Pseudo-Relevance Feedback Method based on the Cross Product of Irrelevant Documents

Rabeb Mbarek¹, Mohamed Tmar¹, Hawete Hattab², Mohand Boughanem³
¹Multimedia Information systems and Advanced Computing Laboratory
High Institute of Computer Science and Multimedia
University of Sfax, Sfax, Tunisia
²Umm Al-Qura University Kingdom of Saudi Arabia
³University of Toulouse - IRIT lab France
rabeb.hattab@gmail.com, mohamedtmar@yahoo.fr, hattab.hawete@yahoo.fr, bougha@irit.fr
http://www.miracl.rnu.tn

ABSTRACT: Pseudo-Relevance Feedback assumes that the top-ranked k documents of the initial retrieval are relevant, and then terms of these documents are used to re-weight the terms of the initial query (add new terms and/or change the weights of existing terms in the query). In this paper, we propose a new approach for query expansion for ad hoc search, by using an absorbing document which is the cross product of irrelevant documents. This document will be orthogonal to irrelevant ones. We show how this absorbing document can extract better expansion terms from the top-ranked k documents. The experiments show that our approach gives improvements for both collections, TREC-7 and TREC-8, over known models.

Keywords: Pseudo-Relevance Feedback, Absorption of irrelevance, Cross Product

Received: 30 December 2015, Revised 24 January 2016, Accepted 02 February 2016

© 2016 DLINE. All Rights Reserved

1. Introduction

To refine the Information Retrieval (IR) process, it is required to reformulate the query using query expansion [13], query substitution [8], and other refinement techniques [2, 9]. Pseudo-Relevance Feedback (PRF) is a well-studied query expansion technique which assumes that the top-ranked documents of the initial retrieval are relevant and expansion terms are then extracted from them [6]. If there are only a few or no relevant documents in the top-ranked documents, then we can add terms which have no relationships with the topic of relevance of the query and so the PRF only improves the performance of queries which have good initial results. Thus, to improve the PRF technique it suffices to effectively select from top-ranked documents those terms that are most likely relevant to the query topic.

In this paper, we propose to solve this problem by exploiting the role of irrelevant documents in selecting better expansion terms from the top-ranked documents. In particular we build an absorbing document which is the cross product of linearly independent irrelevant documents. This document will be orthogonal to irrelevant ones.

How to automatically identify irrelevant documents is an open question. We propose to exploit documents at the bottom of
the rank. This strategy is widely used in IR [17, 3].

In this paper, the Vector Space Model (VSM) [15] is adopted to rank the documents. The VSM showed good feedback performance on most collections whereas the probabilistic model had problems with some collections [7].

This paper is organized as follows. Section 2 presents the related work. In Section 3 we define the cross product. Section 4 describes our approach based on the absorption of irrelevance. Experiments performed for evaluating our approach are presented in Section 5. The conclusion and future work are presented in section 6.

2. Related work

Croft and Harper [5] first suggested the technique of PRF for estimating the probabilities within the probabilistic model. Due to the sensitivity to the quality of top \( k \) documents, PRF is not robust to the quality of the initial retrieval. Several approaches have been proposed to improve the robustness of PRF; see for example [18, 4, 1, 21, 25, 11, 27, 28].

Several works have investigated the role of irrelevant documents on PRF performance. Rocchio’s original formulation explicitly includes a component of irrelevant documents [16]. The aim of Rocchio’s approach is to boost the terms from relevant documents and reduce the weights of terms from irrelevant ones. Note that we can also use this approach to perform negative feedback by ignoring its component of relevant documents. Singhal et al. [20] achieved an interesting result for the learning routing query problem: they showed that using irrelevant documents close to the query, in place of those in the whole collection, is more effective. Raman et al. [14], introduced the notion of pseudo-irrelevant documents, i.e., high-scoring documents outside of top \( k \) that are highly unlikely to be relevant. They show how pseudo-irrelevant documents can be used to extract effective expansion terms from the top-ranked \( k \) documents. A successful use of irrelevant documents for negative pseudo-relevance feedback has been carried out in [23], where Wang et al. pointed out the effectiveness of their approach with poorly performing queries. Basile et al. [3], proposed a novel approach to document re-ranking, which relies on the concept of negative feedback represented by irrelevant documents. In their work, the concept of irrelevance is defined as a quantum operator.

