Linear Operators in Information Retrieval*

Hawete Hattab[1] and Rabeb Mbarek[2]

1 **Umm Al-qura University, Al-Jumum University College, Department of Mathematics, Makkah, KSA**
hshattab@uqu.edu.sa

2 **Sfax University, Multimedia Information Systems and Advanced Computing Laboratory, Sfax, Tunisia**
rabeb.hattab@gmail.com

Abstract

In this paper, we propose a pseudo-relevance feedback approach based on linear operators: vector space basis change and cross product. The aim of pseudo-relevance feedback methods based on vector space basis change IBM (Ideal Basis Method) is to optimally separate relevant and irrelevant documents. Whereas the aim of pseudo-relevance feedback method based on cross product AI (Absorption of irrelevance) is to effectively exploit irrelevant documents. We show how to combine IBM methods with AI methods. The combination methods IBM+AI are evaluated experimentally on two TREC collections (TREC-7 ad hoc and TREC-8 ad hoc). The experiments show that these methods improve previous works.

1998 ACM Subject Classification H.3.3 Information Search and Retrieval

Keywords and phrases Pseudo-relevance feedback, vector space basis change, Cross product

Digital Object Identifier 10.4230/OASIcs.SLATE.2017.23

1 Introduction

The main goal of an Information Retrieval System (IRS) is to find the subset of documents potentially most relevant to a given query. Most IRS compute a numeric score which measures the relevance of an object with respect to a query, and rank the objects according to this value. Several Information Retrieval (IR) models, including Vector Space Model (VSM) [15], probabilistic models [12] and language models [11], have been proposed to model this scoring function.

In the VSM, documents and queries are represented by vectors. Each component in a vector represents the weight of a term in the document and so the set of index terms (original vector space basis) generates documents and queries.

The idea of Relevance Feedback (RF) is to take the results that are initially returned from a given query and to use information about whether or not these results are relevant to perform a new query. The most commonly used RF methods aim to rewrite the user query. In the VSM, RF is usually undertaken by re-weighting the query terms without any modification in the vector space basis. With respect to the initial vector space basis (index terms), relevant and irrelevant documents share some terms (at least the terms of the query which selected these documents). The Vector Space Basis Change (VSBC) is the algebraic operator responsible for change of basis and it is parameterized by a transition matrix. If

* This work was supported financially by the Deanship of Scientific Research at Umm Al-Qura University to Dr. Hawete Hattab (Grant Code: 15-COM-3-1-0018).

we change the vector space basis, then each vector component changes depending on this matrix. According to [9], the VSBC causes vector behavior changes. The best framework that could make the VSBC technique into application is RF: the user selects relevant and irrelevant documents in an initial ranking and instead of reformulating the query, we change the vector space basis in which it is written (as well as the documents). The strategy of VSBC has been shown to be effective in separating relevant document and irrelevant ones. Recently, using this strategy, some feedback algorithms have been developed [6, 8, 9]. These techniques are called IBM (Ideal Basis Model).

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 [4]. 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 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 [7], Mbarek et al. proposed to solve this problem by exploiting the role of irrelevant documents in selecting better expansion terms from the top-ranked documents. In particular they built an absorbing document which is the cross product of linearly independent irrelevant documents. This document is orthogonal to irrelevant ones. This method is called Absorption of Irrelevance (AI).

In [10], Mbarek et al. investigated the role of irrelevant documents in document re-ranking. This proposed re-ranking strategy is based on a negative RF approach which takes into account irrelevant documents in the initial document ranking. The key idea behind this approach is to use the absorbing document [7], which is orthogonal to irrelevant documents, to re-rank documents on the ground of their similarity with respect to the absorbing document.

Score vectors from two different scoring methods can be combined to yield a new score vector, and thereby a new scoring method. If the two scoring methods have complementary advantages, the combined scoring method may perform better than either scoring method alone. In this paper, we propose to combine IBM with AI methods.

