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 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 across requests, four FPRF runs which vary the number of expansion terms across requests, and traditional runs with and without PRF.

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 $d$ 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 $d$ 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 $D$ between a given test request and each group is derived from individual $d$’s, for example, by taking the average. Finally, the optimal parameter values for the group that minimizes $D$ are used for the test request. Relying on the request-to-group distance ($D$) instead of the request-to-request distance ($d$) 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 $d$. 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 output, 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 (Normalized) Document Score Curve.

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 Robertson/Walker normalized Document Score Curve. The Appendix also describes some variations of this distance measure, such as $RDSC^c$ and $RDSC^n$.

**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 Relevance Probability Curve.

**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 Average (Normalized) Document Score, the Average Robertson/Walker normalized Document Score and the Average Relevance Probability, 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 Highest (Normalized) Document Score, the Highest Robertson/Walker normalized Document Score and the Highest Relevance Probability, 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 (Normalized) Offer Weight Curve.

**AOW, ANOW, HOW, HNOW** Equations 23–26 define distances based on the Average (Normalized) Offer Weight and the Highest (Normalized) Offer Weight, 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 Highest/Lowest Initial Collection frequency Weights, 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 Highest/Lowest Relevance Weights, 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 Initial Collection Frequency Weight Curve and the Initial Relevance Weight Curve, 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 resuffle the requests. This yielded eight additional subtasks, with subsets \( Q_3 \) through \( Q_{10} \).

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.)

<table>
<thead>
<tr>
<th>Table 1. Official NTCIR-2 test results.</th>
<th>AveP</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R)-cNDSC (CAMUK5)</td>
<td>0.2146 (0.1990)</td>
</tr>
<tr>
<td>PRF</td>
<td>0.2139 (0.1977)</td>
</tr>
<tr>
<td>( T)-dAOW (CAMUK2)</td>
<td>0.2138 (0.1919)</td>
</tr>
<tr>
<td>( T)-cAOW (CAMUK4)</td>
<td>0.2123 (0.1961)</td>
</tr>
<tr>
<td>( T)-cHOW (CAMUK3)</td>
<td>0.2117 (0.1932)</td>
</tr>
<tr>
<td>( T)-cNOW (CAMUK1)</td>
<td>0.2067 (0.1865)</td>
</tr>
<tr>
<td>NOPRF (CAMUK7)</td>
<td>0.1720 (0.1505)</td>
</tr>
</tbody>
</table>

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 $R$-идеал:

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 statistically 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, $T$-cNOWC 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 $T$-cNOWC 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-$R$). |

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

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

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


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:

\[
nds_q(r, R') = ds_q(r) - ds_q(R' + 1)
\]  

(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:

\[
d_{DSC,R'}(q', q(i)) = \frac{1}{R'} \sum_{r=1}^{R'} |ds_q(r) - ds_{q(i)}(r)|
\]  

(2)

\[
d_{NDSC,R'}(q', q(i)) = \frac{1}{R'} \sum_{r=1}^{R'} |nds_q(r, R') - nds_{q(i)}(r, R')|
\]  

(3)
The following is the Robertson/Walker normalized document score taken from the TREC-8 filtering track [4]:

\[ rds_q(r) = \frac{ds_q(r)}{ads_I_q + 0.22maxds_q - 1.3q_I} \]  
(4)

where \( ds_I_q \) is the average document score of the top 1% of initially retrieved documents, \( maxds_q \) is the theoretical maximum document score, and \( q_I \) is the initial query length in terms. The following are some variations:

\[ rds^I_q(r) = \frac{ds_q(r)}{ads_I_q} \]  
(5)

\[ rds^II_q(r) = \frac{ds_q(r)}{maxds_q} \]  
(6)

\[ rds^{III}_q(r) = \frac{ds_q(r)}{ads_I_q} \]  
(7)

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

\[ d_{RDSC, f}(q', q(i)) = \frac{1}{R^f} \sum_{r=1}^{R^f} |rds_q(r) - rds_{q(i)}(r)| \]  
(8)

Similarly, \( d_{RDSC,R^f} \), \( d_{RDSC,R^f}^{*} \), and \( d_{RDSC^{*},R^f} \) 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]:

\[ cds_q(r, \beta, \gamma) = \beta + \gamma \frac{ds_q(r)}{ads_I_q} \]  
(9)