The modelling of terms negation in pseudo-relevance feedback by quantum logic operators is due to Widdows [24]. In his work, Widdows has shown that negation in quantum logic is able to remove, from the result set, not only un-wanted terms but also their related meaning. The concept of vectors orthogonality is exploited to express queries like ‘Retrieve documents that contain term A NOT term B’. Widdows suggested that vectors which represent unrelated concepts should be orthogonal to each other. Indeed, orthogonality prevents vectors from sharing common features.

3. Cross Product

Let \( E \) be a vector space of dimension \( n \) and let \( u_1, \ldots, u_{n-1} \) be \( n-1 \) vectors of \( E \). For each vector \( x \) of \( E \) there exists a unique vector \( w \) such that

\[
det(u_1, \ldots, u_{n-1}, x) = w^T \cdot x
\]

where \( det \) is the determinant of \( n \) vectors, \( w^T \) is the transpose of \( w \) and \( w^T \cdot x \) is the classical inner product.

\( w \) is called the cross product of \( u_1, \ldots, u_{n-1} \) and is denoted by \( u_1 \wedge \ldots \wedge u_{n-1} \) (for \( n = 3 \), see Figure 1). We have the following properties:

1. The vector \( u_1 \wedge \ldots \wedge u_{n-1} \) is orthogonal to each vector \( u_i \).
2. The vector \( u_1 \wedge \ldots \wedge u_{n-1} \) is orthogonal to the subspace \( F \) of \( E \) generated by the family \( (u_1 \wedge \ldots \wedge u_{n-1}) \). Indeed, if \( u \) is a vector of \( F \), there exists \( n-1 \) scalars \( \alpha_1, \ldots, \alpha_{n-1} \) such that \( u = \alpha_1 u_1 + \ldots + \alpha_{n-1} u_{n-1} \).
3. \( u_1 \wedge \ldots \wedge u_{n-1} = 0 \) if and only if \( u_1, \ldots, u_{n-1} \) are linearly dependent.
4. If \( u_1, \ldots, u_{n-1} \) are linearly independent then \( (u_1, \ldots, u_{n-1}, u_1 \wedge \ldots \wedge u_{n-1}) \) is a basis of \( E \).
In the following we compute the cross product of $u_1, ..., u_{n-1}$.

Let $A = (u_1, ..., u_{n-1})$ be the matrix of $n$ rows and $n - 1$ columns. Let $A_i$ be the matrix obtained from the matrix $A$ by deleting the $i$-th row ($1 \leq i \leq n$).

The cross product of $u_1, ..., u_{n-1}$ is the vector:

$$u_1 \wedge ... \wedge u_{n-1} = \begin{pmatrix}
det A_1 \\
-\det A_2 \\
... \\
(1)^{n-1} \det A_n
\end{pmatrix}$$

(1)

The Equation 1 generalizes the definition of cross product of two vectors in dimension 3.

We give an example of cross product of three vectors in dimension 4 and we show that properties (1) and (2) are satisfied. Let $u_1 = (1, 0, 1, -1)^T$, $u_2 = (0, 2, 1, 1)^T$ and $u_3 = (1, 3, 1, 0)^T$ be three vectors. The cross product is $u_1 \wedge u_2 \wedge u_3 = (4, -1, -1, 3)^T$.

Since $(4, -1, -1, 3), (1, 0, 1, -1)^T = (4, -1, -1, 3), (0, 2, 1, 1)^T = (4, -1, -1, 3), (1, 3, 1, 0)^T = 0$, the property (1) follows.