The paper is structured as follows. Section 2 describes the two strategies IBM and AI and also the combination methods IBM+AI. Experiments performed for evaluating our combined approaches are presented in Section 3. The last section concludes.

2 Linear Operators in information retrieval

2.1 Ideal Basis Methods: IBM

In the IBM approches, the optimal matrix M^* puts the relevant documents gathered to their centroïd g_R and the irrelevant documents far from it. g_R is done by:

$$g_R = \frac{1}{|R|} \sum_{d \in R} d$$

where R is the set of relevant documents.

The optimal matrix M^* should minimize the sum of squared distances between each relevant document and g_R i.e.:

$$M^* = \underset{M \in M_n(\mathbb{R})}{\arg\min} \sum_{d \in R} (d - g_R)^T . M^T M . (d - g_R) \,. \tag{1}$$

By the same, matrix M^* should maximize the sum of squared distances of each irrelevant document and g_R, which leads on the following:

$$M^* = \underset{M \in M_n(\mathbb{R})}{\arg\max} \sum_{d \in S} (d - g_R)^T . M^T M . (d - g_R) \tag{2}$$

where S is the set of irrelevant documents.

According to [6] (IBM1 method), the optimal matrix M^* should minimize the quotient and so equations 1 and 2 result on the following single equation:

$$M^* = \underset{M \in M_n(R)}{\arg\min} \frac{\sum_{d \in R} (d - g_R)^T . M^T M . (d - g_R) + \theta}{\sum_{d \in S} (d - g_R)^T . M^T M . (d - g_R) + \theta} \tag{3}$$

where θ is a real parameter.

According to [8] (IBM2 method), the optimal matrix M^* should maximize the difference and so equations 1 and 2 result on the following single equation:

$$M^* = \underset{M \in M_n(\mathbb{R})}{\arg\max} \left(\sum_{d \in S} (d - g_R)^T . M^T M . (d - g_R) - \sum_{d \in R} (d - g_R)^T . M^T M . (d - g_R) \right) . \tag{4}$$

In [9], Mbarek et al. built the transition matrix M^* using an algebraic method (IBM3 method). In the IBM3 method, if $d \in R$, the optimal matrix M^*, which satisfies equation 1, should contract the vector $d - g_R$ which implies that there exists a real parameter $0 < \gamma < 1$ such that:

$$M(d - g_R) = \gamma(d - g_R) .$$

Then $d - g_R$ is an eigenvector of the matrix M associated to the eigenvalue γ.

If $d \in S$, the optimal matrix M^*, which satisfies equation 2, should dilate the vector $d - g_R$ which implies that:

$$M(d - g_R) = (1 + \gamma)(d - g_R) .$$

Then $d - g_R$ is an eigenvector of the matrix M associated to the eigenvalue $1 + \gamma$. M^* is a diagonalized matrix (similar to a diagonal matrix) having two distinct eigenvalues γ and $1 + \gamma$. Therefore:

$$M = V.D.V^{-1} \tag{5}$$

where D is the eigenvalues matrix of M, and V is the eigenvectors Matrix of M.

2.2 Absorption of Irrelevance: AI

One of the main problem in information retrieval is how to exploit effectively the irrelevant documents? To give a satisfactory answer we need to build a solid structure which represents the irrelevant content. In the theory of vector space, the cross product is a very suitable candidate. Mbarek et al. used this approach in [10].

This section describes the PRF approach based on irrelevant documents. The main idea is to build an absorbing document noted $\widetilde{d}$, 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:

$$Q_{new} = \delta.Q_{int} + (1 - \delta).\widetilde{d} \tag{6}$$

where δ is a real parameter. Note that if $\delta = 0$ we obtain a re-ranking method studied in [10].

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 D_{init}. Let m 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 $\widetilde{d}$ which is the cross product of u_1,...,u_m, each vector must be written as a linear combination of $m + 1$ indexing terms [10]. For this reason $p = m + 1$.

