Journal of Engineering and Applied Sciences 14 (Special Issue 8): 10398-10401, 2019
ISSN: 1816-949X

Query Expansion Based Proximity and Distributional Methods

Hadeel D. Abdulameer and Wafaa M. Saeed
Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

Abstract: Automatic query expansion is an effective way used to improve the efficiency of the Information Retrieval system (IR) based on the Pseudo Relevance Feedback (PRF). The first-pass retrieval contains relevant and irrelevant documents. In this study, we used different algorithms to select the terms from top retrieved documents which used for expansion query. Where proximity term frequency methods based on kernel function was used to show the semantic relationship between two words and distributional methods KLD and CHI square methods were used to show the divergence of the terms in two spaces. Then we used borda combined method. After that, we used reweighting method to calculate the new weight of the candidate terms which used to expand the query. Our result produced tangible improvement in IR system.

Key words: Information retrieval, pseudo relevance feedback, query expansion, proximity term frequency, KLD, CHI square

INTRODUCTION

Information Retrieval (IR) is a device that satisfies user's information needs which were represented by a query, by retrieving most relevant documents from a collection (Manning *et al*., 2010). With the exponential growth in the amount of information sources on the internet, the need for retrieval system has increased. The quality of the search depends on the quality of the user information represented by the queries. Commonly in web search, the queries provided by the user are often short, two or three word at most with a huge heterogeneous document collections. These led to increase limitations of current information retrieval systems because of the difficulties of dealing with synonymy (describing the same things by different word) and polysemy (describing the different things by the same word). Consequently, the system may flop to retrieve the relevant documents by retrieving a very little relevant documents and irrelevant documents (Carpineto *et al*., 2001).

A simple approach to solve this problem is automatic query expansion which based on the gathering extra information from the relevance feedback documents and reformulation the original query. As result creating new query automatically. Relevance feedback is an effective approach used for improving IR system effectiveness and modify the query. There are three type of relevance feedback implicit, explicit and Pseudo Relevance Feedback (PRF) (Chen *et al*., 2012).

Different approaches used to improve the effectiveness of IR system. One of these approaches was query expansion. There are two approaches for expansion query. The first is a global query expansion which is relies on the external resource such as WordNet. Crimp and Trotman (2018) used WordNet for query expansion and applied on different versions of TREC collections. The other one is local query expansion which based on relevance feedback. Hui *et al*. (2011) proposed different pseudo relevance feedback methods based on Rocchio's model and compare between them to find the effective one which makes improvement on the IR system. Singh and Sharan (2016) used different combining method to merge the result of different selection terms algorithms and Word2Vec approach to extract term from relevance feedback documents and used for expansion query. Pal *et al*. (2013) combined distribution methods with association methods to expand the query. Jun Miao 1, Jimmy Xiang ji Huang 2, Zheng Ye 2 used different ways of proximity to select expansion terms based on relationship between words (Miao *et al*., 2012).

In the proposed method a pseudo relevance feedback or local feedback was used to query expansion. The system returns an initial set of retrieval documents (first pass retrieval) and assumed that the top k documents are relevant and the rest irrelevant documents. From these top retrieved relevant documents different algorithms approached to select terms which used later to expand the original query. Borda aggregation method proposed then to combine the different results produced from many selection terms methods. Later method of reweighting was suggested to reweight the candidate terms that used to expand the query. Returned documents from the second retrieval were more relevant and optimize the performance of the system.

MATERIALS AND METHODS

In this study, we improved \the effectiveness of IR system based on relevance feedback. In the first stage of retrieval, we compared between two similarity measures

Corresponding Author: Hadeel D. Abdulameer, Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

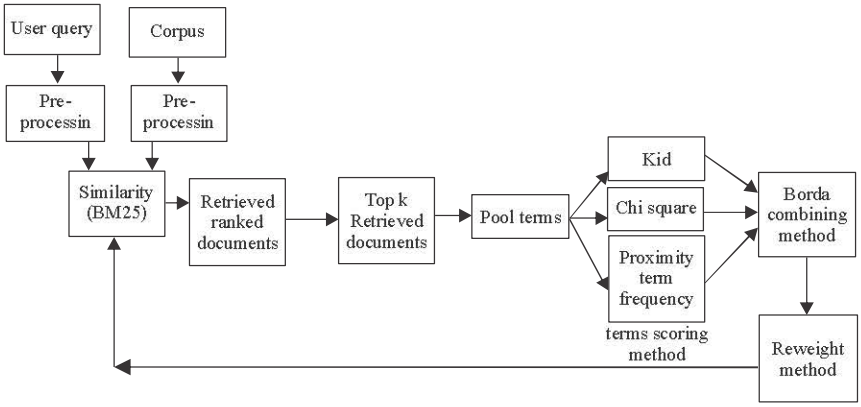

Fig. 1: Our proposed system

which are cosine and BM25 similarity measures. We noticed that the result of BM25 was better. So, we proposed it as similarity measure in our system. Then we used different term selection methods to extract terms from the top retrieved documents, then borda method merges the result from each terms selection method. Since, the higher rank of candidate terms reweighted, then it used to expand the query. Figure 1 shows our proposed model. The following sections will explain each method used in this study with some detail.