Let $u_4$ be a linear combination of $u_1, u_2$ and $u_3$, i.e. there exist $\alpha, \beta$ and $\gamma$ such that $u_4 = \alpha u_1 + \beta u_2 + \gamma u_3$. Since $(4, -1, -1, 3), u_4 = \alpha (4, -1, -1, 3). u_1 + \beta (4, -1, -1, 3). u_2 + \gamma (4, -1, -1, 3). u_3 = 0 + 0 + 0 = 0$, the property (2) follows.

4. Absorption of Irrelevance (AI)

This section describes our PRF approach based on irrelevant documents. The main idea is to build an absorbing document noted $\overrightarrow{\theta}$, as the cross product of linearly independent irrelevant documents and then terms of this document are used to re-weight the terms of the original query in the following way:

$$Q_{new} = \alpha Q_{init} + (1 - \alpha). \overrightarrow{\theta}$$

Where $\overrightarrow{\theta}$ is the absorbing document and $\alpha$ is a real parameter between 0 and 1.

Let $D_{init}$ be the initial set of ranked documents and let $n$ be the number of indexing terms of $D_{init}$.

Identifying relevant documents $D^+$ is quite straightforward, we assume that the top-ranked $k$ documents in $D_{init}$ as relevant. Let $p$ be the number of expansion terms of the top-ranked $k$ documents.

Identifying irrelevant documents is not trivial, we propose to select the set of irrelevant documents $D^-$ from the bottom of
Let $m < p$ be the number of linearly independent documents of $D^-$. Let $u_1, ..., u_m$ denote these irrelevant documents. Each irrelevant document of $D^-$ is a linear combination of $u_1, ..., u_m$.

To compute the absorbing document $\vec{d}$ which is the cross product of $u_1, ..., u_m$, each vector must be written as a linear combination of $m+1$ indexing terms (see Section 3). For this reason, we must choose $m+1$ terms from the $p$ expansion ones, i.e., $p = m + 1$.

The absorbing document is $\vec{d} = u_1 \wedge ... \wedge u_m$. By Section 3 item (2), $\vec{d}$ is orthogonal to each irrelevant document.

### 5. Experiments

In this section we give the different experiments and results obtained to evaluate our approach.

#### 5.1 Evaluation Methodology

We set up a baseline system based on the BM25 formula proposed in [17]. BM25 parameters are $b = 0.5, k_1 = 1.2, k_2 = 0$ and $k_3 = 8$. (For parameters definitions and use please refer 17). The TREC-7 and TREC-8 collections were used for test [22]. They consist of the same set of documents ($t_4$, i.e., TREC disks 4 and 5, containing approximately 2 gigabytes of data) and different query sets (topics 351-400 and topics 401-450, respectively). The full topic statement was considered, including title, description, and narrative. TREC-7 topics were described with an average of 57.6 terms, while the average on TREC-8 topics was 51.8 terms.

To generate a query $Q_{int}$, the title of a topic was used, thus falling into line with the common practice of TREC experiments; description and narrative title were not used. Using $Q_{int}$ the top 1000 documents are retrieved from the collections.

The set of relevant documents $D^+$ is the set of top-ranked $k$ documents, while the set of irrelevant documents $D^-$ is the set of retrieved documents 501-1000, assumed to be irrelevant.

The experiments consist of re-ranking the results of the Baseline Model. For our approach AI the reformulated query is:

$$Q_{new} = \alpha Q_{int} + (1-\alpha) \vec{d}$$  \hspace{1cm} (2)

Where $\vec{d}$ is the absorbing document and $\alpha$ is a real parameter between 0 and 1.

We compare our approach AI to the Baseline model BM25 and to the traditional combination of BM25 and Rocchio’s feedback model1 (BM25+Rocchio).

The following improved version [19] of the original Rocchio’s formula [16] is used:

$$Q_{new} = \alpha Q_{int} + \beta \frac{1}{|D^+|} \sum_{d \in D^+} d$$ \hspace{1cm} (3)

Here, $\alpha$ and $\beta$ are tuning constants controlling how much we rely on the original query and the feedback information. In practice, we can always fix $\alpha$ at 1, and only study $\beta$ in order to get better performance.