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

where \( cds_q(r) \) is the calibrated document score, and \( \beta \) and \( \gamma \) 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 \( (\beta = -0.77, \gamma = 2.68) \), and then adjusted them empirically. The distance based on the Relevance Probability Curve (RPC) can be defined as follows:

\[ d_{RPC,R^f}(q', q(i)) = \frac{1}{R^f} \sum_{r=1}^{R^f} |rp_q(r, \beta, \gamma) - rp_{q(i)}(r, \beta, \gamma)| \]  
(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:

\[ d_{ADS,R^f}(q', q(i)) = \left| \frac{1}{R^f} \sum_{r=1}^{R^f} ds_q(r) - \frac{1}{R^f} \sum_{r=1}^{R^f} ds_{q(i)}(r) \right| \]  
(12)

\[ d_{ANDS,R^f}(q', q(i)) = \left| \frac{1}{R^f} \sum_{r=1}^{R^f} nds_q(r, R^f) - \frac{1}{R^f} \sum_{r=1}^{R^f} nds_{q(i)}(r, R^f) \right| \]  
(13)

\[ d_{ARDS,R^f}(q', q(i)) = \left| \frac{1}{R^f} \sum_{r=1}^{R^f} rds_q(r) - \frac{1}{R^f} \sum_{r=1}^{R^f} rds_{q(i)}(r) \right| \]  
(14)

\[ d_{ARP,R^f,\beta,\gamma}(q', q(i)) = \left| \frac{1}{R^f} \sum_{r=1}^{R^f} rp_q(r, \beta, \gamma) - \frac{1}{R^f} \sum_{r=1}^{R^f} rp_{q(i)}(r, \beta, \gamma) \right| \]  
(15)

Similarly, \( d_{ARDS,R^f}^{*} \), \( d_{ARDS^*,R^f} \), and \( d_{ARDS^{*},R^f} \) can be defined using Equations 5–7.

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

\[ d_{HDS}(q', q(i)) = d_{HDS,1}(q', q(i)) = |ds_q(1) - ds_{q(i)}(1)| \]  
(16)

\[ d_{HND,R^f}(q', q(i)) = |nds_q(1, R^f) - nds_{q,1}(1, R^f)| \]  
(17)

\[ d_{HRDS}(q', q(i)) = |rds_q(1) - rds_{q(i)}(1)| \]  
(18)

\[ d_{HRP,\beta,\gamma}(q', q(i)) = |rp_{q}(1, \beta, \gamma) - rp_{q(i)}(1, \beta, \gamma)| \]  
(19)

Similarly, \( d_{HRS^{*},R^f} \), \( d_{HRS^{*},R^f}^{*} \), and \( d_{HRS^{*},R^f}^{**} \) 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:

\[ now_q(r, T') = ow_q(r) - ow_q(T' + 1) \]  
(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:

\[ d_{OWC,T'}(q', q(i)) = \frac{1}{T'} \sum_{r=1}^{T'} |ow_q(r) - ow_{q(i)}(r)| \]  
(21)
\[
\mathbf{d}_{\text{NOW}, \text{T}}(q', q(i)) = \frac{1}{T} \sum_{r=1}^{T'} |\text{now}_{q}(r, T') - \text{now}_{q(i)}(r, T')| \tag{22}
\]

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

\[
\mathbf{d}_{\text{AOOW}, \text{T}}(q', q(i)) = \frac{1}{T} \sum_{r=1}^{T'} \text{ow}_{q}(r) - \frac{1}{T} \sum_{r=1}^{T'} \text{ow}_{q(i)}(r) \tag{23}
\]

\[
\mathbf{d}_{\text{ANOOW}, \text{T}}(q', q(i)) = |\frac{1}{T} \sum_{r=1}^{T'} \text{now}_{q}(r, T') - \frac{1}{T} \sum_{r=1}^{T'} \text{now}_{q(i)}(r, T')| \tag{24}
\]

\[
\mathbf{d}_{\text{HOW}, \text{T}}(q', q(i)) = \mathbf{d}_{\text{OWC}, \text{I}}(q', q(i)) = |\text{ow}_{q}(1) - \text{ow}_{q(i)}(1)| \tag{25}
\]

\[
\mathbf{d}_{\text{HNOW}, \text{T}}(q', q(i)) = |\text{now}_{q}(1, T') - \text{now}_{q(i)}(1, T')| \tag{26}
\]

