Query Expansion using PRF-CBD approach for Documents Retrieval

R.Rajendra Prasath* and Sudeshna Sarkar†

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur - 721302, India.

Abstract

Query Expansion has been widely used to improve the effectiveness of documents retrieval. In this work, we have attempted to identify additional terms for query expansion, from the initial set of documents retrieved for the original query, with the help of Clustering-by-Directions (CBD) algorithm proposed by Kaczmarek[1]. The CBD algorithm is based a tag cloud of associated terms that are located in a radical arrangement and provides a clue to the direction of user intent in which search can be continued effectively. The output of the CBD approach gives rise to a set of terms in which we have considered top k terms for expanding the given query. The importance (weighting) of these selected expansion terms are computed with respect to the number of terms in the radical of the selected directions. The experiments were conducted on FIRE 2012 adhoc data collection and we have performed monolingual documents retrieval in 3 major languages: Bengali, Hindi and English.

Keywords: Pseudo Relevance Feedback, Query Terms Expansion, Clustering-By-Directions, Query Terms Weighting

1 Introduction

"Search for information" is one of the main activities for every human in his/her day to day life. It is difficult to imagine a world with the absence of search engines that strieve to satisfy the information needs of different users on different aspects. When a user enters his information need in the form of a query composed of a set of keywords, a search engine is required to retrieve relevant information with respect to the actual intent of the query, not just based on weighted / non-weighted term matching. Thus, on many searches, the expectation of the people searching for specific information is

*e-mail: rajendra@cse.iitkgp.ernet.in;
†e-mail: sudeshna@cse.iitkgp.ernet.in

not fulfilled merely by the subset of documents retrieved for the supplied set of keywords. In contrast, the actual user intent of query has to be captured from the supplied keywords. So this tasks leads us to the problem of query expansion so as to have a deep understanding on the actual user intent.

In this work, we have considered the initial set of documents as Pseudo Relevance Feedback and then applied the tag cloud based clustering by directions approach, for finding the terms for query expansion, that shows potential directions in which the search can be continued.

2 Query Expansion

Inspiried by the tag cloud based clustering-by-directions approach, we proposed documents retrieval system in which the PRF-CBD approach is employed for query terms expansion.

2.1 Pseudo Relevance Feedback

Documents retrieved by an IR system pertaining to a query may either be relevant or irrelevant. Relevance Feedback (RF) systems require users to mark the initial set of retrieved documents as relevant or irrelevant. IR system uses such manual feedback from users to further improve the performance of documents retrieval. To get improved retrieval performace without an extended interaction of users, Pseudo Relevance Feedback (PRF) [2] based methods have been proposed. The idea of PRF is to provide a mechanism for automatic local analysis that helps to find a better representative set of documents for the given query[2]. The PRF does documents retrieval to find an initial set of documents for a given query. The evidence captured from the content analysis of the initial set of retrieved documents is considered as PRF to CBD algorithm that identifies a good set of representative terms for query expansion.

2.2 Clustering-by-Directions algorithm

The primary objective of the clustering-by-directions algorithm[1] is neither to divide search results into clusters nor to show the related terms on the basis of some thesauri. It is applied to indicate the directions in which the search can be continued. To achieve this, the CBD algorithm first selects different directions, and afterward, it determines how the user can move in each direction. It is done regardless if there are subsets of web pages with the similar subject or not.

The basic steps in the clustering by directions algorithm are as follows:

1. calculate vectors which represent documents and distances between these vectors;

2. select different directions;
3. assign documents to directions and select terms which represent directions;
4. select top k terms as candidate terms for query expansion

Since documents are represented as vectors in the Vector Space Model(VSM), choosing the candidate direction is complicated due to a large number of dimensions in VSM. In order to choose representations serving as directions, the subset in which the sum of distances is the greatest is selected.

