

Automatic Control of Multiple Pseudo-Relevance Feedback Pass

Chakir Mokhtari

Computer Science Department
University Mušapha Stambouli of Mascara,
Mascara, Algeria
mokhtari.c@univ-mascara.dz

Mohammed Debakla

Computer Science Department
University Mušapha Stambouli of Mascara,
Mascara, Algeria
debakla_med@yahoo.fr

Boudjelal Meftah

Computer Science Department
University Mušapha Stambouli of Mascara,
Mascara, Algeria
meftahb@yahoo.fr

Abstract—Automatic query expansion (AQE) based on Pseudo Relevance Feedback (PRF) is a useful technique for enhancing the effectiveness of information retrieval systems. In this article, we propose the study of the behavior of a set of features based on the query and the set of feedback documents in the aim to control automatically the convergence of the query expansion process to the pertinent set of documents. It is a new look for which is commonly known in IR community as prediction of query performance or query ambiguity. The results obtained show that is very interesting to use these features as good predictors for any PRF methods.

Index Terms— Information retrieval, Automatic query expansion, pseudo-relevance feedback, query performance.

I. INTRODUCTION

Information retrieval (IR) system aims to find documents that are relevant to a user's information need. In general, the user's information need is typically expressed as a query consisting of a small number of words due to the difficult to formulate it by the user. In fact, the user's original query consisting of just a few terms is too short to describe the information need accurately leading to make answer really difficult and a poor coverage of the relevant documents [1]. To overcome this limitation, the user's original query is augmented by new features with a similar meaning which is commonly known in information retrieval community as automatic query expansion (AQE). The main purpose of AQE is to improve the retrieval effectiveness of document ranking by understanding automatically the query and produce related terms that are close to the original query [2]. AQE is based on Feedback documents that can come from user relevance judgments (relevance feedback, RF), the top retrieved documents for the initial query (pseudo-relevance feedback, PRF), or even the whole collection of documents or external resources such as thesaurus or ontology (global feedback) [3]. One of the most popular AQE technique inspired directly from relevance feedback is PRF method. PRF [4, 5, 6, 7, 8, 9, 10] aims to learn the user's requirements during a search session and by implicit user involvement adapt the behavior of the retrieval process. The basic idea is to reformulate the query such that it gets closer to the relevant documents by using the first documents retrieved in response to the original query as a more detailed description of the underlying query topic, from which to extract the most important terms to be used as expansion features. Since PRF is an important technique to improve the effectiveness of IR systems, it is widely proposed and studied for various retrieval models [4, 5, 11, 12, 13] and still attract the attention of IR community [14, 15, 6, 16, 17, 8].

The implementation of PRF methods always requires two rounds of retrieval. In the first round, an initial set of documents is retrieved with respect to an original query. And in the second round, another set of documents is retrieved based on an expanded query, which is generated by expanding the original query with relevant terms selected from the previously retrieved doc-

ument set. This second round could be repeated more than one time, but the challenge is how to ensure the effectiveness of the IR system at each time. In other word, we still perform the second round of feedback while the IR system retrieves more and more relevant documents. In the other hand, there have been interests in applying machine learning and clustering methods to select a subset of documents deemed representative of the subject of the query from the feedback documents [16, 18, 19, 20, 21]. Although it is well observed that cluster-based PRF improves effectiveness, however, it is interesting to reuse feedback documents more than once in order to bring more relevant and novel information to the user.

For this purpose, in this paper, we estimate the effectiveness of any PRF methods by studying the behavior of a set of features. This set of features can help us predict the effectiveness of PRF rounds and to avoid query drift. It is a new look for which is commonly known in IR community as prediction of query performance or query ambiguity [22]. The experimental results show that the proposed features have significant correlation with IR effectiveness in term of mean average precision. One of the major issues in PRF is selecting the most appropriate feedback documents to be used in the expansion phase. In our work, we consider PRF based on documents clustering.

