].

3.1.2 Subtasks

For generating our official FPRF runs, we addressed the following two questions:

(a) What are the most reliable *distance measures* for defining the test-to-training mapping?

(b) Which is more reliable, *direct mapping* or *categorization*?

Moreover, some FPRF strategies required parameter tuning (See the **Appendix**). For these purposes, we devised a set of *subtasks* by dividing the set of NTCIR-1 requests into two subsets, each containing 30 requests. Our first *division method* put the first NTCIR-1 request into a subset Q_1, the second into Q_2, and so on. Then, firstly Q_2 was used as the training set, and Q_1 was used as the test set (this will be referred to as *Subtask 1*). Secondly, the subsets were interchanged (this will be referred to as *Subtask 2*). However, because our results were not consistent across the two subtasks, we devised four additional division methods to reshuffle the requests. This yielded eight additional subtasks, with subsets Q_3 through Q_{10}.

Table 1. Official NTCIR-2 test results.

		AveP
R-cNDSC	(CAMUK5)	0.2146 (0.1990)
PRF	(CAMUK6)	0.2139 (0.1977)
T-dAOW	(CAMUK2)	0.2138 (0.1919)
T-cAOW	(CAMUK4)	0.2123 (0.1961)
T-cHOW	(CAMUK3)	0.2117 (0.1932)
T-cHNOW	(CAMUK1)	0.2067 (0.1865)
NOPRF	(CAMUK7)	0.1720 (0.1505)

For each subtask, each FPRF strategy was assessed using the following three *conditions*:

1. It must outperform PRF in terms of average performance.

2. It must outperform PRF for at least n requests if PRF outperforms it for other n requests. We used this rather weak condition because we could not obtain any statistically significant differences regarding Condition 1.

3. It must correctly estimate the best PRF parameter value for at least 9 requests out of 30. This is because, since both Problem-R and Problem-T are essentially selection from 4 possible values, the expected number of correct guesses by a random selection strategy is $30/4 = 7.5$.

Because the final results were still inconsistent across the subtasks, we selected all FPRF strategies that satisfied all of the above three conditions for at least 3 subtasks out of 10, and used them for generating the official runs.

3.2 Official Results with the NTCIR-2 Test Requests

Table 1 shows the results of our 7 official runs. "R-" and "T-" represent FPRF runs that deal with Problem-R and Problem-T, respectively, while "c" and "d" represent Categorization and Direct mapping, respectively. Thus, for example, "R-cNDSC (CAMUK5)" is an FPRF run for Problem-R, based on the Normalized Document Score Curve through categorization. However, recall that these results are based on pre-submission experiments which did *not* use standard TREC average precision. (Table 1 also shows the average precision values based on S-relevant, A-relevant *and B-relevant* documents, in parentheses.)

4 Post-Submission Experiments

After the NTCIR-2 submission, our FPRF strategies were re-run and re-evaluated based on TREC average precision, but by re-using the parameter values of k_1, b, R, T, w mentioned in **Section 3.1.1** rather than re-tuning them. The FPRF parameters were re-tuned using Subtasks 1 and 2 (See the **Appendix**).

Tables 2 and 3 show the corrected results with the NTCIR-2 test requests, for all FPRF strategies that outperformed PRF on average in either Subtask 1 or 2 or both (i.e. in the training experiments). The official runs PRF and NOPRF, which did not require any correction, are included for comparison. Here, R-IDEAL and T-IDEAL are the theoretically best possible FPRF runs, with 100% parameter estimation accuracy. The three table columns correspond to the conditions mentioned in **Section 3.1.2**: For example, Table 2 provides the following information regarding R-IDEAL:

1. Its average precision is 0.2506.

2. It outperforms PRF for 31 requests, while PRF never outperforms it (since R-IDEAL selects the best value of R for each request). The "**" indicates that this difference is statisticallly significant with the sign test ($\alpha = 0.01$).

3. The number of its correct guesses is 49 out of 49 (by definition of an *ideal* FPRF run).

The overall results are quite disappointing: Most of our FPRF strategies failed to outperform PRF, except R-dHRP, R-dRPC, T-dNOWC, T-cAOW and T-cNOWC, and even these exceptions are nowhere near the ideal runs. Moreover, it can be observed that R-dHRP and T-dNOWC are in fact less effective than PRF in terms of per-request "win-or-lose" comparison. On the other hand, the "#correct" column shows that some of our FPRF strategies are more accurate than the random selection strategy, whose expected number of correct guesses in this case is 49/4=12.3. Our most recent work [9] contains re-examination of these FPRF strategies using recent TREC data, as well as more details on the NTCIR-2 results.

