Rocchio's Model Based on Vector Space Basis Change for Pseudo Relevance Feedback

Rabeb Mbarek[1], Mohamed Tmar[2], and Hawete Hattab[3]

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

2 **Sfax University**
Multimedia Information systems and Advanced Computing Laboratory
Sfax, Tunisia
mohamedtmar@yahoo.fr

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

Abstract

Rocchio's relevance feedback model is a classic query expansion method and it has been shown to be effective in boosting information retrieval performance. The main problem with this method is that the relevant and the irrelevant documents overlap in the vector space because they often share same terms (at least the terms of the query). With respect to the initial vector space basis (index terms), it is difficult to select terms that separate relevant and irrelevant documents. The Vector Space Basis Change is used to separate relevant and irrelevant documents without any modification on the query term weights. In this paper, first, we study how to incorporate Vector Space Basis Change into the Rocchio's model. Second, we propose Rocchio's models based on Vector Space Basis Change, called VSBCRoc models. Experimental results on a TREC collection show that our proposed models are effective.

1998 ACM Subject Classification H.3.3 Information Search and Retrieval

Keywords and phrases Rocchio model, vector space basis change, pseudo relevance feedback

Digital Object Identifier 10.4230/OASIcs.SLATE.2014.215

1 Introduction

In the Vector Space Model (VSM), each component of the vector represents a term in the document [18] i.e. each component in the vector represents the weight of the term in the document. The set of all index terms is called the original vector space basis. For the most vector space based Information Retrieval (IR) and feedback models, the original vector space basis generates documents and queries. Although several term weighting and feedback methods have been proposed, only a few approaches [4, 11, 8, 9] consider that changing the vector space basis from the original vector space basis into another basis is an issue of investigation. The Vector Space Basis Change (VSBC) consists of using a transition matrix[1]. By changing the vector space basis, each vector coordinate changes depending on this matrix. If we change the basis, then the inner product changes and so the Cosine

[1] The algebraic operator responsible for change of basis.

© Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab;
licensed under Creative Commons License CC-BY

3^{rd} Symposium on Languages, Applications and Technologies (SLATE'14).
Editors: Maria João Varanda Pereira, José Paulo Leal, and Alberto Simões; pp. 215–224
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

function behavior changes [10]. By the same Dice, Jaccard and Overlap functions behavior changes.

Pseudo Relevance Feedback (PRF) is known as a useful method for enhancing retrieval performance. It assumes that the top-ranked n documents (pseudo-documents) of the initial retrieval are relevant and extracts expansion terms from them. PRF has been shown to be effective in improving IR performance [2, 3, 6, 7, 13, 14, 16, 17, 19, 20, 21]. PRF can also fail in some cases. For example, when some pseudo-documents contain terms of the irrelevant contents, then these terms misguide the feedback models by importing noisy terms into the queries. This could influence the retrieval performance in a negative way.

With respect to the original vector space basis, relevant and irrelevant documents share some terms (at least the terms of the query which selected these documents). To avoid this problem, it suffice to separate relevant and irrelevant documents. VSBC is an effective method for the separation of relevant and irrelevant documents. This method has been studied in the past few years [4, 11, 8, 9]. In [8, 9], the authors have been found a basis which gathers the relevant documents and the irrelevant ones are kept away from the relevant ones. These approaches have been evaluated on a Relevance Feedback (RF) framework.

Rocchio's model [16] is a classic framework for implementing (pseudo) RF via improving the query representation. It models a way of incorporating (pseudo) RF information into the VSM in IR. In this paper, first, we study how to incorporate VSBC into the Rocchio's model. Second, we propose Rocchio's models based on the VSBC, called VSBCRoc models.

This paper is organized as follows. section 2 presents the related work. Sections 3 describes our approach based on the VSBC. In Section 4, the experimental results are presented and discussed. A direct comparison is made to compare VSBCRoc models with the classical Rocchio's model. Finally, we conclude our work with a brief conclusion and future research directions in Section 5.

2 Related Work

The VSM [18] is adopted to rank the documents. This model showed good feedback performance on most collections whereas the probabilistic model had problems with some collections [5].