For our approach AI and the BM25+Rocchio model, the retrieved documents are re-ranked by the inner product done by:

$$< Q_{new}, d > = Q_{new}^T d$$ \hspace{1cm} (4)

#### 5.2 Parameters Settings

The experiments and the evaluations are as follow. Comparison between our AI model, the BM25 model and the BM25+Rocchio model.

---

1 According to [26], BM25 [17] term weighting coupled with Rocchio feedback remains a strong baseline.
We vary the BM25 parameters $k_1$ from 1 to 3 in steps of 0.1 and $b$ from 0.05 to 1 in steps of 0.05.

We vary the parameters in Equation 2 and in Equation 3 from 0 to 1 in steps of 0.1.

The two models AI and BM25+Rocchio depend on the RF parameters. One parameter was the number $k$ of relevant documents. The other parameter was the number $p$ of expansion terms. We varied these two parameters in the following way: $k \in \{1, 2, 3, 4, 5\}$ et $p \in \{10, 20, 30, 50\}$. Note that the number $m$ of linearly independent irrelevant documents must be equal to $p - 1$.

### 5.3 Results

To evaluate the performance of our approach we use MAP, R-Precision, P@5, P@10 and P@20 as evaluation measures. These measures are the most commonly used measures of overall retrieval performance.

We present here the behavior of evaluation measures on TREC-7 and TREC-8 for the three models: AI, BM25 and BM25+Rocchio.

The optimal results are illustrated in Tables 1 and 2.

These results are obtained for the following values of parameters:

- For our approach, $k = 2$, $p = 50$, $m = 49$ and $\alpha = 0.3$;
- For the BM25 model, $k_1 = 1$ and $b = 0.25$ for TREC-7. $k_1 = 1$ and $b = 0.4$ for TREC-8;
- For the Rocchio model, $k = 2$, $p = 30$ and $\delta = 1$.

<table>
<thead>
<tr>
<th></th>
<th>BM25</th>
<th>BM25+Rocchio</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.1907</td>
<td>0.2533</td>
<td>0.2831</td>
</tr>
<tr>
<td>R-Precision</td>
<td>0.246</td>
<td>0.295</td>
<td>0.355</td>
</tr>
<tr>
<td>P@5</td>
<td>0.468</td>
<td>0.451</td>
<td>0.6</td>
</tr>
<tr>
<td>P@10</td>
<td>0.43</td>
<td>0.44</td>
<td>0.56</td>
</tr>
<tr>
<td>P@20</td>
<td>0.364</td>
<td>0.381</td>
<td>0.513</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the performances between the models AI, BM25 and BM25+Rocchio on TREC-7 collection

### 5.4 Comparison with BM25 and BM25+Rocchio

From Tables 1 and 2, we can clearly see that the average performance of our AI model is superior to the basic models BM25 and BM25+Rocchio.

<table>
<thead>
<tr>
<th></th>
<th>BM25</th>
<th>BM25+Rocchio</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.2113</td>
<td>0.2641</td>
<td>0.3023</td>
</tr>
<tr>
<td>R-Precision</td>
<td>0.2626</td>
<td>0.313</td>
<td>0.383</td>
</tr>
<tr>
<td>P@5</td>
<td>0.4727</td>
<td>0.462</td>
<td>0.65</td>
</tr>
<tr>
<td>P@10</td>
<td>0.4045</td>
<td>0.425</td>
<td>0.59</td>
</tr>
<tr>
<td>P@20</td>
<td>0.3568</td>
<td>0.392</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 2. Comparison of the performances between the models AI, BM25 and BM25+Rocchio on TREC-8 collection
On TREC-7 (Table 1), the AI model obtains significant improvements of MAP of 32.6% and 10.5% over BM25 model and BM25+Rocchio model respectively. The R-Precision is also improved, the improvement is of 30.7% and 16.9% over BM25 model and BM25+Rocchio model respectively. We remark also significant improvements of P@5, P@10 and P@20. For example, the model AI achieves improvements of P@5 of 22% and 24.8% over BM25 model and BM25+Rocchio model respectively.