Figure 1 shows the actual values of R used for each NTCIR-2 test request by R-IDEAL, R-dHRP and R-dRPC. The height of each dot is slightly adjusted to avoid collision. For example, for Request ID=105, R-dHRP used $R = 15$ while R-dRPC used $R = 10$, but the true best value (used by R-IDEAL) is $R = 0$. That is, PRF is in fact not effective for this request. A similar graph for Problem-T (not included in this paper due to lack of space) suggests that T-cNOWC and T-cAOW favour large values of T.

Regarding Question (a) mentioned in **Section 3.1.2** (What are the most reliable distance measures?), our NTCIR-2 experiments appear to suggest that the relevance probability may be useful

Table 2. Corrected NTCIR-2 test results (Problem-R).

	AveP	>/< PRF		#correct
R-IDEAL	0.2506	31/0	**	49
R-dHRP	**0.2207**	14/18		15
R-dRPC	**0.2179**	17/16		18
PRF (CAMUK6)	0.2139	-		15
R-cARP	0.2114	3/5		14
R-dARDS	0.2011	10/22	*	8
R-cHICW	0.1992	16/28		9
R-dADS	0.1987	9/22	*	11
R-cANDS	0.1978	15/23		13
R-dARP	0.1970	10/18		10
R-dRDSC	0.1965	12/20		13
R-cNDSC	0.1879	15/24		13
R-cICWC	0.1873	11/22		13
R-cRDSC	0.1721	14/30	*	9
NOPRF (CAMUK7)	0.1720	13/33	**	10
R-cARDS	0.1718	14/31	*	8

Table 3. Corrected NTCIR-2 test results (Problem-T).

	AveP	>/< PRF		#correct
T-IDEAL	0.2408	36/0	**	49
T-dNOWC	**0.2177**	12/17		11
T-cAOW	**0.2167**	18/14		17
T-cNOWC	**0.2147**	23/22		17
PRF (CAMUK6)	0.2139	-		11
T-dOWC	0.2138	14/13		15
T-dRPC	0.2128	17/19		14
T-cANOW	0.2095	23/23		15
T-cOWC	0.2067	20/20		16
T-cANDS	0.2064	2/3		11
T-dANOW	0.2063	13/21		12
T-dLIRW	0.2053	17/21		13
T-cHOW	0.2044	23/22		19
T-cHNDS	0.2035	11/20		10
T-dAOW	0.2019	15/23		13
T-cHRDS	0.2018	17/28		11
T-cHNOW	0.2013	21/22		16
T-dRDSC	0.2004	15/23		15
T-dHDS	0.1995	11/19		8
T-dHRDS$'$	0.1992	13/21		7
T-dHRP	0.1985	12/22		11
T-cHIRW	0.1968	16/31	*	9
T-dADS	0.1967	16/20		14
T-dARP	0.1946	13/17		11
T-cHRP	0.1865	13/33	**	5
T-cRPC	0.1842	13/32	**	5
T-cARP	0.1842	13/33	**	5
T-cARDS$''$	0.1731	15/31	*	12
NOPRF (CAMUK7)	0.1720	13/33	**	12
T-cNDSC	0.1717	15/29	*	9

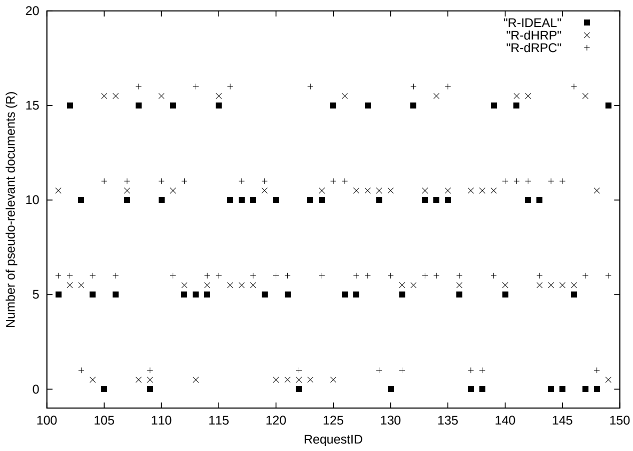

Figure 1. The values of R used for each NTCIR-2 test request.