2.1 Vector Space Basis Change

The Latent Semantic Indexing (LSI) [4] exploits the hypothesis that the term-document frequency matrix encloses information about the semantic relations between terms and between documents. This technique is based on Singular Value Decomposition (SVD) aiming at decomposing the matrix and disclosing the principal components used to represent fewer independent concepts than many inter-dependent index terms. This method results on a new vector space basis, with a lower dimension than the original one (all index terms), and in which each component is a linear combination of the indexing terms.

When using a term to express a query or a document, the user gave to the term a semantics which is different from the semantics of the same term used by another user or by the same user in another place, time, need. In other words, the use of a term depends on context. Therefore, context influences the selection of the terms, their semantics and inter-relationships. A vector space basis models a document or query terms. The semantics of a document or query term depends on context. A vector space basis can be derived from a context. Therefore, a vector space basis of a vector space is the construct to model context.

Also, change of context can be modeled by linear transformations from one base to another which is a VSBC [10, 11].

Recently, Mbarek et al. [8, 9] developed a RF algorithms based on a vector space basis change. These RF algorithms improve the results of known models (BM25 model, Rocchio model). They built a basis which gives a better representation of documents. This basis should minimize the sum (S_1) of squared distances between each relevant document and g_R (g_R is the centroïd of relevant documents) and should maximize the sum (S_2) of squared distances between each irrelevant document and g_R. And so this basis should minimize the quotient $\frac{S_1}{S_2}$ [8] and maximize the difference $S_2 - S_1$ [9].

2.2 Pseudo-Relevance Feedback

In IR, PRF via query expansion is referred to as the techniques that reformulate the original query by adding new terms into the query, in order to achieve a better retrieval performance. There are a large number of studies on the topic of PRF. Here we mainly review the work about PRF which is the most related to our research. A classical RF technique was proposed by Rocchio in 1971 for the Smart retrieval system [16]. It is a framework for implementing (pseudo) RF via improving the query representation, in which a set of documents are utilized as the feedback information. Unique terms in this set are ranked in descending order of their *tf* ∗ *idf* weights. In the following decades, many other RF techniques and algorithms were developed, mostly derived under Rocchio's framework. A popular and successful automatic PRF algorithm was proposed by [14] in the Okapi system; Amati et al. [1] proposed a query expansion algorithm in his divergence from randomness retrieval framework; Carpineto et al. [2] proposed to compute the weight of candidate expansion terms based on the divergence between the probability distributions of terms in the top ranked documents and the whole collection; Miao et al. [12] studied how to incorporate proximity information into the Rocchio's model, and proposed three proximity based Rocchio's models.

In this paper, first, we will incorporate VSBC into the Rocchio's model. Second, we propose Rocchio's models based on VSBC, called VSBCRoc models.

3 Rocchio's Models based on Vector Space Basis Change

3.1 Rocchio's Formula

Rocchio's model [16] is a classic framework for implementing (pseudo) RF via improving the query representation. It models a way of incorporating (pseudo) relevance feedback information into the VSM in IR. In case of PRF, the Rocchio's model (without considering negative feedback documents) has the following steps:

- All documents are ranked for the given query using a particular retrieval model. This step is called initial retrieval, from which the $|R|$ highest ranked documents are used as the feedback set.
- The representation of the query is finally refined by taking a linear combination of the initial query term vector with the feedback document vector, this initial formula is denoted by VSBCRoc1:

$$\mathit{VSBCRoc1} : Q_1 = \alpha * Q_0 + \beta * \sum_{d \in R} \frac{d}{|R|} \tag{1}$$

where Q_0 represents the original query vector, Q_1 represents the first iteration query vector, d is the document weight vector, and α and β are tuning constants controlling how much we rely on the original query and the feedback information. In practice, we can always fix α at 1, and only study β in order to get better performance.

3.2 Vector Space Basis Change

In [8, 9], the authors built a new vector space basis which separates relevant and irrelevant documents without any modification on the query term weights. That is, this basis gathers the relevant documents and the irrelevant ones are kept away from the relevant ones. It can be viewed as a representation that keeps the relevant documents gathered to their centroïd and the irrelevant ones far from it. Each document d_i is represented in a vector space by $d_i = (w_{i1}, w_{i2}, ...w_{iN})^T$ where w_{ij} is the weight of term t_j in document d_i and N is the number of index terms[2]. As for us our approach is independent of the term weighting method.