3 Proposed IR system

The architecture of the proposed IR system with PRF-CBD approach is given in Figure. 1. The proposed IR system consists of two major

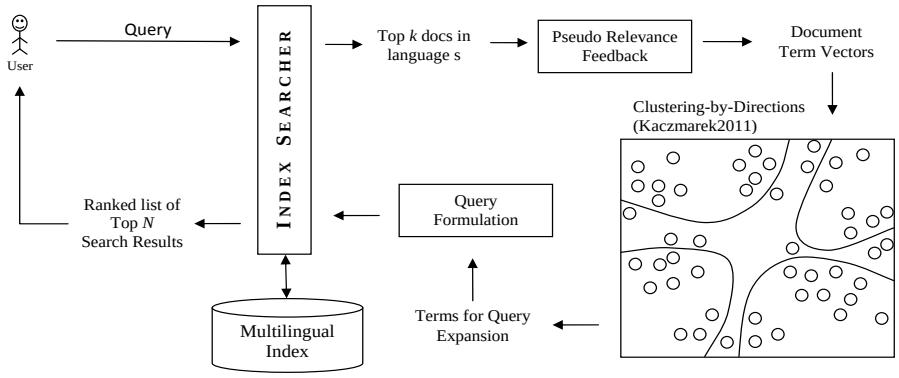

Figure 1: The proposed PRF-CBD based IR system

components:

(a) *Query Expansion:* Use the Clustering-by-Directions algorithm to choose the probable query terms and perform term weighting to formulate the expanded query

(b) *Ranking Documents:* We apply BM25 as the ranking function to compute the score of a document given the query.

3.0.1 Query Expansion

Intially we retrieve a set of representative documents pertaining to the given query. These initial set of documents are considered to be relevant. Then we apply CBD algorithm on the terms vectors of these documents and get a list of terms for each of the selected directions among d directions. Using these list of terms and their term statictics in top k documents, we

formulate the weighted query, ensuring that atleast one top ranking term from each direction exists in the reformulated query.

3.0.2 Ranking Documents

We have used Okapi BM25 [4, 3] as our ranking function. BM25 retrieval function ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. Given a query Q, containing keywords $q_1, q_2, \cdots, q_n$, the BM25 score of a document D is computed as:

$$score(Q, D) = \sum_{i}^{n} idf(q_i) \cdot \frac{tf(q_i, D) \cdot (k_1 + 1)}{tf(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdoclength}})}$$

where $tf(q_i, D)$ is the term frequency of q_i in the document D; $|D|$ is the length of the document D and *avgdoclength* is the average document length in the text collection; k_1, $k_1 \in \{1.2, 2.0\}$ and b, $b = 0.75$ are parameters; and $idf(q_i)$ is the inverse document frequency of the query term q_i.

The inverse document frequency $idf(q_i)$ is computed as:

$$idf(q_i) = \log \frac{N - df(q_i) + 0.5}{df(q_i) + 0.5}$$

where N is the total number of documents and $df(q_i)$ is the number of documents containing the terms q_i.

4 Experimental Results

4.1 Corpus and Topics

We have performed experiments with FIRE 2012 adhoc data collection on three languages: Bengali, Hindi and English. The coverage of documents in each collection is listed in Table. 1. We have used the query topics provided for each of these languages and also listed the details in Table. 1. We have submitted two runs (using title, desc fields) for each of the monolingual retrievals in Bengali, Hindi and English.

Language	**# documents**	**# terms**	Topic IDs
Bengali	500,122	2,497,978	176-225
Hindi	331,599	1,164,526	176-225
English	392,577	1,427,986	176-225

Table 1: FIRE 2012 adhoc dataset

4.2 Analysis

In this section, we illustrate our observations in detail. We report here only the runs of the monolingual documents retrieval in Bengali, Hindi and English. Since the relevance judgments of many Bengali documents are missing in evaluated pool, we are unable to get a clear picture from the partial results of Bengali monolingual runs. So we included only the partial results in the Precision - Recall plot. Figure. 2 shows the P-R curve of three monolingual retrievals: Bengali, Hindi and English, for the set of fire topics ranging from 176 to 225. In this experiment, we have used title as the query and expanded it with PRF-CBD approach.

