José Luis Vicedo
Patricio Martínez-Barco
Rafael Muñoz
Maximiliano Saiz Noeda (Eds.)

LNAI 3230

Advances in Natural Language Processing

4th International Conference, EsTAL 2004
Alicante, Spain, October 2004
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

José Luis Vicedo
Patricio Martínez-Barco
Rafael Muñoz
Maximiliano Saiz Noeda
Universidad de Alicante
Departamento de Lenguajes y Sistemas Informáticos
Carretera de San Vicente del Raspeig
03690 San Vicente del Raspeig, Alicante, Spain
E-mail: {vicedo;patricio;rafael;max}@dlsi.ua.es

Library of Congress Control Number: 2004113295

CR Subject Classification (1998): I.2.7, F.4.2-3, I.2, H.3, I.7

ISSN 0302-9743
ISBN 3-540-23498-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11334347 06/3142 5 4 3 2 1 0

The Merging Problem in Distributed Information Retrieval and the 2-Step RSV Merging Algorithm

Fernando Martínez-Santiago, Miguel Angel García Cumbreras, and L. Alfonso Ureña López

Department of Computer Science, University of Jaén, Jaén, Spain

Abstract. This paper[1] presents the application of the 2-step RSV (retrieval status value) merging algorithm to DIR environments. The reported experiment shows that the proposed algorithm is scalable and stable. In addition, it obtains a precision measure higher than the well known CORI algorithm.

1 Introduction

Usually, a Distributed Information Retrieval (DIR) system must rank document collections by query relevance, selecting the best set of collections from a ranked list, and merging the document rankings that are returned from a set of collections. This last issue is the so-called "collection fusion problem" [8, 9], and it is the topic of this work. We propose an algorithm called 2-Step RSV [3, 4]. This algorithm works well in Cross Language Information Retrieval (CLIR) systems based on query translation, but the application of 2-Step RSV in DIR environments requires an additional effort: learning of collection issues such as document frequency, collection size and so on. On the other hand, since 2-Step RSV makes up a new global index based on query terms and the whole of retrieved documents, it makes possible the application of blind feedback at a global level by means of the DIR monitor, rather than a local level by means of each individual Information Retrieval (IR) engine. Previous works have researched into the application of Pseudo-Relevance Feedback (PRF) to improve the selection process of the best set of collections from a ranked list [5]. This work emphasizes the effectiveness of PRF applied to the collection fusion problem. Finally, a second objective is to study the stability of 2-Step RSV against weighting function variances in the local indices.

The rest of the paper is organized as follows. Firstly, we present a brief revision of DIR problems. Section 2 describes our proposed method which is integrated into CORI model[1]. In section 3, we detail the experiments carried out and the results obtained. Finally, we present our conclusions and future lines of work.

---

[1] This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

J. L. Vicedo et al. (Eds.): EsTAL 2004, LNAI 3230, pp. 442–453, 2004.

2 The 2-Step RSV Algo

The basic 2-step RSV idea is straig
several selected collections, their d
this way, the method recalculates t
frequency of each query term. Give
will be calculated by means of the
term for each selected collection. G

1. The document pre-selection ph
locally for each selected collecti
collection of preselected docum
the top retrieved documents fo
2. The re-indexing phase consists
considering solely the query voc
given a term, its document freq
document frequency of each te
new index is created by using t
is carried out against the new i
selected, I_1 and I_2, and the te
the new global frequency is df_{I_1}

In this work, we have used OKA
the second step of the 2-step RSV
document length, the term frequenc
elements in order to calculate OKA
frequency- are learned by means of
[7] algorithms. Term frequency requ
2- step RSV creates the index at
is created step by step. For exampl
documents per selected collection,
to the user. If more documents are
downloads the next two or three d
After some documents (no more t
reindexed, the application of blind
documents at global level, the Ro
applied. In this work, PRF is appl
the top ten terms obtained from th
query is applied to reweigh every dow
query is only applied at a global an

3 Experiments and Resu

3.1 Experimental Methods

The steps to run each experiment ar

The 2-Step RSV Algorithm

e basic 2-step RSV idea is straightforward: given a query term distributed over eral selected collections, their document frequencies are grouped together. In s way, the method recalculates the document score by changing the document quency of each query term. Given a query term, the new document frequency l be calculated by means of the sum of the local document frequency of the m for each selected collection. Given a user query, the two steps are:

The document pre-selection phase consists of searching relevant documents locally for each selected collection. The result of this previous task is a single collection of preselected documents (I' collection) as result of the union of the top retrieved documents for each collection.