The absorbing document is $\widetilde{d} = u_1 \wedge ... \wedge u_m$. By [10], $\widetilde{d}$ is orthogonal to each irrelevant document.

2.3 Combination methods IBM+AI

This section describes the PRF combination methods IBM+AI. We propose a novel PRF approach based on linear operators: vector space basis change and cross product, i.e. we propose to combine the IBM methods and AI method.

Score vectors from two different scoring methods (IBM and AI) can be combined to yield a new scoring method (IBM+AI). If the two scoring methods have complementary advantages, the combined scoring method may perform better than either scoring method alone. Note that VSBC-based methods IBM and cross product based method AI have complementary advantages. Indeed, the main goal of VSBC-based methods IBM is to optimally separate relevant and irrelevant documents and the main goal of cross product based method AI is to build an absorbing document, named $\widetilde{d}$, as the cross product of linearly independent irrelevant documents selected from the bottom. For these reasons, we propose to combine IBM and AI methods in the following way:

$$Q_{new} = \alpha.M^{*T}.M^*.Q_{int} + (1 - \alpha).\widetilde{d} \tag{7}$$

where M^* is the transition matrix (Equations 3, 4 and 5), $\widetilde{d}$ is the absorbing document and α is a real parameter.

3 Experiments

The main goal of our experiments is to investigate whether the new combination methods IBM+AI described above perform better than BM25, BM25+Rocchio and linear operators based approaches (IBM and AI).

3.1 Evaluation Methodology

We set up a baseline system based on the BM25 formula proposed in [13]. BM25 parameters are b and k_1. The TREC-7 ad hoc and TREC-8 ad hoc collections were used for test. They consist of the same set of documents (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. Note that we choose the TREC-7 ad hoc and TREC-8 ad hoc test collections as they have the highest frequency of scores reported for ad hoc retrieval in recent years [1].

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. This strategy is widely used in IR [13, 3].

The experiments consist of re-ranking the results of the Baseline Model. For our combined approach IBM+AI the reformulated query is done in Equation 7.

We compare our approach IBM+AI to the baseline model BM25 and to the traditional combination of BM25 and Rocchio's feedback model[1] (BM25+Rocchio).

The following improved version [16] of the original Rocchio's formula [14] is used:

$$Q_{new} = \lambda.Q_{int} + \beta.\frac{1}{|D^+|}\sum_{d \in D^+} d\,. \tag{8}$$

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

For the linear operators based approaches and the BM25+Rocchio model, the retrieved documents are re-ranked by the inner product done by:

$$< Q_{new}, d >= Q_{new}^T.d\,. \tag{9}$$

3.2 Parameter settings

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

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, and vary the parameters θ in Equation 3, γ in Equation 5, δ in Equation 6, α in Equation 7 and β in Equation 8 from 0 to 1 in steps of 0.1.

The models IBM, AI, IBM+AI and 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 $p \in \{10, 20, 30, 50\}$. The number m of linearly independent irrelevant documents must be equal to $p - 1$[2].

3.3 Results

To evaluate the performance of our approaches IBM+AI 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 [2].

We present here the behavior of evaluation mesures on TREC-7 and TREC-8 for all models. The optimal results are illustrated in Tables 1 and 2.

[1] According to [17], BM25 [13] term weighting coupled with Rocchio feedback remains a strong baseline.
[2] In a vector space of dimension n, we compute the cross product of $n - 1$ vectors.

Table 1 Comparison of the performance on TREC-7 collection.

