

A Cluster Based Pseudo Feedback Technique Which Exploits Good and Bad Clusters

Javier Parapar and Álvaro Barreiro

IRLab, Computer Science Department
University of A Coruña, Spain
{javierparapar,barreiro}@udc.es

Abstract. In the last years, cluster based retrieval has been demonstrated as an effective tool for both interactive retrieval and pseudo relevance feedback techniques. In this paper we propose a new cluster based retrieval function which uses the best and worst clusters of a document in the cluster ranking, to improve the retrieval effectiveness. The evaluation shows improvements in some standard TREC collections over the state-of-the-art techniques in precision and robustness.

1 Introduction and Motivation

Several strategies were studied in the history of the Information Retrieval in order to improve the retrieval models effectiveness. One technique that has been demonstrated successful is relevance feedback. In this family of techniques is particularly interesting the so called pseudo relevance feedback [3], where relevance of the feedback documents is assumed.

Clustering has been considered as an useful tool in the retrieval process since the formulation of the cluster hypothesis in 1979 [16]. This hypothesis states that very related documents tend to be relevant to the same query. Indeed, several experiments [5,18] have demonstrated that clustering algorithms working at pseudofeedback time can obtain clusters with a high percentage of relevant documents, still the automatic identification of these clusters between the whole set of them is still a challenge.

Although initial experiments using query specific clustering [12] in order to improve the retrieval effectiveness were not conclusive, after improving the cluster representation [13] and with the use of clustering algorithms that support overlapping [9], finally the quality of the initial ranking was significant improved with cluster based re-ranking [13,7].

It was only recently when a cluster based retrieval approach was used to improve the quality of the pseudo relevance set, for further use in query expansion methods [11]. This approach takes advantage of the better initial ranking produced by the cluster based retrieval to select a better pseudo relevance set, improving in this way the effectiveness, sometimes significantly. But, although this kind of methods tend to improve the effectiveness in average, one known problem of them is the lack of robustness, i.e., still a significant amount of queries

are negatively affected by them. One of the main factors of this behaviour is the presence of non-relevant documents in the feedback set.

In this paper we present a new cluster based retrieval method that exploits bad clusters in order to reduce the amount of non-relevant documents in the feedback set. We consider not just if a document is present inside a “good” cluster to update its score, but also the presence of the document in “bad” (low relevance score) clusters. As far as we know this kind of negative information has not been exploited yet in the context of pseudo relevance feedback.

We tested our approach in several TREC Collections and compared with a Language Modelling retrieval approach [20], a query expansion based model [1] and with the resampling method presented in [11]. The evaluation shows that the results in terms of MAP are so good or better than the resampling approach but our method consistently improves the robustness in all the collections.

The paper is presented as follows. Section 2 presents our proposal explaining the different steps of the model. Section 3 explains the evaluation methodology and comments the results. In Section 4 we describe the related work and finally conclusions and future work are reported in Section 5.

2 Cluster Based Pseudo Relevance Feedback

In order to get a better pseudo relevance set we formulated a new cluster based re-ranking function. The first step of our method is to perform an initial document ranking, in this case we chose as base a Language Model (LM). After that, we cluster the top N documents (d_{init}), we chose in this case a clustering algorithm with overlapping. Once the top documents are clustered the query likelihood is calculated for the resulting clusters. After that, the clusters query likelihoods are combined by the retrieval formula re-ranking the documents. And finally these new top documents are used to feed a query expansion process.

Initial Ranking. In Language Models, the probability of a document given a query, $P(d|q)$, is estimated using Bayes’ rule as presented in Eq. 1.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \quad (1)$$

In practice $P(q)$ can be dropped for document ranking purposes. The prior $P(d)$ encodes a-priori information on documents and the query likelihood, $P(q|d)$, incorporates some form of smoothing. In this paper we consider uniform priors and unigram Language Models with Dirichlet smoothing [20], see Eq. 2.

$$P(q|d) = \prod_{i=1}^n \frac{tf(q_i, d) + \mu \cdot P(q_i|Col)}{|d| + \mu} \quad (2)$$

where n is the number of query terms, $tf(q_i, d)$ is the raw term frequency of q_i in d , $|d|$ is the document length expressed in number of terms, and μ is a parameter for adjusting the amount of smoothing applied. $P(q_i|Col)$ is the probability of the term q_i occurring in the collection Col that we obtained with the maximum likelihood estimator computed using the collection of documents.