The re-indexing phase consists of re-indexing the global collection I', but considering solely the query vocabulary. Only the query terms are re-indexed: given a term, its document frequency is the result of grouping together the document frequency of each term from each selected collection. Finally, a new index is created by using the global document frequency, and the query is carried out against the new index. Thus for example, if two collections are selected, I_1 and I_2, and the term "government" is part of the query, then the new global frequency is $df_{I_1}(government) + df_{I_2}(government)$.

In this work, we have used OKAPI BM-25 [6] to create the global index for e second step of the 2-step RSV approach. The collection size, the average cument length, the term frequency and the document frequency are required ments in order to calculate OKAPI BM-25. These elements -except the term quency- are learned by means of *Capture-recapture* [2] and *sample-resample* algorithms. Term frequency requires the document to be downloaded before step RSV creates the index at query time. Note that the merging process created step by step. For example, the DIR monitor downloads two or three cuments per selected collection, it applies 2-step RSV and shows the result the user. If more documents are required by the user, then the DIR monitor wnloads the next two or three documents per selected collection and so on. ter some documents (no more than 10 or 20) have been downloaded and indexed, the application of blind feedback is easy. Given the R top-ranked cuments at global level, the Robertson-Spark Jones PRF approach [6]] is plied. In this work, PRF is applied by expanding the original query using e top ten terms obtained from the top ten documents. Then, the expanded-ery is applied to reweigh every downloaded document. Note that the expanded ery is only applied at a global and not a local level.

Experiments and Results

1 Experimental Methods

ne steps to run each experiment are the followings:

1. Create one index for each collection (OKAPI or a random index).
2. Use CORI to score each collection for that query.
3. When all collections have been scored, choose the most relevant using a *clustering* algorithm or by selecting the N best scored.
4. Run the query using the selected indices.
5. In some experiments, apply pseudo-relevance feedback in each collection, running each expanded query against its index.
6. Merge the document rankings using CORI or 2-Step RSV .
7. In some experiments, apply again pseudo-relevance feedback using the *on-line* index created by 2-Step RSV .
8. Evaluate the quality of the results in terms of precision and recall. The precision measure used is the average precision at 5,10,20 and 100 documents, and the average precision at eleven coverage points.

The parameters of each experiment are the set of queries used, whether or no ? have used feedback and the fusion strategy. Another variable is the weighting ?nction used with each index (always OKAPI in the multilingual scene). This ?eans that there are some experiments where we have only used OKAPI, but ? others we have used mixed indices since this scene shows in a real manner a ?stributed system: two collections that belong to the same distributed system ?nnot use the same weighting function.

? Test Collection Description

?e experiments are carried out using three partitions of the conferences TREC1 ?d TREC2 and two sets of queries for the queries 51-100 and 101-150 of ? TREC collections. The TREC1 and TREC2 collections belong to the text ?blished between 1987 and 1990 in various newspapers, news agencies and ?torials. They contain more than two gigabytes of data, divided into 740.000 ?uments. Over these thirteen collections we have made three tests, described ?he 1:

TREC-1. The thirteen collections indexed with only one index. It shows the best case.
TREC-13. Each collection of the thirteen has been indexed separately.
TREC-80. The original thirteen collections have been divided into eighty collections, and indexed separately. The procedure to create these eighty collections is the following:

1. Each source (AP,DOE,FR,WSJ,ZIFF) has a number of subcollections, according to its size. Therefore, AP has 19 subcollections, DOE has 7, FR has 18, WSJ has 20 and ZIFF has 16.
2. With a random value and according to its source each document has been copied in one of its subcollections.

?he queries used appear in the first two editions of the TREC conferences, 100 ?ies, in total. We have used, in the experiments, only the title and description ?.