for determining R, and that the offer weight may be useful for determining T. However, our subsequent experiments using TREC data [9] do not show any advantage of using the relevance probability. As for Question (b) (Which is better, direct mapping or categorization?), our TREC data experiments do suggest that categorization may be more reliable than direct mapping, although this is not clear from our NTCIR-2 results alone. Whether a specific distance calculation method is better than others (e.g. Is *normalization* useful at all?) remains an open question.

5 Conclusions

This paper described our pre-submission and post-submission NTCIR-2 experiments for enhancing the reliability of pseudo-relevance feedback. Although the parameter estimations of our flexible pseudo-relevance feedback strategies are not yet accurate enough for any practical use, our NTCIR-2 results, together with our subsequent TREC data results, suggest that the *offer weight* may be useful for determining the number of expansion terms, and that *categorization* may be more reliable than direct mapping [9].

Our future work includes investigations of other approaches to FPRF such as the use of *optimization tables* [7], which first form training request groups based on evidence such as the initial document scores, and then perform *per-group* optimization of the PRF parameters rather than per-request optimization.

Acknowledgments

We are grateful to Karen Sparck Jones for her comments on the earlier draft of this paper.

References

[1] Harman, D.: What We Have Learned, and Not Learned, from TREC, *BCS IRSG '2000 Proceedings*, 2000, pp. 2–20. http://irsg.eu.org/irsg2000online/papers/harman.htm

[2] Nakayama, K. and Kumano, A.: Collection of Dictionary Data through Internet Translation Service, *AAMT Machine Translation Summit VIII Proceedings*, 1999, pp. 586–592.

[3] Robertson, S. E. and Sparck Jones, K: Simple, Proven Approaches to Text Retrieval, Computer Laboratory, University of Cambridge, 1994.

[4] Robertson, S. E. and Walker, S.: Okapi/Keenbow at TREC-8, *TREC-8 Proceedings*, 2000.

[5] Robertson, S. E. and Walker, S.: Microsoft Cambridge at TREC-9, *TREC-9 Proceedings*, 2001.

[6] Sakai, T. *et al.*: Cross-Language Information Retrieval for NTCIR at Toshiba, *NTCIR Workshop 1 Proceedings*, 1999, pp. 137–144. http://research.nii.ac.jp/˜ntcadm/workshop/OnlineProceedings/017-IR-Sakai.pdf

[7] Sakai, T., Kajiura, M. and Sumita, K.: A First Step towards Flexible Local Feedback for Ad hoc Retrieval, *IRAL 2000 Proceedings*, 2000, pp. 95–102.

[8] Sakai, T.: MT-based Japanese-English Cross-Language IR Experiments using the TREC Test Collections, *IRAL 2000 Proceedings*, 2000, pp. 181–188.

[9] Sakai, T., Robertson, S. E. and Walker, S.: Flexible Pseudo-Relevance Feedback via Direct Mapping and Categorization of Search Requests, *BCS-IRSG ECIR 2001 Proceedings*, 2001, pp. 3–14.

[10] Sparck Jones, K., Walker, S. and Robertson, S. E: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, *Information Processing and Management* 36, 2000, pp. 779–808 (Part I) and 809–840 (Part II).

Appendix: Distances

Distances based on Initial Document Scores

Let $ds_q(r)$ denote the *document score* at rank r in the initial ranked output for a search request q. For a given R', the corresponding *normalized document score* can be defined as follows:

$$nds_q(r, R') = ds_q(r) - ds_q(R'+1) \qquad (1)$$

The distance between a given test request q' and each training request $q(i)$ can be defined based on the *Document Score Curve* (**DSC**) or the *Normalized Document Score Curve* (**NDSC**), as follows:

$$\mathbf{d}_{DSC,R'}(q', q(i)) = \frac{1}{R'} \sum_{r=1}^{R'} |ds_{q'}(r) - ds_{q(i)}(r)| \qquad (2)$$

$$\mathbf{d}_{NDSC,R'}(q', q(i)) = \frac{1}{R'} \sum_{r=1}^{R'} |nds_{q'}(r, R') - nds_{q(i)}(r, R')| \qquad (3)$$