The Euclidian distance between documents d_i and d_j is given by:

$$\begin{aligned} dist(d_i, d_j) &= \sqrt{\sum_{k=1}^{N} (w_{ik} - w_{jk})^2} \\ &= \sqrt{(w_{i1} - w_{j1}...w_{iN} - w_{jN}) \cdot (w_{i1} - w_{j1}...w_{iN} - w_{jN})^T} \\ &= \sqrt{(d_i - d_j)^T \cdot (d_i - d_j)}\,. \end{aligned}$$

By changing the basis using a transition matrix M, the distance between 2 vectors d_i^* and d_j^* which are respectively d_i and d_j rewritten in the new basis is given by:

$$\begin{aligned} dist(d_i^*, d_j^*) &= dist(M.d_i, M.d_j) \\ &= \sqrt{(M.d_i - M.d_j)^T \cdot (M.d_i - M.d_j)} \\ &= \sqrt{(d_i - d_j)^T \cdot M^T M \cdot (d_i - d_j)}\,. \end{aligned}$$

The vector space basis which optimally separates relevant and irrelevant documents is represented by a matrix M^* called the optimal transition 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.

By the same, using a transition matrix M, we obtain:

$$M.g_R = M \cdot (\frac{1}{|R|} \sum_{d \in R} d) = \frac{1}{|R|} \sum_{d \in R} M \cdot d\,.$$

[2] x^T is the transpose of x

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

$$\begin{aligned} M^* &= \arg \min_{M \in M_n(\mathbb{R})} \sum_{d \in R} dist^2(M \cdot d, M \cdot g_R) \qquad (2) \\ &= \arg \min_{M \in M_n(\mathbb{R})} \sum_{d \in R} (Md - Mg_R)^T \cdot (Md - Mg_R) \\ &= \arg \min_{M \in M_n(\mathbb{R})} \sum_{d \in R} (d - g_R)^T \cdot M^T M \cdot (d - g_R)\,. \end{aligned}$$

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

$$\begin{aligned} M^* &= \arg \max_{M \in M_n(\mathbb{R})} \sum_{d \in S} dist^2(M \cdot d, M \cdot g_R) \qquad (3) \\ &= \arg \max_{M \in M_n(\mathbb{R})} \sum_{d \in S} (Md - Mg_R)^T \cdot (Md - Mg_R) \\ &= \arg \max_{M \in M_n(\mathbb{R})} \sum_{d \in S} (d - g_R)^T \cdot M^T M.(d - g_R) \end{aligned}$$

where S is the set of irrelevant documents.

In [8], the authors have been showed that the Equations 2 and 3 result on the following single equation:

$$M^* = \arg \min_{M \in M_n(\mathbb{R})} \frac{\sum_{d \in R} (d - g_R)^T \cdot M^T M \cdot (d - g_R) + \alpha}{\sum_{d \in S} (d - g_R)^T \cdot M^T M \cdot (d - g_R) + \alpha} \qquad (4)$$

where α is real parameter close to 0.

And in [9], the authors have been showed that the Equations 2 and 3 result on the following single equation:

$$M^* = \arg \max_{M \in M_n(\mathbb{R})} [\sum_{d \in S} (d - g_R)^T \cdot M^T M \cdot (d - g_R) - \sum_{d \in R} (d - g_R)^T \cdot M^T M \cdot (d - g_R)]\,. \qquad (5)$$

Let M_1^* be a solution of Equation 4 and M_2^* be a solution of Equation 5. These matrices separate relevant and irrelevant documents. The proposed Rocchio's models based on VSBC are:

$$VSBCRoc2 : Q_2 = Q_0 + \beta * \sum_{d \in R} \frac{M_1^* d}{|R|} \qquad (6)$$

$$VSBCRoc3 : Q_3 = Q_0 + \beta * \sum_{d \in R} \frac{M_2^* d}{|R|} \qquad (7)$$

We remark that the initial Rocchio's formula, VSBCRoc1 (Equation 1), corresponds to incorporating the identity matrix[3] (there is no basis change).