Table 1. Description

	# of docs	
	Min.	Avg.
TREC-1	741.991	741.991
TREC-13	10.163	57.066
TREC-80	2473	9273

3.3 Experiments Description

Using the sets of indices, TREC-1, TRE experiments, according to the following

- Weighting functions used in the indic of the proposed model, and for this p sets, *OKAPI* and *random*. With the with *OKAPI* weighting method. In th weighting function from those of the Z has been indexed using only OKAPI.
- Collection selection method. In a distr each collection for a given query, and us to value each collection, but does follow to choose more or fewer collecti or setting a fixed value of recovery c have used clustering and also a fixed
- Application of local pseudo-relevance automatically the query, with the in Nonetheless, since the added terms a receptionist or the central system, so of the possible local increase. Even so, popular method can affect in the fina made by adding the most relevant ter the original query.
- Application of global pseudo-relevance 2- step RSV generates a new index in a new index to apply PRF, although it documents, to analyze them and to ext power of these N documents. In this w Therefore, this method introduces a lo

Experiments without Query Expansic this section CORI and 2-step RSV are c indexed using the same weighting method local nor global query expansion.

Tables 2 and 3 show the results obta step RSV, whose increase is better, in avera

Table 1. Description of the collections sets

	# of docs.			Size (in MB)		
	Min.	Avg.	Max.	Min.	Avg.	Max.
TREC-1	741.991	741.991	741.991	2168	2168	2168
TREC-13	10.163	57.066	226.087	33	159.23	260
TREC-80	2473	9273	32.401	23	25,81	30

Experiments Description

the sets of indices, TREC-1, TREC-13 and TREC-80, we have run some iments, according to the following parameters:

/eighting functions used in the indices. We decided to measure the stability : the proposed model, and for this purpose we have worked with two index :ts, *OKAPI* and *random*. With the first, all collections have been indexed ith *OKAPI* weighting method. In the second case we have chosen a random eighting function from those of the ZPRISE system. The TREC-1 collection as been indexed using only OKAPI.

'ollection selection method. In a distributed system it is necessary to weight ach collection for a given query, and choose the most relevant. CORI allows s to value each collection, but does not say which judgement we should ›llow to choose more or fewer collections. [1] suggests using cluster methods r setting a fixed value of recovery collections. We have applied both; we ave used clustering and also a fixed value N (with N=5,10,15 and 20).

ıpplication of local pseudo-relevance feedback. Each librarian can expand utomatically the query, with the increase of the local results in mind. Ionetheless, since the added terms are local, these are not known by the eceptionist or the central system, so CORI and also 2-step RSV lose some f the possible local increase. Even so, it is very important to study how this opular method can affect in the final result [5]. The expansion has been nade by adding the most relevant ten terms of the ten first documents to he original query.

\pplication of global pseudo-relevance feedback. Since the estimation of the ?- step RSV generates a new index in a query time, it is possible to use this ıew index to apply PRF, although it is necessary to download the first N locuments, to analyze them and to extract the terms with the most relevant ɔower of these N documents. In this work we use the first ten documents. Therefore, this method introduces a low cost.

ıeriments without Query Expansion and Homogeneous Indices. In section CORI and 2-step RSV are compared. All collections have been xed using the same weighting method, OKAPI. We have applied neither l nor global query expansion.

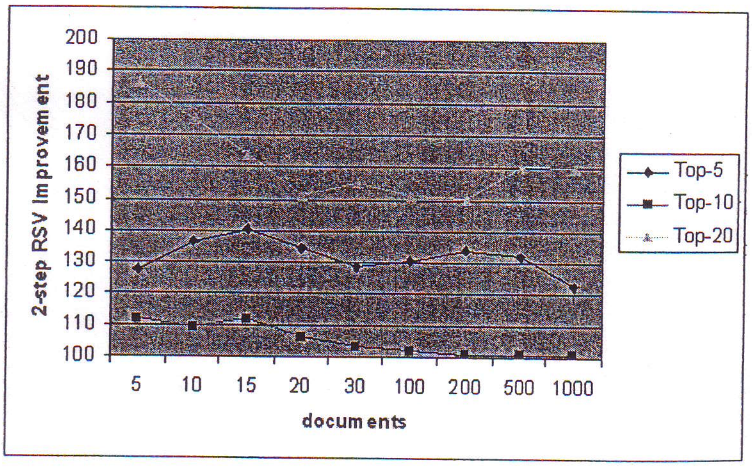
'ables 2 and 3 show the results obtained. The best results are from 2- RSV, whose increase is better, in average precision terms, between 19,4%, able 2. DIR experiments without feedback and homogeneous indices (set TREC-13, ueries 101-150)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
Top-10					
CORI	0,484	0,416	0,360	0,268	0,134
2-Step RSV	0,492	0,478	0,445	0,348	0,199
Centralized	0,492	0,492	0,444	0,346	0,194