The remainder of this paper has the following organization. We first give a background on pseudo-relevance feedback and query predicting effectiveness in Section II. Section III focuses on the formulation of the problem. Section IV and V present FPF classification algorithm and the set of the proposed features. Section VI proposes our study on how to predict the effectiveness of any PRF method. The evaluations are presented in section VII and we conclude this paper with a look at the future work in section VIII.

II. BACKGROUND

Pseudo relevance feedback learn from the top retrieved documents set to expand terms to the original query by assuming this set of documents as relevant documents when there are no user relevance judgments available. It is a significant technique in IR and is proved to improve the average effectiveness of a number of retrieval models for the past decades [14, 23, 24, 25, 26, 27]. But can also fail, particularly for ambiguous queries [28] [29]. And, it is observed in several works [18, 16] that the quality of selected expansion terms from top ranked documents (which may contain irrelevant information) have a strong effect on the effectiveness of PRF methods. Thus, the major issues are selecting the most appropriate documents from feedback documents. Clustering is one of techniques adopting to improve PRF and to overcome this problem. It is well studied in many researches [20, 19, 30, 31]. The main idea of clustering methods is to exploit the relationship between documents of the top ranked list and to detect clusters

with similar topics to the query.

In the other hand, estimating the quality of the retrieved documents set for a query is regarded as predicting the query performance. It is an important and challenging task in IR research [22, 32]. The central goal of this field of research is to identify features that can be used for predicting query performance. These features are categorized into pre-retrieval methods and post-retrieval methods [33].

Our work is between the field of PRF methods and the field of query performance prediction. Through this study, we want to identify a number of post-retrieval features to guide the evolution of any PRF method in the case where it is performed more than once which means if the first pass of PRF retrieve more relevant document (according to features) we perform another pass and so on. But, once no more relevant documents retrieved we stop the process of PRF.

III. PROBLEM FORMULATION

Let be an IR system R with any information model. For any given query Q and a document collection C, R returns a ranked list of documents L. Any PRF method assumes that the set S containing the N top-ranked documents of L are relevant and uses them to enhance the IR effectiveness by adding to the original query terms derived from these first N documents. Due to the fact that PRF has proven to be an effective strategy for improving retrieval accuracy in all retrieval models, why not don't repeat this strategy more than one time to retrieve more and more relevant documents. In the below, the circle at the left represents the retrieved documents, the other circle at the right the set of the relevant documents for the given query Q. As described in this figure, after the first PRF pass we are faced with two situations. In the situation A, our system R retrieves more relevant documents, but in the situation B no more relevant documents are retrieved. So, if this IR system detect that he is in the situation A, he still performs more and more PRF pass until to be in the situation B. In this case, the system stops.

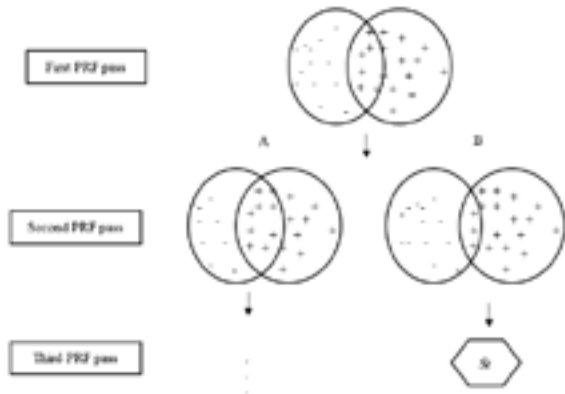


Fig.1. Cases of PRF pass.

Our goal is to estimate the contribution of any PRF method by studying the quality of the N top-ranked documents of L in relation to the query which will be rephrased it in order to retrieve more relevant documents. Since we use PRF as AQE method, we have in the top-ranked documents relevant and irrelevant documents. So, in order to detect the most useful subset from S for query expansion, we use clustering methods.