Clustering Algorithm. Once that the initial ranking is obtained, clustering is performed over the top N documents. The use of clustering algorithms with overlapping has already been demonstrated successful [9] in cluster based retrieval. Indeed, initial approaches to query specific cluster [12] were not conclusive and it was only after incorporating clustering algorithms with overlapping [13] when the results were improved. As we explained one of the main points of our method is to use the information provided by bad clusters to avoid non-relevant documents in the pseudo relevance set. In order to do this we used a clustering algorithm that supports overlapping, i.e. one document can belong to one or more clusters.

The straightforward selection based in previous works could be using a k -nearest neighbours (k -NN) algorithm, but the k -NN forces to each document to have k neighbours. This aspect is not desired in our approach because we will exploit that a document belongs to a low scored cluster. If we had used k -NN, a non-relevant document with low query likelihood and no close neighbours could attract other documents that, although they are not close to that document, they are the closest ones.

So we decided to cluster the documents in base to a given threshold t , grouping for each document those neighbours that are more similar than t . Let's call this way of grouping thr -N. The purpose of this algorithm is that non-relevant documents could be isolated in singletons [14]. Standard $tf \cdot idf$ document representation was used with the well-known cosine similarity function in order to calculate distances between the documents.

Cluster Query Likelihood. In order to exploit the cluster information in our retrieval approach we need a way of estimating cluster the query likelihood. In the origin the first approaches to cluster retrieval considered the clusters as meta-documents, i.e. one cluster is represented as the concatenation of the documents that belong to it [9,12], or the centroid of the cluster [19]. But these representations suffer from several problems because of the document and cluster sizes. As demonstrated by Liu and Croft in [13], the geometric mean is a better cluster representation in order to calculate the cluster query likelihood, so it was chosen in our approach. The cluster query likelihood based on the geometric mean representation was calculated combining equations 3 and 4.

$$P(q|C) = \prod_{i=1}^n P(q_i|C) \quad (3)$$

$$P(w|C) = \prod_{i=1}^{|C|} P(w|d_i)^{\frac{1}{|C|}} \quad (4)$$

where n is the number of query terms, $|C|$ is the number of documents in the cluster, and $P(w|d_i)$ was computed using a Dirichlet estimate. Finally the cluster query likelihood applying logarithmic identities can be calculated as in Eq. 5

$$P(q|C) = \prod_{i=1}^n e^{\frac{\sum_{j=1}^{|C|} \log P(w_j|d_j)}{|C|}} \quad (5)$$

Cluster Based Reranking. Previous approaches to cluster based re-ranking only used the presence of a document in a good cluster as indicator of its relevance. As previous explained these approaches when using to construct pseudo relevance sets for further processing with, for instance query expansion, suffer from the problem that even the good clusters are not one hundred percent composed of relevant documents. The inclusion of non-relevant documents in the relevance set will produce a poor performance of the query expansion process resulting in effectiveness degradation for that query.

The final objective of our approach is to reduce the number of non-relevant documents in the pseudo relevance set. To achieve that point we decided to use the information given by the bad clusters. Our hypothesis is that given two documents d_1 and d_2 , and being C_{1max} , C_{1min} , C_{2max} and C_{2min} the clusters with best and worst query likelihood to which d_1 and d_2 belong respectively, if $P(q|C_{1max}) = P(q|C_{2max})$ and $P(q|d_1) = P(q|d_2)$ then if $P(q|C_{1min}) > P(q|C_{2min})$ should indicate that d_1 is likely to be more relevant than d_2 . In other words if a document belongs to low clusters in the cluster ranking, it should be a pseudo negative indicator about its relevance.

So in order to produce a document ranking we decided to combine the document query likelihood, with the pseudo positive information in terms of best cluster, and the negative in terms of the worst cluster to which the document belongs. The query likelihood combination is presented in Eq. 6.

$$P'(q|d) = P(q|d) \times \max_{d \in C_i} P(q|C_i) \times \min_{d \in C_i} P(q|C_i) \quad (6)$$

where $P(q|d)$ was estimated as in Eq. 2 and $P(q|C_i)$ was estimated as in Eq. 5.

Ideally removing all the non-relevant documents from the relevant set would have a great impact in order to get better expanded queries and, as a consequence, to improve the final retrieval effectiveness. Even although some relevant documents could be penalised because they group with other ones which appear low in the ranking, this effect will be extensively compensated by the benefit of removing the non-relevant documents from the relevance set.