[3] A square matrix with ones on the main diagonal and zeros elsewhere.

4 Experiments

In this section we give the different experiments and results obtained to evaluate our approach. We describe the environnement of evaluation and the experimental conditions.

4.1 Environment

The test collection TREC-7 was used for the experiments in this article. Data was preprocessed through stop-word removal and Porter's stemming, and one–word terms were stored; the initial rankings of documents (Baseline Model) were weighted by the $BM25$ formula proposed in [15]. BM25 parameters are $b = 0.5$, $k_1 = 1.2$, $k_2 = 0$ and $k_3 = 8$.

- The initial query Q_0 is made from the short topic description, and using it the top 1000 documents are retrieved from the collections (weighted $\alpha = 1$).
- R is the set of top ranking n documents, assumed to be relevant.
- S is the set of retrieved documents $501 - 1000$, assumed to be irrelevant.

For the three approaches, the retrieved documents are ranked by the inner product done by:

$$RSV(Q_i, d) = Q_i^T \cdot d \qquad 1 \leq i \leq 3 \tag{8}$$

4.2 Results

To evaluate the performance we execute several runs using the topics provided by TREC. In detail, the TREC-7 collection has 50 topics. Topics are structured in three fields: title, description and narrative. To generate a query, the title of a topic was used, thus falling into line with the common practice of TREC experiments; description and narrative were not used.

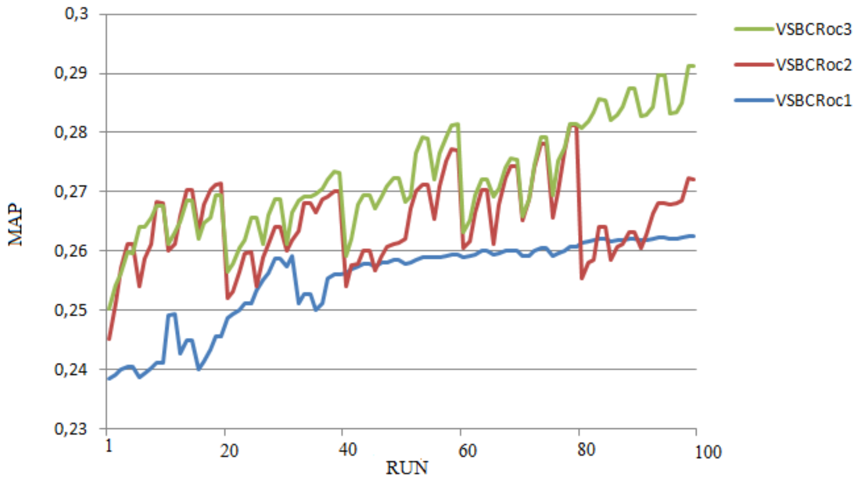
We perform 100 runs by considering all possible combinations of the three parameters involved in the three models. In particular, we take into account: n (the cardinality of R), m (the number of expansion terms) and the parameter β (used for the linear combination): see Equations 1, 6 and 7. We select different ranges for each parameter: n ranges in $(1, 2, 3, 4, 5)$, m ranges in $(10, 20, 30, 50)$ and β ranges in $(0.1, 0.2, 0.5, 1, 2)$.

We evaluate each run in terms of Mean Average Precision (MAP). The experiments and the evaluations are articulated around the comparison between VSBCRoc1, VSBCRoc2 and VSBCRoc3.

Figure 1 plots the MAP values for each run and approach: VSBCRoc1 is the original Rocchio model, VSBCRoc2 and VSBCRoc3 are the new Rocchio models obtained by incorporating the VSBC strategy. These graphs highlights as the system performance vary according to parameters changes. It is possible to note that:

- VSBCRoc2 and VSBCRoc3 models have better performance than VSBCRoc1 model.
- The MAP value of VSBCRoc1 is similar for $\beta = 1$ and $\beta = 2$ (the same remark for VSBCRoc2 and VSBCRoc3).
- The MAP values of VSBCRoc1, VSBCRoc2 and VSBCRoc3 increase if the number of expansion terms increase.
- The MAP values of VSBCRoc1 and VSBCRoc3 increase if the number of pseudo-documents increase.