**Distances based on Initial Search Terms Weights**

For a request \(q\), let \(\text{ic}_{wq}(r)\) denote the \(r\)-th highest value among the collection frequency weights \([3][10]\) of the initial search terms. Then, \(\text{ic}_{wq}(1)\) and \(\text{ic}_{wq}(q_{i})\) are called the Highest/Lowest Initial Collection Frequency Weights (HICW/LICW), respectively, where \(q_{i}\) is the initial query length in terms. Likewise, let \(\text{ir}_{wq}(r)\) 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, \(\text{ir}_{wq}(1)\) and \(\text{ir}_{wq}(q_{i})\) are called the Highest/Lowest Initial Relevance Weights (HIRW/LIRW), respectively. The following distances are based on the above values:

\[
\mathbf{d}_{\text{HICW}}(q', q(i)) = |\text{ic}_{wq}(1) - \text{ic}_{wq(i)}(1)| \tag{27}
\]

\[
\mathbf{d}_{\text{LICW}}(q', q(i)) = |\text{ic}_{wq}(q_{i}) - \text{ic}_{wq(i)}(q_{i}(i))| \tag{28}
\]

\[
\mathbf{d}_{\text{HIRW}}(q', q(i)) = |\text{ir}_{wq}(1) - \text{ir}_{wq(i)}(1)| \tag{29}
\]

\[
\mathbf{d}_{\text{LIRW}}(q', q(i)) = |\text{ir}_{wq}(q_{i}) - \text{ir}_{wq(i)}(q_{i}(i))| \tag{30}
\]

For convenience, let \(\text{ic}_{wq}(r) = \text{ir}_{wq}(r) = 0\) for \(r > q_{i}\). Then, the distances based on the Initial Collection Frequency Weight Curve (ICWC) and the Initial Relevance Weight Curve (IRWC) can be defined as follows:

\[
\mathbf{d}_{\text{ICWC}}(q', q(i)) = \frac{1}{\text{maxq}'} \sum_{r=1}^{\text{maxq}'} |\text{ic}_{wq}(r) - \text{ic}_{wq(i)}(r)| \tag{31}
\]

\[
\mathbf{d}_{\text{IRWC}}(q', q(i)) = \frac{1}{|G(j)|} \sum_{q(i) \in G(j)} \mathbf{d}_{x}(q', q(i)) \tag{32}
\]

**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).

**Table 4. FPRF parameter values.**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R^{+})</td>
<td>10</td>
</tr>
<tr>
<td>(R^{-})</td>
<td>30</td>
</tr>
<tr>
<td>(R^{0})</td>
<td>10</td>
</tr>
<tr>
<td>(R_{\text{CP}})</td>
<td>10, (\beta = -10, \gamma = 2.68)</td>
</tr>
<tr>
<td>(R_{\text{ADS}})</td>
<td>50</td>
</tr>
<tr>
<td>(R_{\text{ANDS}})</td>
<td>30</td>
</tr>
<tr>
<td>(R_{\text{ARDS}})</td>
<td>10</td>
</tr>
<tr>
<td>(R_{\text{ARP}})</td>
<td>50, (\beta = -6.77, \gamma = 2.68)</td>
</tr>
<tr>
<td>(R_{\text{HRP}})</td>
<td>50, (\beta = -6.77, \gamma = 2.68)</td>
</tr>
</tbody>
</table>

\(R_{\text{NDSC}}(\text{CAMUK})\) \(R^{+} = 10\)
\(R_{\text{NDSC}}\) \(R^{+} = 10\)
\(R_{\text{DRDC}}\) \(R^{+} = 10\)
\(R_{\text{RDCS}}\) \(R^{+} = 10\)
\(R_{\text{DRPC}}\) \(R^{+} = 10, \beta = -10, \gamma = 2.68\)
\(R_{\text{ADS}}\) \(R^{+} = 50\)
\(R_{\text{ANDS}}\) \(R^{+} = 10\)
\(R_{\text{ARDS}}\) \(R^{+} = 10\)
\(R_{\text{ARP}}\) \(R^{+} = 50, \beta = -6.77, \gamma = 2.68\)
\(R_{\text{HRP}}\) \(R^{+} = 50, \beta = -6.77, \gamma = 2.68\)