Query Expansion for Pseudo Relevance Feedback. Once that we obtained a ranking with, hopefully, less amount of non-relevant documents in high positions, we take the first $|r|$ documents as the pseudo relevance set. With this relevance set we feed a query expansion approach. We chose to use Kullback Leiber Divergence (KLD) as the scoring function to select expansion terms [1,15], so the e terms with highest KLD score, calculated as in Eq 7, are selected to expand the original query.

$$kld_{score}(t) = \frac{tf(t,r)}{NT_r} \times \log \frac{tf(t,r) \times |Col|}{NT_r \times tf(t,Col)} \quad (7)$$

where $tf(t,r)$ is the term frequency of t in the pseudo relevance set, NT_r is the number of terms in the pseudo relevance set r , $|Col|$ is the total number of terms in the collection and $tf(t,Col)$ is the term frequency of t in the whole collection.

We interpolated the e terms selected as results of the KLD scoring formula for expansion terms with the original query. That was already demonstrated successful in RM3 [4], obtaining a high performance query expansion model. Therefore, the final rank is processed with the expanded query presented in 8:

$$\lambda \times (q_1, \dots, q_n), (1 - \lambda) \times (kld_{score}(t_1)t_1, \dots, kld_{score}(t_e)t_e) \quad (8)$$

where q_i are the original query terms, $kld_{score}(ti)ti$ are the expanded terms with the weight corresponding to their KLD score and λ is a parameter $0 \leq \lambda \leq 1$ than combines the original query and the expanded one.

3 Experiments and Results

The evaluation of our approach was performed over four TREC collections comparing with a baseline retrieval model, a baseline feedback model and a baseline cluster based feedback model. The results of an upper-bound model are also reported.

3.1 Settings and Methodology

Collections. We tested our method (BWCluster_PF) in four collections, two text collections: AP and WSJ, and two web collections: WT10g and GOV. We decided to use cross-validation evaluation: we performed training in a set of topics in the AP for both text collections and individual training for the each web collections. We chose to use short queries (title only). All the collections were preprocessed with standard stopword removal and Porter stemmer. In Table 1 are summarised the evaluation settings.

Table 1. Collections and topics for training and test

Col.	# of Docs	Topics	
		Train	Test
AP	242,918	51-150	151-200
WSJ	173,252		151-200
WT10g	1,692,096	451-500	501-550
GOV	1,247,753	WT04.1-50	WT04.151-200

Baselines

- **LM:** The baseline LM retrieval model with Dirichlet smoothing that was used as base of the other methods.
- **KLQE:** a pseudo relevance model based on the query expansion [1]. The selection of expansion terms and the construction of the expanded query which was explained in section 2. This was also the query expansion approach used for the next approaches.

- **Resampling:** The cluster based resampling method presented by Lee *et al.* in [11], but using the Geometric Mean instead of the document concatenation to compute the cluster query likelihood and KLQE instead of the estimation of the Relevance Model [10] to compute the expanded query.
- **TrueRF:** An upper-bound of all the pseudo relevance methods, was computed feeding the KLQE approach with all the relevant documents present in the d_{init} .

Training and Evaluation. As commented we performed cross-validation strategy, more precisely we perform training for the text collection with a set of topics with AP dataset, and testing in the AP and WSJ with different topics. For the web collections we trained in each collection with 50 queries each, and testing with other sets of queries. Training was performed optimising Mean Average Precision (MAP).

There are several parameters to train. Namely, the smoothing parameter μ was tuned in the baseline LM retrieval model ($\mu \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$) and its value was used for every other retrieval model. The parameters $|r|$, the size of the pseudo relevance set, e , the number of expansion terms, and λ , the interpolation factor, for the pseudo feedback based query expansion were trained in the KLQE model ($|r| \in \{5, 10, 25, 50, 75, 100\}$, $e \in \{5, 10, 25, 50, 75, 100\}$ and $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$) and also used in the cluster based approaches. The cluster based retrieval models, resampling and our method, need apart of μ , $|r|$, e and λ , the parameters k that was set to 5, t that was set to 0.05 and N , the size of the d_{init} , that was set to 100. In [11] Lee *et al.* demonstrated that with exhaustive tuning the resampling method obtains significant improvements over pseudo feedback based query expansion methods. In this paper we decided to avoid excessive tuning effort and we fix the values of μ , $|r|$, e and λ to the one trained in LM and KLQE respectively.

Therefore we have to remark that both cluster based approaches can be improved in terms of effectiveness by specifically tuning every parameter in each of the retrieval approaches. Also, our evaluation methodology allows see more clearly the effect of the cluster information without depending on excessive tuning effort. For the TrueRF upper-bound model we maintained the same parameter set as in the KLQE model only changing r by the set of relevant documents in the d_{init} .