Figure 1 Plot of MAP values on TREC-7.



For the VSBCRoc1, VSBCRoc2 and VSBCRoc3 models, the lowest MAP value is 0.2385, 0.2451 and 0.2503, respectively. This value occurs when only one relevant document and 10 expansion terms are involved. The highest MAP value for VSBCRoc1 is 0.2625, while for VSBCRoc3 is 0.2913. Both values are obtained with 5 relevant documents and 50 expansion terms. The highest MAP value for VSBCRoc2 is 0.2813. This value occurs when 4 relevant documents and 50 expansion terms are involved.

4.3 Significance of Our Results

Statistical significance is the probability that an effect is not due to just chance. These tests are based on a pre-specified low probability threshold called p-values. P-values are always coupled to a significance level, usually at 0.05. Thus, if a p-value was found to be less than 0.05, then the result would be considered statistically significant. To study the statistical significance of our result we use a free software environment, R, for statistical computing and graphics[4]. Before applying the student's t-test we compute a R data frame in which each row has a measurement and a categorical system identifier.

Listing 1 t-test of significance of the difference of results of VSBCRoc1 and VSBCRoc2.

```
>  MAP<-c(0.2385,0.2392,0.2401,...,0.2685,0.2723,0.2722)
> Sys<-c("VSBCRoc1","VSBCRoc1","VSBCRoc1","VSBCRoc1",...,"VSBCRoc2",
"VSBCRoc2","VSBCRoc2")
> X<-data.frame(MAP=MAP,Sys=Sys)
> X
     MAP      Sys
1   0.2385  VSBCRoc1
2   0.2392  VSBCRoc1
3   0.2401  VSBCRoc1
.     .         .
.     .         .
.     .         .
```

[4] http://www.r-project.org/

```
100 0.2625 VSBCRoc1
101 0.2451 VSBCRoc2

.    .        .
.    .        .
.    .        .
198 0.2685 VSBCRoc2
199 0.2723 VSBCRoc2
200 0.2722 VSBCRoc2
> t.test(MAP ~ Sys, paired=T, data=X)

        Paired t-test

data:  MAP by Sys
t = -11.7418, df = 99, p-value < 2.2e-16
```

Listing 2 t-test of significance of the difference of results of VSBCRoc1 and VSBCRoc3.

```
>  MAP<-c(0.2385,0.2392,0.2401,...,0.2851,0.2913,0.2913)
> Sys<-c("VSBCRoc1","VSBCRoc1","VSBCRoc1",...,"VSBCRoc3",
"VSBCRoc3","VSBCRoc3")
> X<-data.frame(MAP=MAP,Sys=Sys)
> X
     MAP      Sys
1   0.2385 VSBCRoc1
2   0.2392 VSBCRoc1
3   0.2401 VSBCRoc1
.    .        .
.    .        .
.    .        .
100 0.2625 VSBCRoc1
101 0.2503 VSBCRoc3
.    .        .
.    .        .
.    .        .
198 0.2851 VSBCRoc3
199 0.2913 VSBCRoc3
200 0.2913 VSBCRoc3
> t.test(MAP ~ Sys, paired=T, data=X)

        Paired t-test
data:  MAP by Sys
t = -26.4026, df = 99, p-value < 2.2e-16
```

Listing 3 t-test of significance of the difference of results of VSBCRoc2 and VSBCRoc3.

```
>  MAP<-c(0.2451,0.2511,0.2572,...,0.2851,0.2913,0.2913)
> Sys<-c("VSBCRoc2","VSBCRoc2","VSBCRoc2",...,"VSBCRoc3",
"VSBCRoc3","VSBCRoc3")
> X<-data.frame(MAP=MAP,Sys=Sys)
> X
     MAP      Sys
1   0.2451 VSBCRoc2
2   0.2511 VSBCRoc2
3   0.2572 VSBCRoc2
```

```
.    .        .
.    .        .
.    .        .
100  0.2722  VSBCRoc2
101  0.2503  VSBCRoc3
.    .        .
.    .        .
.    .        .
198  0.2851  VSBCRoc3
199  0.2913  VSBCRoc3
200  0.2913  VSBCRoc3
> t.test(MAP ~ Sys, paired=T, data=X)

        Paired t-test
data:  MAP by Sys
t = -8.6917, df = 99, p-value = 7.741e-14
```

In listings 1, 2 and 3 we have the p-values < 0.05, then our results are statistical significant.