Since in PRF, the main unit is the expansion terms, we present each documents d in S by a vector of the term weights in S. We rank unique terms in S according to their Kullback-Leibler Divergence (KLD) weights. It has been shown that KLD is an ef-

fective weighting mechanism in many state-of-art AQE methods [34] [35]. Terms with higher KLD score are the more informative ones and then they will be well to represent documents. Other terms with lower score are ignored.

KLD measures divergence term's distribution in the set of feedback documents S from its distribution in the whole collection. The KLD weight is given by:

$$KLD(t) = p(t|S) \log_2 \left(\frac{p(t|S)}{p(t|C)} \right) \quad (1)$$

Where $p(t|S)$ is the probability of presence of term t in the set S of the N top-retrieved document, given by

$$p(t|S) = \frac{c(t, S)}{c(S)}$$

$c(t, S)$ is the frequency of the term t in S, and $c(S)$ is the count of terms in S. And $p(t|C)$ is the probability of presence of the term t in the entire document collection C. It is defined as follow:

$$p(t|C) = \frac{c(t, C)}{c(C)}$$

$c(t, C)$ is the frequency of the term t in collection C, and $c(C)$ is the count of terms in the whole collection C.

One of the major issues in PRF is selecting the most appropriate feedback documents to be used in the expansion phase. Document clustering is one way to select appropriate documents for AQE. Many clustering techniques have been developed and can be applied to clustering documents. The document clustering used in this study has been carried out using FPF (Furthest Point First) clustering algorithms which are presented below [36].

IV. FPF ALGORITHM

FPF algorithm is a polynomial clustering algorithm introduced by Gonzalez [36]. Unlike K-means algorithm, FPF algorithm doesn't choose randomly the k-centroids but it builds them iteratively. At each iteration, it calculates the cosine similarity defined above between documents and centroid within their cluster. The furthest document from the current set of centroids is chosen as new centroid and it will be added to this set of centroids. We can summaries the FPF algorithm in two steps, step of initialization and step of iteration as follow:

- Initialization
 - Pick an arbitrary document as centroid1, centroid of cluster 1,
 - Assign all documents of S to cluster 1,
- Iteration I ($i=2, \dots, K$)
 - Designate the furthest document from the current set of centroids as centroidi,
 - Assign to cluster i each document closer to centroidi than to its current centroid.

The similarity used in this work is Cosine distance. The Cosine distance between a document d and a center cte is defined as:

$$\text{Cosine}(d, cte) = \frac{a^*cte}{|d||cte|} \quad (2)$$

V. FEATURES

In order to predict the robustness of any PRF method and to perform it more than one time, we propose some features that capture intrinsic characteristics of the query and of the feedback documents selected from the N top-ranked documents via document clustering algorithm and their relationship. The proposed list of features is inspired from previous works [37] [38] and it is intuitive, simple to implement and have a very low computational cost. These features are mainly based on the query, the retrieved documents and the relationship between them.

- **Query Clarity (QC).** Query clarity is an intrinsic feature of a query. It has been shown that it predicts query ambiguity well [38, 39]. So, we expect it to predict the effectiveness of PRF method. In the definition, the clarity of a query is the Kullback-Leibler divergence of the query model from the collection model. To overcome the computational cost of this definition, we follow the work of [39]. It is a simplified formula of clarity and doesn't need the computation of the relevance scores for the query model which is defined as:

$$QC = \sum_{t \in Q} \log \frac{c(t, Q)}{c(Q)} \quad (3)$$

In this definition

$$p(t|Q) = \frac{c(t, Q)}{c(Q)}, \text{ where } c(t, Q) \text{ is}$$

the count of term t in the query Q , and $c(Q)$ is the total number of terms in Q .