Finally test values are reported for MAP and RI. The Robustness Index (RI) ($-1 \leq RI(q) \leq 1$) also called Reliability of Improvement Index of one model respect a baseline was presented by Sakail *et al.* in [17] and it is formulated as in Eq 9:

$$RI(q) = \frac{n_+ - n_-}{|q|} \quad (9)$$

where q is the set of queries over the RI has to be calculated, n_+ is the number of improved queries, n_- the number of degraded queries and $|q|$ the total number of queries in q .

3.2 Results

Analysing the MAP values for the test topics (see Table 2) it has to be notice that our approach significantly outperforms the baseline LM for every collection, a fact that neither the KLQE nor the Resampling method achieve. Our method only achieves statistically significant improvements over the query expansion method in the WT10g collection, this is explained in part because the KLQE values were obtained with the best parameter settings meanwhile the values of our method did not receive individual parameter tuning. The values of resampling method do not achieve statistical significant improvements over the query expansion method but again the same reason as previously explained applies; we have to remark that such achievements are reported in [11]. In WSJ, WT10g and GOV collection our method outperforms the resampling method being the improvements significant in two of them. Of course the upper-bound model outperforms any other method.

Table 2. Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon $p < 0.05$) with respect to the Dirichlet smoothed LM, KLQE, Resampling and Our method are superscripted with a, b, c and d respectively.

<i>Col.</i>	MAP				
	<i>LM</i>	<i>KLQE</i>	<i>Resampling</i>	<i>BWCluster_PF</i>	<i>TrueRF</i>
AP	.2078	.2918 ^a	.2889 ^a	.2884 ^a	.4569 ^{abcd}
WSJ	.3366	.3940 ^a	.3927 ^a	.3982 ^{ac}	.5614 ^{abcd}
WT10g	.2101	.2166	.2173	.2224 ^{abc}	.3002 ^{abcd}
GOV	.1949	.2198 ^a	.1994	.2128 ^a	.2200 ^{abcd}

Query robustness values measured with RI over the LM baseline model are reported in Table 3. The first fact to mention is that our method is better or equal than the KLQE approach except in the AP collection, where MAP values were also worse. In the case of the resampling approach the RI values are worse than the KLQE even in the case of the WT10g collection, where the resampling method MAP is better than KLQE. In this case, in the WT10g collection, the RI is negative, this means than more queries are penalised than benefited. Comparing both cluster based methods we have to remark that our method outperforms the resampling method in every collection but the AP where both methods report the same values. Again as expected the query robustness

Table 3. Values for Robustness Index (RI) with respect to the LM baseline model for every collection on the test topics

<i>Col.</i>	RI			
	<i>KLQE</i>	<i>Resampling</i>	<i>BWCluster_PF</i>	<i>TrueRF</i>
AP	0.44	0.40	0.40	0.96
WSJ	0.36	0.28	0.44	0.92
WT10g	0.16	-0.08	0.16	0.80
GOV	0.56	0.44	0.60	1.00

of the TrueRF upper-bound model is greater than any other method, although in three of the collections TrueRF still damage some queries.

4 Related Work

Since the formulation of the cluster hypothesis [16] several works tried to exploit clustering information to improve information retrieval tasks. It was only recently when conclusive results were presented improving retrieval effectiveness using query specific clustering. We have to cite the work of Kurland and Domshlak [8] where several features were aggregated to obtain better high precision in the re-ranking of top documents. Kurland and Domshlak used several features related with cluster information, namely *query faithfulness*, *self faithfulness*, *initial list faithfulness* and *peer faithfulness*. The individual feature results seem indicate that *peer faithfulness* is the better indicator, although the aggregation of all the features reports the best values. An approach based on similar facts is presented by Kurland in [7], in this paper the author present several approaches (aspect models and interpolation models) that also combine information about peer clusters. High precision is again improved, although the performance is quite dependent of the settings and MAP values are also reported but only in a cut-off of 50 documents. In [6] Kurland presented several cluster based re-ranking approaches, exploiting in this case clusters with high percentage of relevant documents. several features are combined resulting in improvements in high precision over the initial ranking.

Recently Lee *et al.* [11] proposed query specific clustering in order to improve the quality of the pseudo relevance set used in the query expansion process, in this case a relevance model (RM) [10]. This method uses as cluster re-ranking method the original cluster query likelihood presented by Liu and Croft in [12] but with overlapping clusters. The results show significant improvements over the initial LM based rank and the RM rank in several collections. This approach that we used in order to compare our method still shows some problems with query robustness that we tried to solve with our alternative approach using pseudo negative information. Although the results reported in [11] are higher than the reported here for their method we have again to remark that we did not perform individual parameter tuning for each retrieval method, and also we used KLD based query expansion instead of RM, so the results on this paper can be still improved for both, resampling and our method.