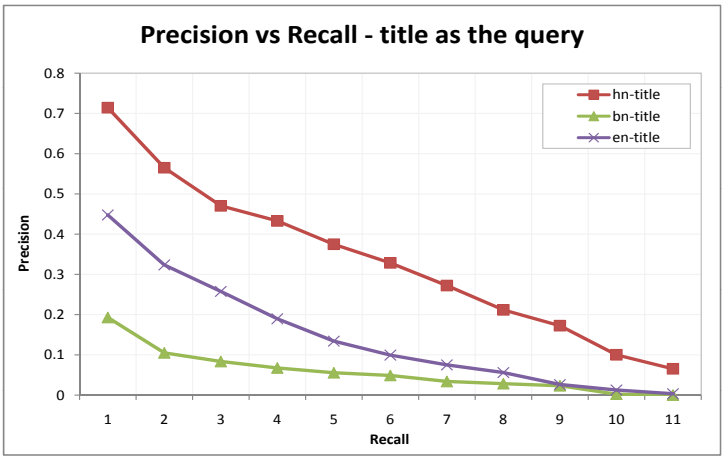

Figure 2: P-R curve: title used as the query in Bengali, Hindi and English monolingual retrieval

In the next experiment, we have used desc field as the query and expanded it using PRF-CBD approach. In bengali Monolingual Retrieval, we observed that for the query id: 225 (equivalent Query in English: "Satanic verses controversy"), PRF-CBD approach is able to present terms that capture the variety of news items on this topic. In Hindi monolingual retrieval with title field as the query, several queries, notably query ids: 179, 182, 184, 191, 199, 222, 223, 225 achieved a mean average precision greater than 0.7 and 60% of the query topics achieved map value of 0.5 or greater. Similar trends were observed when desc is used as the query in Hindi Monolingual Retrieval. In English Monolignial Retrieval, almost 12 queries achieved map value of 0.25 and above with title as the query topic. The distribution of English monolingual retrieval with desc field resulted in a slightly degraded performance in terms of map score.

We will subsequently carry out query wise detailed analysis of retrieval effectiveness of all these monolingual official runs.

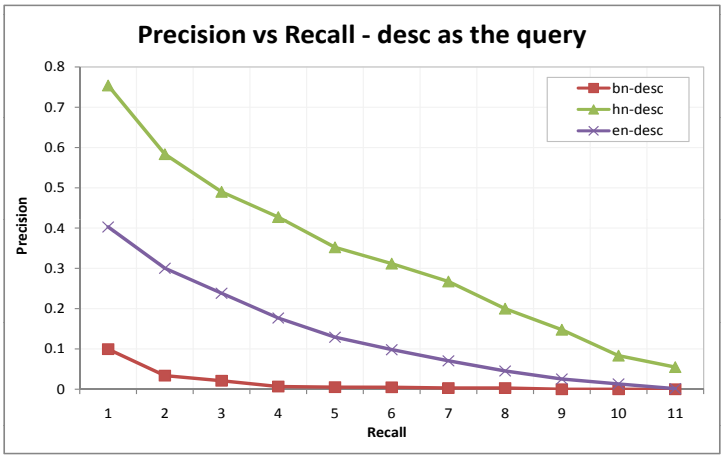

Figure 3: P-R curve: title used as the query in Bengali, Hindi and English monolingual retrieval

5 Conclusion

In this work, we have proposed an approach to find a better representative terms for query expansion for effective document retrieval. We have tested the proposed monolingual documents retrieval system with PRF-CBD approach on FIRE 2012 adhoc data in three languages, viz., Bengali, Hindi and English. The proposed system performs better for many queries by finding a set of good representative terms. However, the derieved results are sensitive to initial retrieved set of documents on which the CBD algorithm is applied.

References

[1] A. Kaczmarek. Interactive query expansion with the use of clustering-by-directions algorithm. *Industrial Electronics, IEEE Transactions on*, 58(8):3168 –3173, Aug. 2011.

[2] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[3] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, Apr. 2009.

[4] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.