	BM25+Roc	AI	IBM1	IBM2	IBM3	IBM1+AI	IBM2+AI	IBM3+AI
MAP	0.253	0.283	0.263	0.269	0.276	0.296	0.311	0.323
R-Prec	0.295	0.355	0.317	0.321	0.327	0.365	0.374	0.387
P@5	0.451	0.600	0.469	0.471	0.477	0.622	0.638	0.649
P@10	0.440	0.560	0.460	0.480	0.510	0.580	0.590	0.598
P@20	0.381	0.513	0.431	0.452	0.463	0.526	0.544	0.570
Para	β=1 k=2 p=30	k=2 p=50 $\delta = 0.3$ $m = 49$	$k = 2$ $p = 50$ $\theta = 0.3$ $m = 49$	$k = 2$ $p = 50$ $m = 49$	$k = 3$ $p = 50$ $\gamma = 0.4$ $m = 49$	$k = 2$ $p = 50$ $\alpha = 0.3$ $m = 49$	$k = 2$ $p = 50$ $\alpha = 0.3$ $m = 49$	$k = 3$ $p = 50$ $\alpha = 0.3$ $m = 49$

Table 2 Comparison of the performance on TREC-8 collection.

	BM25+Roc	AI	IBM1	IBM2	IBM3	IBM1+AI	IBM2+AI	IBM3+AI
MAP	0.264	0.302	0.281	0.284	0.289	0.311	0.324	0.334
R-Prec	0.313	0.383	0.356	0.363	0.371	0.388	0.390	0.398
P@5	0.462	0.650	0.482	0.484	0.487	0.663	0.670	0.685
P@10	0.425	0.590	0.454	0.465	0.471	0.610	0.646	0.654
P@20	0.392	0.530	0.421	0.431	0.442	0.543	0.568	0.590
Para	β=1 k=2 p=30	k=2 p=50 $\delta = 0.3$ $m = 49$	$k = 2$ $p = 50$ $\theta = 0.3$ $m = 49$	$k = 2$ $p = 50$ $m = 49$	$k = 3$ $p = 50$ $\gamma = 0.5$ $m = 49$	$k = 2$ $p = 50$ $\alpha = 0.3$ $m = 49$	$k = 2$ $p = 50$ $\alpha = 0.3$ $m = 49$	$k = 3$ $p = 50$ $\alpha = 0.3$ $m = 49$

Tables 1 and 2 show the performance of BM25, BM25+Rocchio and various methods based on linear operators on TREC-7 ad hoc and TREC-8 ad hoc collections. For all methods, the best parameter values are shown.

The first two columns of Tables 1 and 2 report scores for BM25 and for the traditional combination BM25+Rocchio. It is evident that BM25+Rocchio outperforms BM25 which proves that the BM25 term weighting coupled with Rocchio feedback remains a strong baseline [17]. The traditional combination BM25+Rocchio is the standard Rocchio model using BM25 weights. These weights use the b and k_1 parameters that are optimal for the BM25 method. When BM25 is used as a weighting scheme prior to Rocchio, the optimal parameters may be different, but fixed parameters are used for the sake of simplicity.

Performance results for the VSBC-based methods IBM1, IBM2 and IBM3 in Tables 1 and 2 are based on the computing of the transition matrices M^* (Equations 4, 3 and 5). These matrices use control and feedback parameters. Table 1 and 2 show that these methods outperform the BM25 and BM25+Rocchio methods. In particular, method IBM3, which uses the algebraic properties of the transition matrix, outperforms IBM1 and IBM2.

Performance results for the cross product based method AI, in the third column of Tables 1 and 2, is based on the computing of the cross product of the irrelevant documents $\widetilde{d}$ [10]. Tables 1 and 2 show that this method outperforms the BM25, BM25+Rocchio methods and the VSBC-based methods IBM.

Tables 1 and 2 also show the performance of combination methods IBM1+AI, IBM2+AI and IBM3+AI. For each dataset and combination method, the best values of parameters are shown. The three combination methods, first, outperform BM25 and BM25+Rocchio models. Second, they outperform VSBC-based methods IBM1, IBM2 and IBM3 and also the cross product based method AI. The improvements are big and of practical importance for all evaluation measures.

The results of the proposed approaches IBM+AI are superior and consistent for five measures (MAP, R-Precision, P@5, P@10 and P@20) showed in Tables 1 and 2 in both

collections, TREC-7 and TREC-8. The experimental results prove that our combination methods outperform the BM25, BM25+Rocchio, IBM1, IBM2, IBM3 and AI methods significantly[3].