Similarity measures

Okapi BM25 similarity measure: It is a probabilistic model, widely and effective ranking function which is used in information retrieval. It is the base model which compute the similarity between query and documents in the collection and returns the rank of the matching documents (Singh, 2017). It improves inverse document frequency (idf) by factoring in term frequency. It can be formalized as:

$$\text{Okapi}(Q, D_i) = \sum_{te Q \cap D_i} \frac{(k_1+1)*tf}{K+tf} * \log \frac{N-n+0.5}{n-0.5} * \frac{(k_3+1)*qtf}{k_3+qtf} \quad (1)$$

$$k = k_1 * (1-b) + b * \left(\frac{dl}{avdl}\right) \quad (2)$$

Where:

N	:	The number of documents in the collection
n	:	The number of documents containing a specific term
qtf	:	Query term frequency
tf	:	Term frequency of term
t	:	The document
dl	:	The length of the document
k_1, b and k_3	:	Constant parameters

The values of parameters that we used in this study are based on the Robertson (k_1 = 1.2, b = 0.75 and k_3= 7.0).

Cosine similarity measure: It is a similarity function using to rank the documents in information retrieval system. It is using vector space model by representing each document in collection and query as a vector of weights which correspond the importance of the word in them as soon as weights are computed, document and query vectors are calculated and compared using cosine similarity (Manning *et al*., 2010). It can be formalized as:

$$\cos(v1, v2) = \frac{v1.v2}{\|v1\| x \|v2\|} \quad (3)$$

where "." represent the dot product of two vectors while ||. || represent the vector length.

Selection terms methods to query expansion in PRF: In this segment, we will explain briefly some methods for query expansion terms selection and we will combine the result of this methods then we will compare between them.

Distribution techniques

Kullback-Leibler Divergence (KLD): It is a rank function using for selecting terms from top rank documents retrieved from the first pass and using for query expansion. It is based on how one probability distribution is different from a second, differences between spread the terms in the top retrieved documents and entire corpus (Carpineto *et al*., 2001). The equation of KLD score is:

$$kLD(t) = P_R(t) \log \frac{P_R(t)}{P_C(t)} \quad (4)$$

Where:

$P_R(t)$	:	Represent the probability of occurrence of term in top retrieved documents R
$P_C(t)$	:	The probability of occurrences of term in the corpus

Chi-square: It is a statistical method measured the difference between occurrence of the term in relevant

documents and occurrence in the corpus then return the score of the top term selected (Carpineto *et al*., 2001). The equation of Chi-square score is:

$$Chi_{square} = \frac{(P(t/P) - (P(t/C))^2}{P(t/C)} \quad (5)$$

Where:
P(t/R) : The probability of happening of the term in relevance documents
P(t/C) : The probability of happening of term t in a corpus

Proximity Term Frequency (PTF) measure: Depending on the feedback set expansion words which selected by computing the distance between the expansions words and query words and also taken in the account the frequency of query words which is calculated by inverse document frequency (idf) to calculate the weight for expansion terms. This way called ptf (Miao *et al*., 2012). Different ways found to compute the proximity, in this study we used Gaussian kernel function which compute the association between query and candidate terms depending on their positions and computing the distance between them. Here, kernel function can be calculated as:

$$k(t, q) = \exp\left(\frac{-(p_t - p_q)^2}{2\sigma^2}\right) \quad (6)$$

Where:
P_t : Represent the position of candidate expansion term
P_q : The position of query term

The scale of Gaussian distribution can be controlled by tuning parameter called σ which also represents W-Sizewhich represent the size of document in our system. In this method we take the importance of different terms of query into account, the equation below is show that:

$$Proximity(t) = \sum_{i=1}^{|Q|} k(t, q_i)\, IDF(q_i) \quad (7)\backslash$$

Where:
|Q| : The number of terms in query
q_i : Query term
t : Candidate expansion term, IDF=log N / df
N : All the terms in collection
df ; The document contain the q_i

Rank aggregation methods: In previous section many selection terms algorithms were discussed and each selection method gave us list of candidate terms to expand query in conclusion multi rank lists of candidate terms introduced as a result rank aggregation methods will be used to combine multi rank list and choice the top rank candidate terms to expansion query. There are many voting algorithms some based on query expansion term positions such as borda algorithm and the other based on query expansion term scores in this study, we will use borda algorithm.