able 3. DIR experiments without feedback and homogeneous indices (set TREC-80, ueries 51-100)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
Top-20					
CORI	0,261	0,263	0,293	0,210	0,071
2-Step RSV	0,488	0,460	0443	0,318	0,147
Centralized	*0,556*	*0,514*	*0,492*	*0,371*	*0,210*

we use the five first documents over TREC-13, and 107% if we use the first venty documents over TREC-80. In this last case, in absolute terms, 2-step RSV nproves CORI in 7,6 points, but CORI starts from a low average precision),07), so the increase looks better than it actually is. CORI obtains its best esults when the ten first collections are selected. However its performance uddenly drops if the twenty first collections are selected. In this point 2-step SV is more stable, because the more selected collections, the better the precision. his aspect is shown in figure 1, which shows the performance of 2-step RSV relation to CORI.



ig. 1. Improvement of *2-step RSV* over CORI (coll. TREC-80,queries 51-100). The ase case (100%) shows CORI precision

Finally, in the comparison between model, an interesting result appears. over 13 subcollections, using $Top-10$, the centralized model, and therefore, it all collections, there is not always a loss level obtained with TREC-13 is not th precision obtained is lower, although th of the collections used is lower in com if the twenty first collections are used, average precision obtained is higher th centralized model.

Experiments without Query Exp In this section we have studied how t can affect each subcollection. Whil collections have been indexed with the one each subcollection has been indexe the IR ZPrise system. Tables 4 and 5 s same tendency as the previous section:

- Using the test set TREC-13, 2-step R of CORI, over the run selection pro
- Using TREC-80, this difference is twenty first collections. In this la performance obtained is not even th with 2-Step RSV . In the other cases 40%.

Table 4. DIR Experiments without feedbac queries 101-150)

Fusion	5-prec	10-prec
Top-		
CORI	0,348	0,254
2-Step RSV	0,492	0,480
Centralized	*0,492*	*0,492*

Table 5. Experiments without feedback a queries 51-100)

Fusion	5-prec	10-prec
Top-		
CORI	0,176	0,214
2-Step RSV	0,610	0,619
Centralized	*0,556*	*0,514*

Finally, in the comparison between CORI and 2-step RSV, and a centralized nodel, an interesting result appears. The precision obtained with 2-step RSV ver 13 subcollections, using $Top-10$, is better than the precision obtained with he centralized model, and therefore, it is proved that if the system does not use ll collections, there is not always a loss of precision. The good performance and evel obtained with TREC-13 is not the same with TREC-80. In this case the recision obtained is lower, although the most probable reason is that the ratio f the collections used is lower in comparison with the total available. In fact, the twenty first collections are used, which represents 25% of the total, the verage precision obtained is higher than 70% of the precision obtained with a entralized model.

xperiments without Query Expansion and Heterogeneous Indices. 1 this section we have studied how the use of a random weighting function an affect each subcollection. While in the last section all the sub-ollections have been indexed with the same weighting function OKAPI, in this ne each subcollection has been indexed with a random function, available in ne IR ZPrise system. Tables 4 and 5 show that the results obtained follow the me tendency as the previous section:

- Using the test set TREC-13, 2-step RSV almost doubles the average precision of CORI, over the run selection procedures (Top-5, Top-10 and *clustering*).
- Using TREC-80, this difference is not so important, except if we use the twenty first collections. In this last case, CORI drops again, and the performance obtained is not even the third part of the performance obtained with 2-Step RSV . In the other cases the increase of the performance is about 40%.

ble 4. DIR Experiments without feedback and heterogeneous indices (TREC-13 set, eries 101-150)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
		Top-10			
CORI	0,348	0,254	0,228	0,208	0,100
2-Step RSV	0,492	0,480	0,443	0,350	0,198
Centralized	*0,492*	*0,492*	*0,444*	*0,346*	*0,194*

ble 5. Experiments without feedback and heterogeneous indices (TREC-80 set, eries 51-100)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
		Top-20			
CORI	0,176	0,214	0,231	0,167	0,046
2-Step RSV	0,610	0,619	0,569	0,377	0,160
Centralized	*0,556*	*0,514*	*0,492*	*0,371*	*0,210*