- **Feedback Documents Clarity (FDC).** Similar to the query clarity used above, we define FDC as the Kullback-Leibler divergence of the feedback documents model from the collection model. Intuitively, a large value of this feature possibly means that the set of feedback documents on the same topic and far away from noise. Therefore, we hope that this feature gives us more information to perform positively another PRF pass. So, this feature is defined as:

$$FDC = \sum_{t \in S} p(t|S) \log \frac{p(t|S)}{p(t|C)} \quad (4)$$

- **Query Entropy (QE).** as mentioned in [37], the query is often very short, then, we compute query entropy based on the set of feedback documents S . It is defined as follows:

$$QE = \sum_{t \in S} -p(t|S) \log_2 p(t|S) \quad (5)$$

- **Selected Documents Entropy (SDE).** In our work, the PRF process is increased by a clustering phase in order to detect the most useful documents for query expansion. At this stage, each cluster is treated as a large document constructed by concatenating all documents belonging to it. So after forming clusters, the clusters are ranked by the query-likelihood retrieval model [40]. This retrieval model scores a document d with the respect to a query Q as follows:

$$p(Q|d) = \prod_{t \in Q} p(t|d) \quad (6)$$

$$\text{where } p(t|d) = \frac{|d|}{|d| + \mu} p_{ML}(t|d) + \frac{\mu}{|d| + \mu} p_{ML}(t|C)$$

$$\text{and } p_{ML}(t|d) = \frac{c(t, d)}{c(d)}, p_{ML}(t|C) = \frac{c(t, C)}{c(C)}$$

- $c(t, d)$ is the count of the term t within the document d and $c(d)$ is the total number of the document d . μ is a smoothing parameter used to estimate non-zero values for terms in the query which are not in a document.

The documents belonging to the top ranked cluster constitutes the set of the selected documents for query expansion. We call this set SD . So, we get the entropy of the selected documents by:

$$SDE = \sum_{t \in SD} -p(t|SD) \log_2 p(t|SD) \quad (7)$$

- **Feedback Documents Broadness (FDB).** To compute this feature, we use the set of the N top retrieved documents S . Intuitively, if the documents in this set are more discriminating, it may mean that they focus more on the same topic and away from the noise. The broadness of the set S is defined as the average distance between each document and the centroid of the set S .

$$FDB = \frac{1}{|S|} \sum_{d \in S} \sum_{w \in d} p(w|\theta_d) \log \frac{p(w|\theta_d)}{p(w|\theta_{centroid})}$$

where

$$p(w|\theta_{centroid}) = \frac{1}{|S|} \sum_{d \in S} p(w|\theta_d) \quad (9)$$

$$\text{and } p(w|\theta_S) = \frac{c(w, S)}{|S|}$$

VI. STUDY OF THE DEPENDENCE BETWEEN THE FEATURES AND PRF PASS

The central problem in pseudo relevance feedback is how to select the appropriate documents from the top-retrieved documents since not all these documents are relevant and there is a high chance that we use irrelevant documents for expansion. To overcome this issue, we use clustering techniques which are proved as an effective way. We construct clusters using FPF algorithm [41].

So, after the selection of the most useful documents, the query is expanded with new terms and another set of documents are retrieved. At this stage, if we suppose that the expanded query presents well the user's information need then the novel set of retrieved documents must contain more relevant documents. In this case, another PRF pass will be beneficial. For this purpose we study the behavior of the set of features described above (section 4). After each PRF pass we compute the features QC, FDC, QE, SDE, DSD_S and FDB.

Our studying procedure follows the five steps below:

- Step 1: For a given query Q retrieve a set S of the N top-ranked documents according to their system relevance.
- Step 2: Split the set S in K clusters using the FPF clustering algorithm.
- Step 3: Rank the resulting K clusters according to their score using formula (6).
- Step 4: Use the K' top ranked clusters as the set of documents to be used for feedback pass.
- Step 5: compute the set of features QC, FDC, QE, SDE and FDB.