The following is the *Robertson/Walker normalized document score* taken from the TREC-8 filtering track [4]:

$$rds_q(r) = \frac{ds_q(r)}{ads1_q + 0.22maxds_q - 1.3ql_q} \quad (4)$$

where $ads1_q$ is the average document score of the top 1% of initially retrieved documents, $maxds_q$ is the theoretical maximum document score, and ql_q is the initial query length in terms. The following are some variations:

$$rds'_q(r) = \frac{ds_q(r)}{ads1_q + 0.22maxds_q} \quad (5)$$

$$rds''_q(r) = \frac{ds_q(r)}{maxds_q} \quad (6)$$

$$rds'''_q(r) = \frac{ds_q(r)}{ads1_q} \quad (7)$$

The distance based on the *Robertson/Walker normalized Document Score Curve* (**RDSC**) can be defined as follows:

$$\mathbf{d}_{RDSC,R'}(q', q(i)) = \frac{1}{R'} \sum_{r=1}^{R'} |rds_{q'}(r) - rds_{q(i)}(r)| \quad (8)$$

Similarly, $\mathbf{d}_{RDSC',R'}$, $\mathbf{d}_{RDSC'',R'}$ and $\mathbf{d}_{RDSC''',R'}$ can be defined using Equations 5–7. Our pre-submission subtask experiments suggested that tuning the constants in Equation 4 for the NTCIR-2 task is not particularly beneficial.

The following is the Robertson/Walker estimated *relevance probability* taken from the TREC-7/TREC-9 filtering tracks [5]:

$$cds_q(r, \beta, \gamma) = \beta + \gamma \frac{ds_q(r)}{ads1_q} \quad (9)$$

$$rp_q(r, \beta, \gamma) = \frac{\exp cds_q(r, \beta, \gamma)}{1 + \exp cds_q(r, \beta, \gamma)} \quad (10)$$

where $cds_q(r)$ is called the *calibrated document score*, and β and γ are constants that may be tuned using some training data. In our post-submission subtask experiments, we started with the values used at TREC-7/TREC-9 that were obtained through logistic regression ($\beta = -6.77, \gamma = 2.68$), and then adjusted them empirically. The distance based on the *Relevance Probability Curve* (**RPC**) can be defined as follows:

$$\mathbf{d}_{RPC,R',\beta,\gamma}(q', q(i)) =$$

$$\frac{1}{R'} \sum_{r=1}^{R'} |rp_{q'}(r, \beta, \gamma) - rp_{q(i)}(r, \beta, \gamma)| \quad (11)$$

The following distances are based on the *Average Document Score* (**ADS**), the *Average Normalized Document Score* (**ANDS**), the *Average Robertson/Walker normalized Document Score* (**ARDS**), and the *Average Relevance Probability* (**ARP**), respectively:

$$\mathbf{d}_{ADS,R'}(q', q(i)) = |\frac{1}{R'} \sum_{r=1}^{R'} ds_{q'}(r) - \frac{1}{R'} \sum_{r=1}^{R'} ds_{q(i)}(r)| \quad (12)$$

$$\mathbf{d}_{ANDS,R'}(q', q(i)) =$$

$$|\frac{1}{R'} \sum_{r=1}^{R'} nds_{q'}(r, R') - \frac{1}{R'} \sum_{r=1}^{R'} nds_{q(i)}(r, R')| \quad (13)$$

$$\mathbf{d}_{ARDS,R'}(q', q(i)) =$$

$$|\frac{1}{R'} \sum_{r=1}^{R'} rds_{q'}(r) - \frac{1}{R'} \sum_{r=1}^{R'} rds_{q(i)}(r)| \quad (14)$$

$$\mathbf{d}_{ARP,R',\beta,\gamma}(q', q(i)) =$$

$$|\frac{1}{R'} \sum_{r=1}^{R'} rp_{q'}(r, \beta, \gamma) - \frac{1}{R'} \sum_{r=1}^{R'} rp_{q(i)}(r, \beta, \gamma)| \quad (15)$$

Similarly, $\mathbf{d}_{ARDS',R'}$, $\mathbf{d}_{ARDS'',R'}$ and $\mathbf{d}_{ARDS''',R'}$ can be defined using Equations 5–7.

