Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval

Rekha Vaidyanathan, Sujoy Das, Namita Srivastava
v_rekha@hotmail.com, sujdas@gmail.com, sri.namita@gmail.com
Dept. of Computer Applications, MANIT, Bhopal

ABSTRACT
Query Expansion using Pseudo Relevance Feedback is a useful technique for reformulating the query. In this paper, expansion terms are obtained by combining pseudo relevance feedback and equi-frequency partition of the documents with tf-idf scoring technique. It is observed that the groups of words that have same tf-idf score as that of query terms are better candidate words for query expansion instead of those words that are having highest tf-idf. The experiment is performed on FIRE 2011 Hindi test collections, using Terrier retrieval engine. The result shows an improvement of 2.2% in mean average precision.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, relevance feedback, retrieval models.*

General Terms
Algorithms, Experimentation

Keywords
Query Expansion, Pseudo Relevance Feedback, Partition, tf-idf

1. INTRODUCTION
Query Expansion aims at improving the effectiveness of the search results. The purpose of Query Expansion is to reformulate a query by introducing terms that are closely related to original query [1]. The most challenging task is to *automatically* understand the query and produce related words that are close to the original query. In literature, various techniques have been proposed by researchers. Psuedo Relevance Feedback (PRF) is one such technique where the user submits a query and the search engine retrieves the r*elevant* documents using some retrieval method. The top 10 retrieved documents are then used to reformulate the query using additional terms, apart from original query terms, for better results. In this paper, we propose to obtain expansion terms by combining pseudo relevance feedback and equi-frequency partition with tf-idf scoring technique. The frequency used for partitioning, is derived using equi-width partition technique that gives the maximum frequency of query term(s) in a partition. The terms are then weighted *within* the document using tf-idf before obtaining the weighted expanded query terms. The list of terms and scores are then sorted in descending order. The terms are then selected in two ways, Method-1 and Method-2. In Method-1, the group of words that has the highest score is selected for expansion. In Method-2, the group of words that has the same score as that of the query term (*highest score among the query terms*) is selected for expansion.

2. RELATED WORKS
Automatic query expansion is difficult to perform as reformulation may cause query drift. Normally, tf-idf is measured to extract keywords that appear frequently in a document [2]. Galeas et al. observed that two words that are near to each other in a set of documents may possess some semantic relationship. They have used position and frequency of words in a document to calculate their inter-quartile ranges to get the dispersion of the words within a document using Fourier series expansion [4, 5]. Markus Holi et.al proposed an extension of tf-idf weighting method and observed that their weighting method produced ranking that matched human judgment [3]. In our previous paper [7], frequency and position of query terms was used effectively to find the densest region of a document that has the most relevant terms to that of the query terms.

3. METHODOLOGY
In this paper, expansion terms are obtained by combining pseudo relevance feedback and equi-frequency partition with tf-idf scoring. After the initial retrieval, the top 10 documents are partitioned and terms are weighted using tf-idf before reformulating the query. While calculating tf-idf score, ***partitions obtained through equi-frequency partition*** is considered as a complete entity instead of retrieved document. Thus the idf formula is changed for ***partitions*** than documents to obtain the score of words *within* a document. A sorted list of words based on this score is generated. In Method-1, the group of words that has highest score is selected for expansion. In Method-2, the groups of words that have same score as that of the *query term(s)(highest score among the query terms)* is selected for expansion. It is observed that the groups of words that have the same tf-idf score as that of the query terms are better candidate words for query expansion instead of those words that are having highest tf-idf. The comparative study shows Method-2 improves efficiency marginally whereas in case of Method-1, average precision is dropped.

3.1 Algorithm
1. Partition each retrieved document into 10 partitions (deciles) (d_i *is referred as a partition of a Document D*) **after removing the stopwords.**

   a. For each partition d_i, calculate total frequency of query terms falling within that partition.
   b. Obtain the frequency of the keyword that appears maximum (Set it to $\mathbf{f}_{score_{max}}$).

2. Partition each retrieved document using **equi-frequency partitioning**. The frequency is a derived constant k, calculated

using formula, $\mathbf{k=\sum f_i/f_{scoremax}}$, where $\sum f_i$ is the total frequency of keywords in the document d and $f_{scoremax}$ is *frequency of the keyword that appeared maximum number of times* from equi-width partition. We assume, k, is such that any region of a document will have *atleast* k number of keywords.