The results are similar for the TREC-8 collection (Table 2). Indeed, the AI model obtains significant improvements of R-Precision of 31.4% and 18.2% over BM25 model and BM25+Rocchio model respectively. The MAP is also improved, the improvement is of 30.1% and 12.6% over BM25 model and BM25+Rocchio model respectively. We remark also significant improvements of P@5, P@10 and P@20. The model AI achieves improvements of P@10 of 31.4% and 27.9% over BM25 model and BM25+Rocchio model respectively.

The results of the proposed approach in relation other approach that uses pseudo-irrelevant documents (BM25+Rocchio) are conceptual strength enhanced and consistent for five measures (MAP, R-Precision, P@5, P@10 and P@20) which are shown in the Tables 1 and 2 in both collections, TREC-7 and TREC-8. Our experimental results proved that our model outperforms the BM25 and BM25+Rocchio significantly$^2$.

5.5 Impact of Parameters

Three parameters in our model have been set. The first, $k$, is the number of top ranked relevant documents. The second, $p$, is the number of expansion terms. The second, $p$, is the number of expansion terms. The third one is the controlling parameter $\alpha$. For the irrelevant documents, in practice, among the 500 irrelevant documents, we select $p-1$ linearly independent documents.

In the following we will study the effect of varying the two RF parameters $k$ and $p$ and the parameter $\alpha$.

The experiments show that the variation of $k$ involves the variation of the performance of the AI model. Indeed, if $k$ exceeds 2, then the performance decreases.

The Figure 2 illustrates the variation of the value of R-Precision of the AI model according to $k$ (we fixed $p = 50$ and $\alpha = 0.3$).

![Figure 2. Impact of the parameter $k$ on the performance of AI model](image)

The experiments show also that the variation of the number of expansion terms $p$ involve the variation of the performance of the AI model. Indeed, if $p$ increases, then the performance increases also.

The Figure 3 illustrates the variation of the value of R-Precision of the AI model according to $p$ (we fixed $k = 2$ and $\alpha = 0.3$).

$^2$ Statistically significant improvement over BM25 and BM25+Rocchio models according to the Student t-test at the 0.05 level.
The experiments show also that the variation of the parameter involves the variation of the performance of the AI model.

The Figure 4 illustrates the variation of the value of R-Precision of the AI model according to \( k=2 \) and \( p=50 \).

From Figures 2, 3 and 4, for the TREC-7 and TREC-8 collections, the best R-Precision is obtained when the number of expansion terms \( p \) is 50, the number of relevant documents \( k \) is 2 and \( \alpha = 0.3 \). The best R-Precision are 0.355 and 0.383 on the TREC-7 and TREC-8 collections respectively.

6. Conclusion and Future Work

In this paper, we have introduced a novel approach to PRF using low-ranked documents. The main idea is to build an absorbing document, named \( \mathcal{A} \), as the cross product of linearly independent irrelevant documents and then add the non-zero weight terms of \( \mathcal{A} \) to the original query terms.

Our approach has shown its effectiveness with respect to a baseline system based on BM25 and the traditional combination of BM25 and Rocchio model.

Moreover, the evaluation has proved the robustness of the proposed strategy and its capability to select effective expansion terms. This result was duplicated on two test TREC collections (TREC-7 and TREC-8).

The main outcome of this work is that how absorbing document improves search accuracy for difficult queries?

In a previous work we used transition matrices (i.e. the algebraic operator responsible for changes of basis) to model relevance feedback [12]. In this paper we apply an other algebraic operator (cross product) to build a geometric PRF.

In a future work we intend to apply super linear algebra [10] to solve the problem of semantic relations between terms (such as synonymy, polysemy and so on).

![Figure 3](image.png)

Figure 3. Impact of the parameter \( p \) on the performance of AI model

References