3.4 Discussion

Three parameters in our combination methods have been set. The first, k, is the number of top-ranked relevant documents. The second, p, is the number of expansion terms. The third one is the controlling parameter α. For the irrelevant documents, in practice, we select $m = p - 1$ linearly independent documents.

In the following we will study the impact of varying the two RF parameters k and p and the parameter α .

The experiments show that the variation of k involves the variation of the performance of each IBM+AI method. Indeed, if k exceeds 3, then the performance decreases.

The experiments also show that the variation of the number of expansion terms p involves the variation of the performance of each IBM+AI method. Indeed, if p increases, then the performance increases also. Indeed, if p increases, using the relation: $m = p - 1$, then m increases, and so we can control more irrelevant documents.

The experiments also show that the variation of the parameter α involves the variation of the performance of each IBM+AI method.

4 Conclusion and future work

In this paper, the combination approaches has shown their effectiveness with respect to a baseline system based on BM25 and the traditional combination of BM25 and Rocchio model. Moreover, the evaluation has proved that the combination approaches may perform better than either scoring method alone. These results were duplicated on two test TREC collections (TREC-7 ad hoc and TREC-8 ad hoc).

The main outcome of this work is that how linear operator based methods improve search accuracy for difficult queries?

In this paper we apply linear operator to build a geometric PRF. In a future work we intend to apply super linear algebra [5] to solve the problem of semantic relations between terms (synonymy, polysemy. . .).

Acknowledgements. The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University for the continuous support.

References

1 Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 601–610, 2009.

2 Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu. A geometric interpretation of r-precision and its correlation with average precision. In *28th Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 573–574, 2005.

[3] Statistically significant improvement according to the Student t-test at the 0.05 level.

3 Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. Negation for document re-ranking in ad-hoc retrieval. In *Third International Conference on Advances in Information Retrieval Theory (ICTIR)*, pages 285–296, 2011.

4 Efthimis N. Efthimiadis and Paul V. Biron. Ucla-okapi at TREC-2: query expansion experiments. In *Second Text REtrieval Conference (TREC)*, pages 278–290, 1993.

5 W. B. Vasantha Kandasamy and Florentin Smarandache. *Super Linear Algebra*. InfoLearnQuest, Ann Arbor, 2008. Available in arXiv:0807.3013.

6 Rabeb Mbarek and Mohamed Tmar. Relevance feedback method based on vector space basis change. In *19th International Symposium String Processing and Information Retrieval (SPIRE)*, pages 342–347, 2012.

7 Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab Mohand Boughanem. On improving pseudo-relevance feedback using an absorbing document. *International Journal of Web Applications (IJWA)*, 8(1):8–15, 2016.

8 Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab. A new relevance feedback algorithm based on vector space basis change. In *15th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 355–366, 2014.

9 Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab. Vector space basis change in information retrieval. *Computación y Sistemas*, 18(3), 2014.

10 Rabeb Mbarek, Mohamed Tmar, Hawete Hattab, and Mohand Boughanem. A re-ranking method based on irrelevant documents in ad-hoc retrieval. In *5th Symposium on Languages, Applications and Technologies (SLATE)*, pages 2:1–2:10, 2016.

11 Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *21st Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 275–281, 1998.

12 Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 232–241, 1994.

13 Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at trec. In *The First Text REtrieval Conference*, pages 21–30, 1992.

14 Joseph Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *Information Storage and Retrieval: Scientific Report No. ISR-9*, chapter 23. The National Science Foundation, 1965.

15 Gerard Salton. *Automatic Information Organization and Retrieval*. McGraw Hill Text, 1968.

16 Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. In Karen Sparck Jones and Peter Willett, editors, *Readings in Information Retrieval*, pages 355–364. Morgan Kaufmann Publishers Inc., 1997.

17 Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Tenth International Conference on Information and Knowledge Management*, pages 403–410, 2001.