3. **Assuming that each Partition is equal to a document in the original formula**, the **tf-idf** is calculated for each Partition.
   tf(t,p)=f(t,p)/max{f(w,p):w Є p}
   idf(t,P)=log[|P|/|{pЄP:tЄp}|]
   With |P|= total number of partitions in a Document
   {pЄP:tЄp}= number of partitions where term t appears
   tfidf(t,p,P) =tf(t,p) * idf(t,P)

4. Sort the words with maximum score and write it into a keywords file

5. Extract additional terms or expansion words for reformulating the query
   a. Method-1: Pick the highest score words as expansion words. (This may or may not contain keywords)
   b. Method-2: Pick those words that share the same score as that of ***keyword with highest score*** in the sorted score list. If no keyword appears in the list, we do not expand the query.
6. Resubmit the query using these expansion words.

3.2 Experiment and Discussion

The experiment is performed on FIRE 2011 Hindi test collections, using Terrier retrieval engine. The query is formulated using "*title*" field of FIRE 2011 Hindi test collection.

The expanded terms using the two methods is shown in Table 1 for topic no. 155 for query <title>**मुंबई ताज हमला** </title

Table 1. Expanded query terms along with their tf-idf score for Method-1 and Method-2

Method	Score	Expanded terms
Method-1	0.845098	इलाकों, जगह, छोटे, ब्लास्ट, होटेल
Method-2	0. 477121	**ताज,** आतंकवादी, स्टेशन, जमीन, गोलीबारी, खिलाड़ियों, बहरहाल

Experiments were conducted on FIRE 2011 Adhoc Hindi data and the run is referred as HTRUN.
The Table 2 shows the Mean Average Precision after conducting the experiments using both the methods.

Table 2. Comparison of Average Precision incase of before and after Query Expansion using Method 1 & 2 for HTRUN

	Before Expansion	Method1	Method 2
No. Of Queries	50	50	50
Retrieved	49907	50000	50000
Relevant	2885	2885	2885
Relevant Retrieved	2059	1736	2050
Average Precision	**0.2453**	**0.1895**	**0.2507**

It is observed that **Method-1** resulted in lower mean average precision as compared to **Method-2** and original query where as **Metho-2** proved to be a better one as mean average precision improved by 2.15%.

4. CONCLUSION

In this paper, expansion terms are obtained by combining pseudo relevance feedback and equi-frequency partition with tf-idf scoring technique. While calculating **tf-idf** score, partition is considered as a complete entity instead of the retrieved document. The experimental results show that the group of words that have the same tf-idf score as that of query terms are better candidate words for query expansion. Further, merely taking the highest scored words based on tf-idf, after partitioning the initial top 10 retrieved documents, based on equi-frequency partition and tf-idf may not necessarily improve retrieval effectiveness. We are carrying out further study and analysis to improve the retrieval performance by modeling the same approach in different ways.

5. ACKNOWLEDGMENTS

Our thanks are due to Terrier™ team for providing free software for researchers in the field of IR [6]. Our sincere thanks to FIRE group allowing us to use the data for our experiment.

6. REFERENCES

[1] Rosie Jones, Benjamin Rey and Omid Madani, C.2006. Generating Query Substitutions. *Proceedings of the 15th international conference on World Wide Web (2006) ACM* http://portal.acm.org/citation.cfm?id=1135835

[2] Abdelmgeid Amin Aly, J. 2008. Using a Query Expansion Technique to improve Document Retrieval, *International Journal "Information Technologies and Knowledge*" Vol.2 / 2008 343 (2008). www.foibg.com/ijitk/ijitk-vol02/ijitk02-4-p07.pdf

[3] Markus Holi, Eero Hyvonen and Petri Lindgren, C.2006. Integrating tf-idf weighting with Fuzzy view based search. In *Proceedings of the ECAI 2006 Workshop*. https://www.haiti.cs.uni-potsdam.de/proceedings/ECAI-06/Workshops/w09/holi06.pdf.

[4] Patricio Galeas, Ralph Kretschmer, Bernd Freisleben C. 2009. Document Relevance Assessment via Term Distribution Analysis Using Fourier Series Expansion. In JCDL '09: *Proceedings of the 2009 Joint International Conference on Digital Libraries, pages 277–284, New York,*

*NY, USA, 2009.*ACM dl.acm.org/citation.cfm?id=1555446

[5] Patricio Galeas and Bernd Freisleben. J. 2008. Word Distribution Analysis for Relevance Ranking and Query Expansion. Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, 2008, Volume 4919/2008, 500-511 (2008). dl.acm.org/citation.cfm?id=1787632

[6] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, Christina Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform.*C.2006.OSIR,2006 ACM SIGIR Conference (2006)*

[7] Rekha Vaidyanathan, Sujoy Das, Namita Srivastava, Query Expansion based on Equi-Width and Equi-frequency Partition , Working note at Forum for Information Retrieval Evaluation (FIRE) Workshop -2011, IIT Bombay.

(http://dl.dropbox.com/u/10053028/FIRE2011_working-notes.pdf refer Pg 22)