These steps are performed for many times in order to observe the features behavior according to IR system performance. In our experiments, we have performed five pass.

VII. EVALUATION

In this section, we report our experimental investigation of the proposed scheme. The objective of our experiments is to investigate how we can predict the positive contribution of any PRF method in unsupervised manner by studying the set of the

features described above.

A. Experimental Setup

We utilize in our study the TREC ad-hoc test collection (Disks 4, 5). The data set contains about 528,000 news documents from various sources such as Federal Register, Financial Times, etc..., which are considered as high-quality text data with minimum noise. On average, each document includes 557 terms. The topic set used constitutes the TREC 6 ad-hoc task topics. The topics are numbered from 301 to 350, each one consists of three fields. These are the title, which includes the most related words (1-3 words); the description, which gives a wider explanation about the query; and a narrative, which contains specific conditions on accepting or rejecting documents. The 50 ad-hoc queries were taken from only title field of the corresponding query topics because it is a realistic setting that reflects the concise nature of real user's queries [35]. The average query length in TREC 6 is 2.64.

We use the full-featured text search engine library Lucene¹, an open source IR framework. For all queries and documents in the test collections, each term is stemmed using Porter's English stemmer [42], and stop words are removed.

To measure the correlation of each feature and its behavior with the query performance, we use the Mean Average Precision (MAP) as the focus measure representing the queries performance. MAP is defined as

$$MAP = \frac{1}{|NQ|} \sum_{j=1}^{|NQ|} AP_j \quad (10)$$

Where $|NQ|$ is the number of queries and AP_j is the average precision for a given query Q_j is defined as:

$$AP = \frac{1}{n_j} \sum_{k=1}^{n_j} P@k \quad (11)$$

n_j is the number of relevant documents for Q_j . $P@k$ is the precision at k , it is the fraction of the top k retrieved documents that are relevant.

B. Results

The main aim of this work is to study the correlation between the set of features described above and the IR system effectiveness.

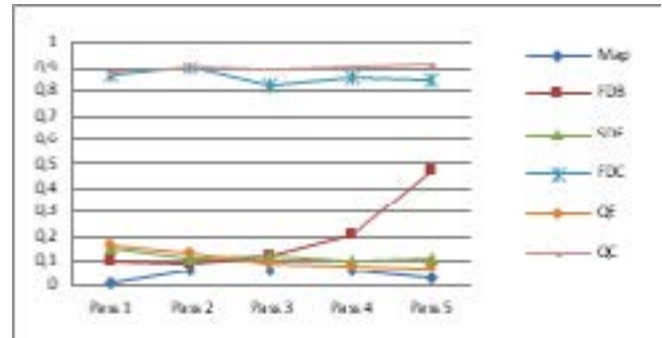
A well-known problem in a typical pseudo relevance feedback algorithm is its instability with respect to the variation of two major parameters, namely, the number of feedback documents and the number of terms [1]. Since our purpose is not to show the contribution of the classification task in the retrieval phase, we have fixed the number of cluster in all experiments at five and we have varied the values of number of documents and terms. Figures 1, 2 and 3 present results when then number of feedback documents is fixed to 20, 50 and 100 respectively. At each figure, the two graphics show the correlation between features and MAP when number of terms is 5 and 10 respectively.

All experiments show that the features forms two classes, one class is formed by FDC and QC and the other class is formed by FDB, QE and SDE. When the MAP increases the features of the first class increases proportionally but features of the second class decreases. We note that, FDB is the most sensitive feature to the MAP variation. It is logical since when the feedback documents forms a cluster of a high quality IR system retrieves more relevant documents and consequently the MAP increases in general.