5 Conclusion

The main problem with Rocchio's approach is that the relevant and the irrelevant documents overlap in the vector space because they often share same terms (at least those of the query). Therefore it is difficult to select terms that separate relevant and irrelevant documents which cause the query drift problem (Croft and Harper). To guide the RF process, the authors of [8, 9] have been computed a vector space basis which gives a better representation of the documents such that the relevant documents are gathered and the irrelevant ones are kept away from the relevant documents. Vector space basis change discriminates irrelevant documents from relevant ones, thus reducing the potential noise in the vector space after produced by query expansion. The combinations of Rocchio's models with vector space basis change improve the results of classic Rocchio's formula.

This paper reports about incorporating transition matrix (i.e. the algebraic operator responsible for change of basis) into the classic Rocchio's model. We intend to incorporate other algebraic operator (like vector product) into the classic Rocchio's model.

References

1 G. Amati. *Probabilistic models for information retrieval based on divergence from randomness*. PhD thesis, Department of Computing Science, University of Glasgow, 2003.

2 Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, January 2001.

3 Kevyn Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM'09, pages 837–846, New York, NY, USA, 2009. ACM.

4 Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

5 Donna Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'92, pages 1–10, New York, NY, USA, 1992. ACM.

6 Xiangji Huang, Yan Rui Huang, Miao Wen, Aijun An, Yang Liu, and J. Poon. Applying data mining to pseudo-relevance feedback for high performance text retrieval. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 295–306, Dec 2006.

7 Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'01, pages 120–127, New York, NY, USA, 2001. ACM.

8 Rabeb Mbarek and Mohamed Tmar. Relevance feedback method based on vector space basis change. In *Proceedings of the 19th International Conference on String Processing and Information Retrieval*, SPIRE'12, pages 342–347, Berlin, Heidelberg, 2012. Springer-Verlag.

9 Rabeb Mbarek, Mohamed Tmar, and Hawete Hattab. A new relevance feedback algorithm based on vector space basis change. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 8404 of *Lecture Notes in Computer Science*, pages 355–366. Springer Berlin Heidelberg, 2014.

10 Massimo Melucci. Context modeling and discovery using vector space bases. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM'05, pages 808–815, New York, NY, USA, 2005. ACM.

11 Massimo Melucci. A basis for information retrieval in context. *ACM Trans. Inf. Syst.*, 26(3):14:1–14:41, June 2008.

12 Jun Miao, Jimmy Xiangji Huang, and Zheng Ye. Proximity-based rocchio's model for pseudo relevance. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'12, pages 535–544, New York, NY, USA, 2012. ACM.

13 Karthik Raman, Raghavendra Udupa, Pushpak Bhattacharya, and Abhijit Bhole. On improving pseudo-relevance feedback using pseudo-irrelevant documents. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval*, ECIR'2010, pages 573–576, Berlin, Heidelberg, 2010. Springer-Verlag.

14 Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. Okapi at trec-4. In *TREC*, 1995.

15 Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at trec. In *TREC*, pages 21–30, 1992.

16 G. Salton. *The SMART retrieval system: experiments in automatic document processing.* Prentice-Hall series in automatic computation. Prentice-Hall, 1971.

17 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., San Francisco, CA, USA, 1997.

18 Cornelis Joost van Rijsbergen. *The geometry of information retrieval.* Cambridge University Press, 2004.

19 Ryen W. White and Gary Marchionini. Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.*, 43(3):685–704, May 2007.

20 Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, January 2000.

21 Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM'01, pages 403–410, New York, NY, USA, 2001. ACM.