The following distances are based on the *Highest Document Score* (**HDS**), the *Highest Normalized Document Score* (**HNDS**), the *Highest Robertson/Walker normalized Document Score* (**HRDS**), and the *Highest Relevance Probability* (**HRP**), respectively:

$$\mathbf{d}_{HDS}(q', q(i)) = \mathbf{d}_{DSC,1}(q', q(i)) =$$

$$\mathbf{d}_{ADS,1}(q', q(i)) = |ds_{q'}(1) - ds_{q(i)}(1)| \quad (16)$$

$$\mathbf{d}_{HNDS,R'}(q', q(i)) = |nds_{q'}(1, R') - nds_{q(i)}(1, R')| \quad (17)$$

$$\mathbf{d}_{HRDS}(q', q(i)) = |rds_{q'}(1) - rds_{q(i)}(1)| \quad (18)$$

$$\mathbf{d}_{HRP,\beta,\gamma}(q', q(i)) = |rp_{q'}(1, \beta, \gamma) - rp_{q(i)}(1, \beta, \gamma)| \quad (19)$$

Similarly, $\mathbf{d}_{HRDS'}$, $\mathbf{d}_{HRDS''}$ and $\mathbf{d}_{HRDS'''}$ can be defined using Equations 5–7.

Distances based on Offer Weights

For a request q with a given set of pseudo-relevant documents, let $ow_q(r)$ denote the r-th highest value among the *offer weights* [3][10] of candidate expansion terms. For a given T', the corresponding *normalized offer weight* can be defined as follows:

$$now_q(r, T') = ow_q(r) - ow_q(T' + 1) \quad (20)$$

The distance between q' and $q(i)$ can be calculated based on the *Offer Weight Curve* (**OWC**) or the *Normalized Offer Weight Curve* (**NOWC**), as follows:

$$\mathbf{d}_{OWC,T'}(q', q(i)) = \frac{1}{T'} \sum_{r=1}^{T'} |ow_{q'}(r) - ow_{q(i)}(r)| \quad (21)$$

$$\mathbf{d}_{NOWC,T'}(q', q(i)) =$$

$$\frac{1}{T'} \sum_{r=1}^{T'} |now_{q'}(r, T') - now_{q(i)}(r, T')| \qquad (22)$$

The following distances are based on the *Average Offer Weight* (**AOW**), the *Average Normalized Offer Weight* (**ANOW**), the *Highest Offer Weight* (**HOW**), and the *Highest Normalized Offer Weight* (**HNOW**), respectively.

$$\mathbf{d}_{AOW,T'}(q', q(i)) =$$

$$|\frac{1}{T'} \sum_{r=1}^{T'} ow_{q'}(r) - \frac{1}{T'} \sum_{r=1}^{T'} ow_{q(i)}(r)| \qquad (23)$$

$$\mathbf{d}_{ANOW,T'}(q', q(i)) =$$

$$|\frac{1}{T'} \sum_{r=1}^{T'} now_{q'}(r, T') - \frac{1}{T'} \sum_{r=1}^{T'} now_{q(i)}(r, T')| \quad (24)$$

$$\mathbf{d}_{HOW}(q', q(i)) = \mathbf{d}_{OWC,1}(q', q(i)) =$$

$$\mathbf{d}_{AOW,1}(q', q(i)) = |ow_{q'}(1) - ow_{q(i)}(1)| \qquad (25)$$

$$\mathbf{d}_{HNOW,T'}(q', q(i)) = |now_{q'}(1, T') - now_{q(i)}(1, T')| \qquad (26)$$