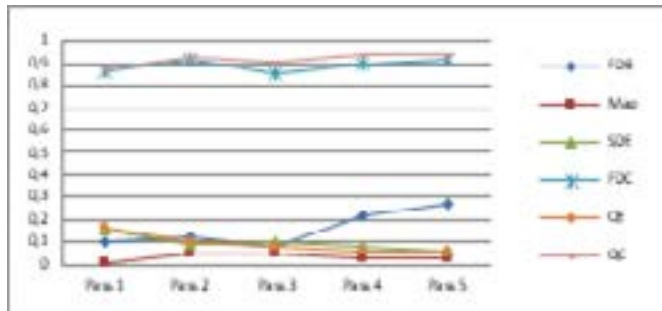
VIII. CONCLUSIONS

This paper presented the results of an experimental study of some features in the aim to control automatically any PRF methods. We have Remarque that we can control this scheme of PRF by just monitoring the behavior of a set of features. We believe that the main reason for this result is that when the PRF method is effective the set of feedback documents contributes positively in the IR system performance and has lower entropy and so the expanded query entropy.

We plan to pursue several directions in our future work. We will investigate the robustness of this proposition by doing more extensive experiments in different ways. Moreover, further works will focus on developing more efficient

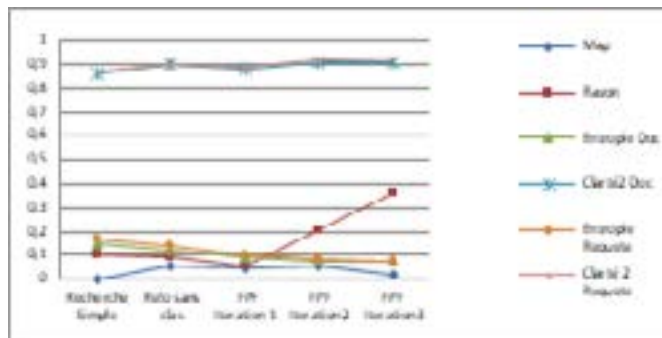


(a): #terms=5



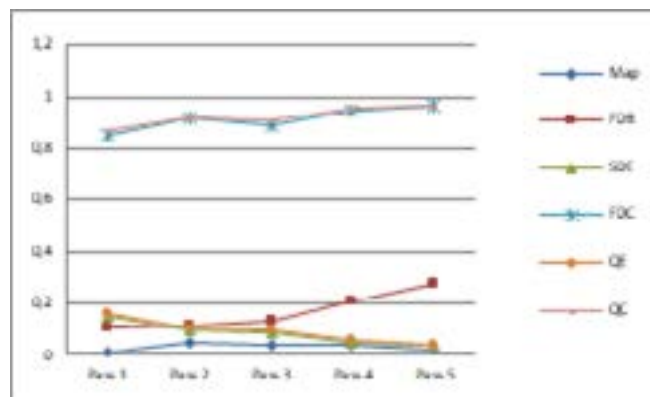
(b): #terms=10

Fig.2. behavior of FDB, SDE, FDC, QE and QC according to MAP, #doc=20.