2-step RSV and also CORI lose precision when a random method is introduced in the selection of the weighting function. This is almost inevitable because the local results obtained with each collection are in some experiments worse than with OKAPI. Nonetheless, this loss of precision is not the same with CORI as with 2-step RSV. If we focus on the results obtained over TREC-80, while 2- step RSV loses in general between 10% and 20%, CORI, with the inclusion of various weighting functions, loses more than 40%. This situation is shown in the figure 2, which shows the quotient between the precision obtained with OKAPI and with the random functions. Using *clustering* and also Top-10, while with 2-step RSV the performance is around 80% of the precision obtained with homogeneous indices, CORI is only around 60%. Therefore, it is possible to arm that 2-step RSV is a stable algorithm against the variation of the weighting functions used in each subcollection.

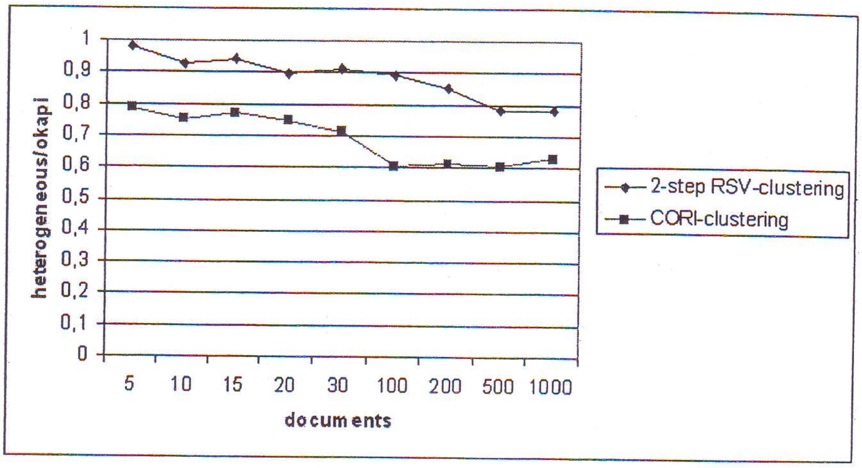


Fig. 2. Impact of the use of heterogeneous indices in the performance of CORI and 2-Step RSV (coll. TREC-80,queries 51-100,clustering). The base case (100%) shows the precision obtained with the algorithm using homogeneous indices (only OKAPI)

3.4 Experiments with Query Expansion

In this section we study the impact of the use of query expansion method based in local and global PRF:

- Local pseudo-relevance feedback. Each librarian applies PRF locally with the purpose of increasing the results obtained over this library.
- Global pseudo-relevance feedback. This case can only be applied to 2-step RSV. Since the receptionist generates a new global index, it is possible to apply PRF over this index.
- Local and global pseudo-relevance feedback. Finally, it is possible to apply PRF first for each librarian, and also for the receptionist later.

Local PRF Experiments. As
not provide an increase in the 2-s
a little better over 13 collections.
the results even worse, but this is
very small (not more than two p
to assume that the PRF does no
with 2-step RSV.

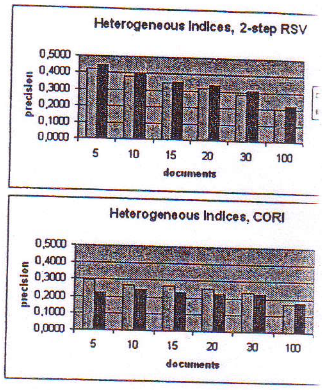


Fig. 3. Local feedba

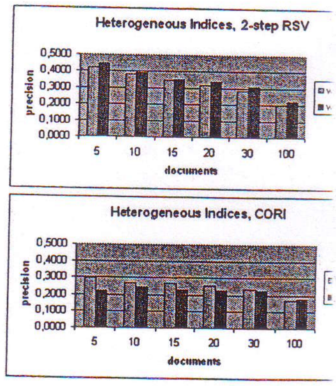


Fig. 4. Local feedback

In systems that do not collaborat
to the added terms of each libraria
[5] shows that the use of query expa
Analysis) does not increase either
selection. A possible cause is the exp

Local PRF Experiments. As figures 3 and 4 show, the local feedback does not provide an increase in the 2-step RSV case. If we use CORI the situation is a little better over 13 collections. The use of the PRF over 80 collections makes the results even worse, but this is not always so. In every case the differences are very small (not more than two points) and the conclusion is that it is possible to assume that the PRF does not affect the final result, with both CORI and with 2-step RSV.