In line with the need of consider negative information, in this case associated with clustering processes, we have to remark the analysis already presented in 1995 by Lu *et al.* in [14]. In this paper it is commented that after running a cluster algorithm over the the top documents of a rank, most of the singletons (clusters with only one document) are non-relevant documents, and should be removed. This data suggested us that the clustering algorithm should allow the creation of singletons. Really not every singleton contains a non-relevant document but, allowing the creation of singletons, the real non-relevant documents will not be promoted in the ranking benefited because they are clustered with relevant ones, while non affecting negatively when they are relevant documents.

Also recently several works approached the task of getting a better pseudo relevance set, in this case to increase the diversity, but none of them show conclusive results. In [2] Collins-Tompson and Callan present sampling over the top documents based on query variants. The objective of having less redundant pseudo relevance set is also approached in [17]. Sakai *et al.* presented in this case a resampling method that it is actually based on clustering, the top documents are clustered based on the common query terms selecting only some of each cluster in order to improve diversity in the relevance set. But again the results presented in the evaluation are not conclusive.

5 Conclusions and Future Work

The proposed method introduces the use of bad clusters in order to achieve pseudo feedback sets with less non-relevant documents. The pseudo negative information is obtained from the belonging of the documents to a “bad” cluster in a cluster re-ranking approach. The results show improvements in MAP over the existing cluster based approaches for pseudo relevance feedback, that in some settings are statistical significant. Another good result is the improvement in terms of query robustness: our approaches penalise less queries than previous ones.

Further analysis of the non-relevant documents that still remain in the pseudo relevance set has to be done. We also want to study the effect of taking a bigger set of top documents d_{init} (evaluation was done with the top 100 documents) that should be an important factor when considering the pseudo negative information. Also we will perform individual parameter tuning for the cluster based approaches in order to report their best values.

Acknowledgements. This work was funded by *Ministerio de Ciencia e Innovación* under project TIN2008-06566-C04-04.

References

1. Carpineto, C., de Mori, R., Romano, G., Bigi, B.: An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.* 19(1), 1–27 (2001)
2. Collins-Thompson, K., Callan, J.: Estimation and use of uncertainty in pseudo-relevance feedback. In: *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 303–310. ACM Press, New York (2007)
3. Croft, W., Harper, D.: Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35, 285–295 (1979)
4. Diaz, F., Metzler, D.: Improving the estimation of relevance models using large external corpora. In: *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 154–161. ACM, New York (2006)
5. Hearst, M.A., Pedersen, J.O.: Reexamining the cluster hypothesis: scatter/gather on retrieval results. In: *SIGIR 1996: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 76–84. ACM, New York (1996)

6. Kurland, O.: The opposite of smoothing: a language model approach to ranking query-specific document clusters. In: *SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171–178. ACM, New York (2008)
7. Kurland, O.: Re-ranking search results using language models of query-specific clusters. *Inf. Retr.* 12(4), 437–460 (2009)
8. Kurland, O., Domshlak, C.: A rank-aggregation approach to searching for optimal query-specific clusters. In: *SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 547–554. ACM, New York (2008)
9. Kurland, O., Lee, L.: Corpus structure, language models, and ad hoc information retrieval. In: *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 194–201. ACM, New York (2004)
10. Lavrenko, V., Croft, W.B.: Relevance based language models. In: *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127. ACM, New York (2001)
11. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: *SIGIR 2008: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 235–242. ACM, New York (2008)
12. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 186–193. ACM, New York (2004)
13. Liu, X., Croft, W.B.: Evaluating Text Representations for Retrieval of the Best Group of Documents. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) *ECIR 2008*. LNCS, vol. 4956, pp. 454–462. Springer, Heidelberg (2008)
14. Lu, X.A., Ayoub, M., Dong, J.: Ad Hoc Experiments using Eureka. In: *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*, pp. 229–240 (1996)
15. Parapar, J., Barreiro, A.: Promoting Divergent Terms in the Estimation of Relevance Models. In: Amati, G., Crestani, F. (eds.) *ICTIR 2011*. LNCS, vol. 6931, pp. 77–88. Springer, Heidelberg (2011)
16. Rijsbergen, C.V.: *Information Retrieval*. Butterworths, London (1979)
17. Sakai, T., Manabe, T., Koyama, M.: Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)* 4(2), 111–135 (2005)
18. Tombros, A., Villa, R., Van Rijsbergen, C.J.: The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Process. Manage.* 38(4), 559–582 (2002)
19. Voorhees, E.M.: The cluster hypothesis revisited. In: *SIGIR 1985: Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 188–196. ACM, New York (1985)
20. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.* 22(2), 179–214 (2004)