Distances based on Initial Search Terms Weights

For a request q, let $icw_q(r)$ denote the r-th highest value among the *collection frequency weights* [3][10] of the *initial* search terms. Then, $icw_q(1)$ and $icw_q(ql_q)$ are called the *Highest/Lowest Initial Collecion frequency Weights* (**HICW/LICW**), respectively, where ql_q is the initial query length in terms. Likewise, let $irw_q(r)$ denote the r-th highest value among the *relevance weights* [3][10] of the *initial* search terms, after term reweighting based on a set of pseudo-relevant documents has been performed. Then, $irw_q(1)$ and $irw_q(ql_q)$ are called the *Highest/Lowest Initial Relevance Weights* (**HIRW/LIRW**), respectively. The following distances are based on the above values:

$$\mathbf{d}_{HICW}(q', q(i)) = |icw_{q'}(1) - icw_{q(i)}(1)| \qquad (27)$$

$$\mathbf{d}_{LICW}(q', q(i)) = |icw_{q'}(ql_{q'}) - icw_{q(i)}(ql_{q(i)})| \qquad (28)$$

$$\mathbf{d}_{HIRW}(q', q(i)) = |irw_{q'}(1) - irw_{q(i)}(1)| \qquad (29)$$

$$\mathbf{d}_{LIRW}(q', q(i)) = |irw_{q'}(ql_{q'}) - irw_{q(i)}(ql_{q(i)})| \qquad (30)$$

For convenience, let $icw_q(r) = irw_q(r) = 0$ for $r > ql_q$. Then, the distances based on the *Initial Collection Frequency Weight Curve* (**ICWC**) and the *Initial Relevance Weight Curve* (**IRWC**) can be defined as follows:

$$\mathbf{d}_{ICWC}(q', q(i)) = \frac{1}{maxql} \sum_{r=1}^{maxql} |icw_{q'}(r) - icw_{q(i)}(r)| \qquad (31)$$

Table 4. FPRF parameter values.

	parameter values
R-cNDSC (CAMUK5)	$R' = 10$
R-cNDSC	$R' = 30$
R-dRDSC	$R' = 10$
R-cRDSC	$R' = 10$
R-dRPC	$R' = 10, \beta = -10, \gamma = 2.68$
R-dADS	$R' = 50$
R-cANDS	$R' = 30$
R-dARDS	$R' = 10$
R-cARDS	$R' = 10$
R-dARP	$R' = 50, \beta = -6.77, \gamma = 2.68$
R-cARP	$R' = 50, \beta = -6.77, \gamma = 2.68$
R-dHRP	$\beta = -10, \gamma = 2.68$
T-cNDSC	$R' = 50$
T-dRDSC	$R' = 30$
T-dRPC	$R' = 10, \beta = -6.77, \gamma = 2.68$
T-cRPC	$R' = 10, \beta = -6.77, \gamma = 1$
T-dADS	$R' = 50$
T-cANDS	$R' = 10$
T-cARDS”	$R' = 50$
T-dARP	$R' = 10, \beta = -6.77, \gamma = 2.68$
T-cARP	$R' = 10, \beta = -6.77, \gamma = 5$
T-cHNDS	$R' = 30$
T-dHRP	$\beta = -10, \gamma = 2.68$
T-cHRP	$\beta = -3, \gamma = 2.68$
T-dOWC	$T' = 30$
T-cOWC	$T' = 10$
T-dNOWC	$T' = 30$
T-cNOWC	$T' = 30$
T-dAOW (CAMUK2)	$T' = 30$
T-dAOW	$T' = 10$
T-cAOW (CAMUK3)	$T' = 30$
T-cAOW	$T' = 30$
T-dANOW	$T' = 10$
T-cANOW	$T' = 10$
T-cHNOW (CAMUK1)	$T' = 10$
T-cHNOW	$T' = 10$

$$\mathbf{d}_{IRWC}(q', q(i)) = \frac{1}{maxql} \sum_{r=1}^{maxql} |irw_{q'}(r) - irw_{q(i)}(r)| \qquad (32)$$

where $maxql = \max(ql_{q'}, ql_{q(i)})$.

Categorization

In the case of *categorization* of requests, the distance between q' and each *group* of training requests $G(j)$ can be determined as follows, based on any of the aforementioned request-to-request distance:

$$\mathbf{D}_X(q', G(j)) = \frac{1}{|G(j)|} \sum_{q(i) \in G(j)} \mathbf{d}_X(q', q(i)) \quad (33)$$

Tuning the FPRF Parameters

Table 4 shows the actual parameter values used with our FPRF strategies in our pre-submission and post-submission experiments, which were selected based on the results with Subtasks 1 and 2. Some of the pre-submission values (labelled with CAMUK) are different from the post-submission ones because (1) they are not based on TREC average precision; (2) they treat B-relevant documents as relevant as well (See **Section 3**).