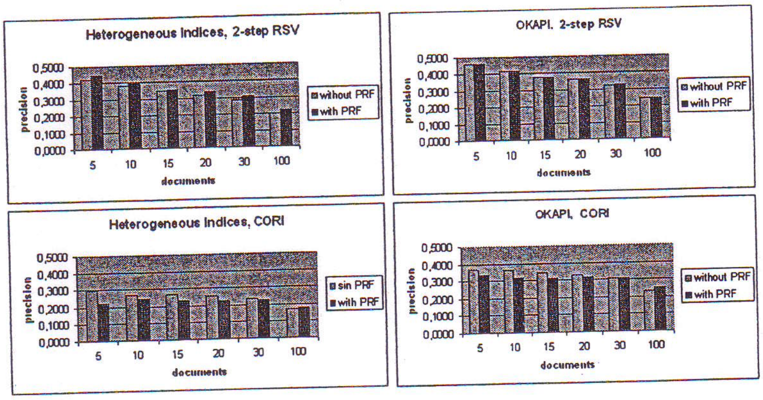


Fig. 3. Local feedback impact (TREC-13, Top-10)

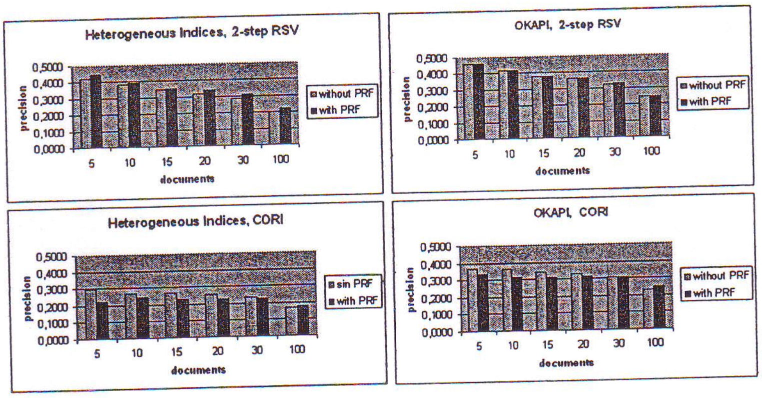


Fig. 4. Local feedback impact (TREC-80, clustering)

In systems that do not collaborate completely (the receptionist cannot access to the added terms of each librarian) this result was already known to CORI. [5] shows that the use of query expansion methods (in their case, *Local Context Analysis*) does not increase either the collections selection or the documents selection. A possible cause is the expanded query length, because this new length makes the score normalization difficult. This reason cannot be applied to 2-step RSV . Some experiments in the last section show that the performance of the 2-step RSV is not different if the length of the query is different. In the case of 2-step RSV there may be two causes:

- The receptionist only works with the original query vocabulary, so the documents that are now relevant because of the query expansion, are not selected.
- 2-step RSV does not use the local score obtained for each document. The sole relevant condition of a document to 2-step RSV is that this document belongs to the list given by the librarian and the query vocabulary that this document contains, and never the local score obtained.

Experiments with Global PRF. Whether or not we use not local feedback, it is possible to apply query expansion methods, not using each collection as a single unit but using the index created in the second phase of the 2-step RSV method. The receptionist uses the expanded query, instead of the original one, in order to evaluate every document received from each local IR system. The computational cost in this case is not very high, if compared with the computational cost of a centralized system, because it only needs to analyze a small number of documents, in our experiments the ten first documents, so in general it is only necessary to download two or three documents per selected collection, using any procedure like the ones described in section 4.