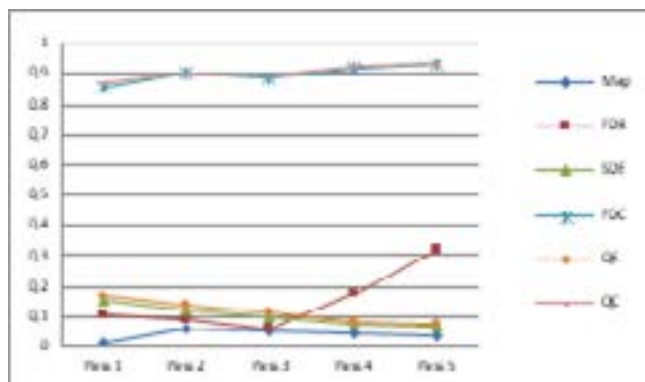
(a): #terms=5

¹ <https://lucene.apache.org>

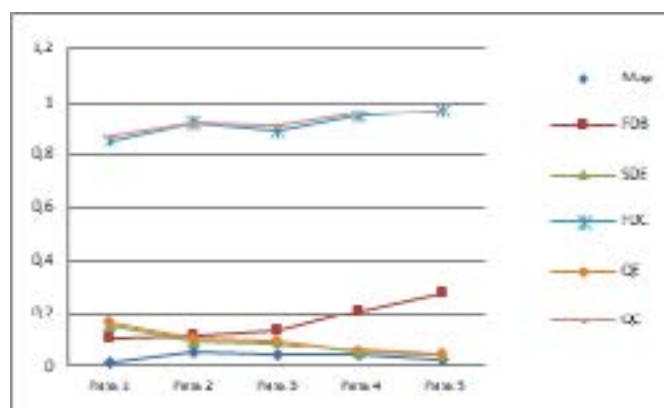


(b): #terms=10

Fig.3. behavior of FDB, SDE, FDC, QE and QC according to MAP, #doc=50.



(a): #terms=5



(b): #terms=10

Fig.4. behavior of FDB, SDE, FDC, QE and QC according to MAP, #doc=100.

REFERENCES

- [1] Carpineto, C. and Romano, G. (2012) 'A Survey of Automatic Query Expansion in Information Retrieval', ACM Computing Surveys, Vol. 44 No. 1, pp. 1-50.
- [2] Manning, D., Christopher, P. R. and Schütze, H. (2008) 'An Introduction to Information Retrieval', Cambridge University Press, England.
- [3] Baeza-Yates, R. A. and Ribeiro-Neto, B. : Modern Information Retrieval. Pearson Education Ltd., 2 edition. 2011.
- [4] V. Lavrenko and B. Croft. Relevance-based language models. In Proceedings of SIGIR'01, pages 120-127, Sept 2001.
- [5] J. Rocchio. Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing, pages 313-323. Prentice-Hall Inc., 1971.
- [6] Y. Lv, C. Zhai, and W. Chen. A boosting approach to improving pseudo-relevance feedback. In Proceedings of SIGIR'11, 2011.
- [7] J. Miao, J. Huang, and Z. Ye. Proximity-based rocchio's model for pseudo relevance. In Proceedings of SIGIR'12, 2012.
- [8] Singh, J., and Sharan, A. "A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach". Neural Computing and Applications, 1-24. 2016.
- [9] Aly, A. A. "Using a query expansion technique to improve document retrieval". International Journal Information Technologies and Knowledge, 2(4), 343-348. 2008.
- [10] Vaidyanathan, R., Das, S., and Srivastava, N. "Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval". arXiv preprint arXiv:1502.05168. 2015.
- [11] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 44(4):288-297, 1990.
- [12] C. Zhai and J. Lafferty. Model-based feedback in the KL-divergence retrieval model. In Tenth International Conference on Information and Knowledge Management (CIKM 2001), pages 403-410, 2001.
- [13] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 18:79-112, 2000.
- [14] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In Proceedings of SIGIR'10, 2010.
- [15] B. He and I. Ounis. Finding good feedback documents. In CIKM '09, pages 2011-2014, 2009.
- [16] Huang, J. X. Miao, J. and He, B. (2013). High performance query expansion using adaptive co-training. Information Processing & Management, Vol. 49 No. 2, pp. 441-453.
- [17] Pal, D. Mitra, M., Bhattacharya, S.,: Improving Pseudo Relevance Feedback in the Divergence from Randomness Model. ICTIR2015: Proceedings of the 2015 International Conference on The Theory of Information Retrieval pp. 325-328, September 27-30, 2015, Northampton, MA, USA.
- [18] Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). 'Selecting good expansion terms for pseudo-relevance feedback'. In Proceedings of 31st annual international ACM conference on research and development in information retrieval. pp. 243-250. July 20 - 24, Singapore, Singapore.
- [19] Lee, K.S.; Croft, W.B. A deterministic resampling method using overlapping document clusters for pseudo-relevance feedback. Information Processing and Management. 2013, 49, 792-806.
- [20] Gelfer K., I.; Kurland, O. Cluster-based query expansion. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09, Boston, MA, USA, 19-23 July 2009; pp. 646-647.
- [21] Levi, O., Raiber, F., Kurland, O., Guy, I. : Selective Cluster-Based Document retrieval. CIKM 2016: 1473-1482.
- [22] Lang, H., Wang, B., Jones, J., Li, J.-T., Ding, F., Liu, Y.-X.: Query Performance Prediction for Information Retrieval Based on Covering Topic Score. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 23(4): 590-601 July 2008.
- [23] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In SIGIR 2006, pages 162-169.
- [24] Buckley, C., Salton, G., Allan, J.: Automatic retrieval with locality information using smart. In: TREC. (1992) 59-72
- [25] R. Udupa, A. Bhole, and P. Bhattacharya. A term is known by the company it keeps: On selecting a good expansion set in pseudo-relevance feedback. In Proceedings of ICTIR 2009, Advances in IR Theory. Springer, 2009.
- [26] E.K.F. Dang , R.W.P. Luk , J. Allan , Fast forward index methods for pseudo-relevance feedback retrieval, ACM Trans. Inf. Syst. 33 (4) (2015) 1-33.
- [27] B. He , I. Ounis , Finding good feedback documents, in: Proceedings of CIKM, 2009, pp. 2011-2014 .
- [28] G. Amati, C. Carpineto, G. Romano, and F. U. Bordoni. Query difficulty, robustness and selective application of query expansion. In Proc. of ECIR, pages 127-137, 2004.
- [29] Singh, J., and Sharan, A. "Relevance Feedback Based Query Expansion Model Using Borda Count and Semantic Similarity Approach". Computational intelligence and neuroscience, 2015.
- [30] Sakai, T., Manabe, T., & Koyama, M. (2005). Flexible pseudo-relevance feedback via selective sampling. ACM Transactions on Asian Language Information Processing (TALIP), 4(2), 111-135.
- [31] Huang, Q.; Song, D.; Ruger, S. Robust query-specific pseudo feedback

- document selection for queryexpansion. In Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, ECIR'08, Milton Keynes, Glasgow, UK, 30 March–3 April 2008; pp. 547–554.
- [32] Raiber, F., Kurland, O., : Query-performance prediction: setting the expectations straight. SIGIR 2014: 13-22
- [33] Hauff, C., Hiemstra, D., de Jong, F., : A survey of pre-retrieval query performance predictors. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pages 1419–1420, New York, NY, USA. ACM.
- [34] 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.
- [35] Z. Ye, X. Huang, B. He, and H. Lin, (2009). York university at TREC 2009: relevance feedback track. In Proceedings of TREC 2009.
- [36] Gonzalez, T. F., : Clustering to minimize the maximum intercluster distance. Theoretical Comput. Sci., 38:293-306, 1985.
- [37] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In CIKM '09, 2009.
- [38] Cronen-Townsend, S., Zhou, Y., and W. B. Croft. Predicting query performance. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 299 - 306, Tampere, Finland, 2002.
- [39] He, B., Ounis, I., Query performance prediction. Inf. Syst. 31(7): 585-594. 2006.
- [40] Liu, X., Croft, W. B.,: Cluster-based retrieval using language models. In Proc. 27thACMSIGIR conference on research and development in information retrieval, pp. 186–193, 2004.
- [41] Filippo Geraci, Marco Pellegrini, Paolo Pisati, and Fabrizio Sebastiani. A scalable algorithm for high-quality clustering of Web snippets. In Proceedings of SAC-06, 21st ACM Symposium on Applied Computing, pages 1058–1062, Dijon, FR, 2006.
- [42] Porter, M. F., : An algorithm for suffix stripping. In Readings in Information Retrieval, 313–316. 1997.