Table 6. DIR Experiments with global feedback and homogeneous indices (TREC-13 set, queries 101-150)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
Top-10					
CORI	0,484	0,416	0,360	0,268	0,134
2-Step RSV	0,492	0,478	0,445	0,348	0,199
2-Step RSV +global PRF	0,432	0,424	0,432	0,375	0,232
Centralized	*0,492*	*0,492*	*0,444*	*0,346*	*0,194*
Centralized+PRF	*0,540*	*0,526*	*0,497*	*0,418*	*0,273*

Table 7. DIR Experiments with global feedback and homogeneous indices (TREC-80 set, queries 101-150)

Fusion	5-prec	10-prec	20-prec	100-prec	Avg-prec
Top-20					
CORI	0,274	0,259	0,272	0,192	0,067
2-Step RSV	0,488	0,446	0,408	0,289	0,131
2-Step RSV +global PRF	0,480	0,440	0,431	0,326	0,155
Centralized	*0,492*	*0,492*	*0,444*	*0,346*	*0,194*
Centralized+PRF	*0,540*	*0,526*	*0,497*	*0,418*	*0,273*

In tables 6 and 7 the results
geneous indices are shown.

The increase of the average pr
to the original 2-step RSV is quite
TREC-13 case, is 16.6% (Top-10
clustering. In this way the results
17.4% (Top-5) and 20% (Top-20
PRF is 41%, much more than the
clear: the centralized model has a
possible to include relevant docu
This is impossible with 2-step R
documents and *some* collections,
but never adds new documents.
run the expanded query over eac
new results the original 2-step R

It is clear that the global PRF
always increase the selection of t
worse precisions with global PRF
only the five or ten first docume
figure 5.

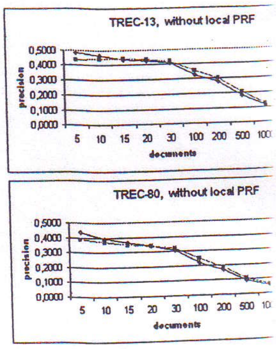


Fig. 5. Global feedbac

The use of global PRF incre
relevant documents between t
precision of the first document
is usual in these cases, the ans
application of PRF worsens a l
and from this number the result
cost of applying PRF is moder
first documents in the query ti
be the reason to apply or not
the user has to wait some mor

In tables 6 and 7 the results obtained for the query set 51-100 and homogeneous indices are shown.

The increase of the average precision introduced using global PRF in relation to the original 2-step RSV is quite signicant, around 15-20%. The lowest increase, TREC-13 case, is 16.6% (Top-10) and the highest is 21%, with TREC- 13 and *clustering*. In this way the results of the experiments with TREC-80 are between 17.4% (Top-5) and 20% (Top-20). The increase of the centralized model with PRF is 41%, much more than the 20% obtained with 2-step RSV . The reason is clear: the centralized model has access to *all* documents of *all* collections, so it is possible to include relevant documents which have not been selected previously. This is impossible with 2-step RSV , because this method only includes *some* documents and *some* collections, and only resorts documents selected previously, but never adds new documents. In any case, this situation could change if we run the expanded query over each selected collection, and then apply over these new results the original 2-step RSV algorithm.

It is clear that the global PRF increases the average precision, but it does not always increase the selection of the first documents, and it is frequent to obtain worse precisions with global PRF compared with the results without PRF, when only the five or ten first documents are considered. This conclusion is shown in figure 5.

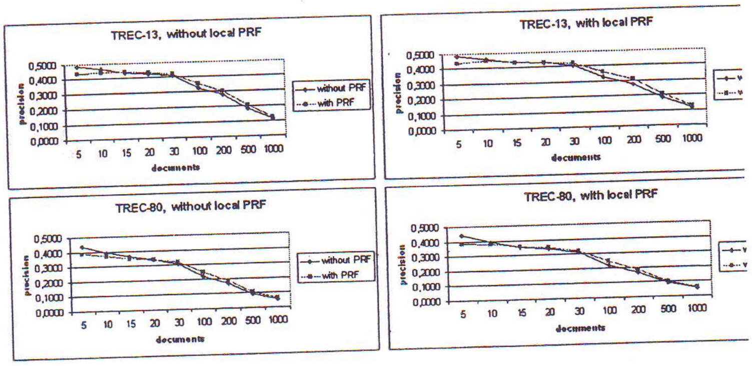


Fig. 5. Global feedback impact (random indices, clustering) (I)

The use of global PRF increases recall in general, because it introduces more relevant documents between the first thousands, but it does not increase the precision of the first documents. Is it always advisable to use global PRF?. As is usual in these cases, the answer depends on the user's needs. In general the application of PRF worsens a little the ranking of the first 10 or 15 documents, and from this number the result increases. On the other hand, the computational cost of applying PRF is moderate but not null, because it needs to analyze the first documents in the query time. It is possible that this computational cost will be the reason to apply or not this method: a little better results in general but the user has to wait some more seconds.

4 Conclusions and Future Work

This paper shows the application of 2-step RSV to DIR environments. We have focused on two questions:

1. Collections with different weighting functions and the improvement of blind feedback applied by the DIR monitor at global level. The experiments about the first question show that 2-step RSV is robust against weighting functions variances. 2-step RSV and also CORI lose precision when a random method is introduced into the selection of the weighting function. The loss of precision with CORI is more than 40%, while with 2-step RSV it is between 10% and 20%.
2. Blind feedback is a useful technique whenever it is applied at global level rather than individually for each IR system. Global PRF applied with 2-step RSV increases the average precision, but it does not always increase the selection of the first documents, and it is frequent to obtain worse precisions with global PRF compared with the results without PRF, when only the five or ten first documents are considered.

We can also test the improvement of the results using global PRF by sending expanded queries for each collection instead of recalculating the score of documents received by means of the original query. The steps of the proposal architecture would be modified as follows:

1. Receptionist receives the user query. Such query is sent to selected collections.
2. Selected collections send a few documents to the receptionist. These documents are used by the receptionist in order to expand the initial user query.
3. The expanded query is sent to the selected collections. The documents received initially are discarded.
4. The score of documents received from the collections is recalculated by using the global index generated by 2-step RSV method and the expanded query.
5. Finally, the receptionist ranks the received documents.

We hope that this approach will achieve an improvement similar to the one obtained using PRF in a centralized IR system. This additional step will have a computational cost that will be also studied.

References

1. J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th International Conference of the ACM SIGIR'95*, pages 21–28, New York, 1995. The ACM Press.
2. K. Liu, C. Yu, W. Meng, A. Santos, and C. Zhang. Discovering the representative of a search engine. In *Proceedings of 10th ACM International Conference on Information and Knowledge Managemenet (CIKM)*, 2001.
3. F. Martínez-Santiago, M. Martín, and L. Ureña. SINAI at CLEF 2002: Experiments with merging strategies. *Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer Verlag*, pages 187–197, 2003.

4. F. Martínez-Santiago, A. Mo
2003: Merging and decompou
M. Kluck, editors, *Proceeding*
System Evaluation Campaign,
5. P. Ogilve and J. Callan. The ef
tion retrieval. In *Proceedings o*
Knowledge Management (CIK
6. S. E. Robertson, S. Walker.,
Okapi at TREC. *Information*
7. L. Si and J. Callan. Relevan
source selection. In T. A. Pres
Conference on Research and D
8. E. Voorhees, N. K. Gupta, and
D. K. Harman, editor, *Procee*
volume 500-225, pages 95–104,
and Technology, Special Public
9. E. Voorhees, N. K. Gupta, and
gies. In ACM, editor, *Procee*
SIGIR Conference on Researc
172–179, Seattle, 1995.

4. F. Martínez-Santiago, A. Montejo-Ráez, L. Ureña, and M. Diaz. SINAI at CLEF 2003: Merging and decompounding. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, *Proceedings of the CLEF 2003 Cross-Language Text Retrieval System Evaluation Campaign*, pages 99–109, 2003.
5. P. Ogilve and J. Callan. The effectiveness of query expansion for distributed information retrieval. In *Proceedings of the Tenth International Conference on Information Knowledge Management (CIKM 2001)*, pages 193–190, 2001.
6. S. E. Robertson, S. Walker., and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*, 1(36):95–108, 2000.
7. L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In T. A. Press., editor, *Proc. of the 26 Annual Int'l ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, 2003.
8. E. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. In D. K. Harman, editor, *Proceedings of the 3th Text Retrieval Conference TREC-3*, volume 500-225, pages 95–104, Gaithersburg, 1995. National Institute of Standards and Technology, Special Publication.
9. E. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In ACM, editor, *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 172–179, Seattle, 1995.