Borda approach: It is rank combining algorithm used to combine different list of candidate term produce from different scoring terms algorithm (voter) to expand query (Kelly, 1988). In each algorithm the higher scoring terms has given m points and the second m-1 and so on. To obtain the final points to each candidate term we will sum the point of the candidate term in each voting list. Since, we rank this list we will obtain borda list of candidate terms.

Reweighting: Reweighting is a process used to reweight the terms in the expanded query with or without considering of the results of the algorithms used to term ranking the standard. Rocchio formula (Rocchio, 1971) and the Robertson/Sparck-Jones formula are traditional methods used to reweight terms depending on the terms in the pseudo-relevance documents.

Max_norm term reweighting: This method was suggested by Carpineto *et al*. (2002). It is a variant of the standard Rocchio formula. This method was used to reweight terms depending on the results of term ranking algorithms. The Rocchio formula can be calculated as:

$$w'q,t = \alpha \cdot wq,t + \beta \cdot max\_norm\_scoret$$

Where:
$w_{q,t}$: Represent the old weight of query term t
$w'_{q,t}$: Represent the new weight of query term t

Both α and β equal to1 and max_norm_scoret can be calculated by dividing the score of term t in the term ranking algorithm by the maximum score of the term in this algorithm.

Test collection: In this study, we used LISA test collection which its sources documents taken from library and information science abstracts database. Provided by Peter Willett of Sheffield University. It contains 5999 documents and 34 queries. The queries are generally long. A document was represented by the title and the abstract.

Evaluation measure: In this study, we used Mean Average Precision (MAP) as a standard measure to evaluate our results and to find the effectiveness of retrieval system (Jangid *et al*., 2014). MAP is computed by calculated the mean value of the average precisions to multiple queries. The Average Precision (AP) is computed by found the mean precision for all the retrieved relevant documents:

$$MAP = \frac{\sum_{i=1}^{Q} AP(i)}{|Q|} \quad (8)$$

Table 1: Comparison between cosine and BM25 similarity measure

Variables	Values
Collection	LISA
Metric	MAP
Cosine	0.318615
BM25	0.352547

Table 2: Comparison between different selection terms methods with BM25 and Borda combining methods

Methods	MAP
BM25	0.352547
KLD	0.364853
PTF	0.368084
Chi square	0.371653
Borda	0.376071

$$AP = \frac{1}{n}\sum_{i=1}^{n} precision(p_i) \quad (9)$$

$$Precision = \frac{|R_r|}{|S_{ret}|} \quad (10)$$

Where:

R_r : The set of the retrieved relevant documents

S_{ret} : All the retrieved documents

RESULTS AND DISCUSSION

For comparison, we implemented two information retrieval baseline similarity methods which are cosine and Okapi BM25 methods we found Okapi BM25 method gives better result than cosine method, so, we depended on this way in our work. The result of this comparison explained in Table 1. Also Table 2 represents the results of different selections methods and combining method. We notice that all selection methods superior on the base similarity method and combining methods gives better results from all of them. This result obtained from top 15 documents retrieved and 20 expansion terms.

CONCLUSION

In our proposed system, we used two distributional methods (KLD, Chi square) and proximity term frequency association method. This methods used to select terms from top retrieved documents which used to expand the query. We found CHI square was better performance from the proximity method. we found that using Borda method which was used to combined the different results produced from many scoring terms methods was an effective role to improve the results produced from each selection methods separately and therefore enhance the performance of the system at all.

REFERENCES

Carpineto, C., G. Romano and V. Giannini, 2002. Improving retrieval feedback with multiple term-ranking function combination. ACM. Trans. Inf. Syst., 20: 259-290.

Carpineto, C., R. De Mori, G. Romano and B. Bigi, 2001. An information-theoretic approach to automatic query expansion. ACM. Trans. Inf. Syst. (TOIS), Vol. 19, No. 1. 10.1145/366836.366860

Chen, C., H. Chunyan and Y. Xiaojie, 2012. Relevance feedback fusion via query expansion. Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT'12) Vol. 3, December 4-7, 2012, IEEE Computer Society, Washington, DC., USA., pp: 122-126.

Crimp, R. and A. Trotman, 2018. Refining query expansion terms using query context. Proceedings of the 23rd Australasian Symposium on Document Computing (ADCS '18), December 11-12, 2018, ACM, New York, USA., pp: 1-4.

Hui, K., B. He, T. Luo and B. Wang, 2011. A comparative study of pseudo relevance feedback for ad-hoc retrieval. Proceedings of the International Conference on the Theory of Information Retrieval (ICTIR'11), September 12-14, 2011, Springer, Berlin, Germany, pp: 318-322.

Jangid, C.S., S.K. Vishwakarma and K.I. Lakhtaria, 2014. Ad-hoc retrieval on FIRE data set with TF-IDF and probabilistic models. Int. J. Comput. Appl., 93: 22-25.

Kelly, J.S., 1988. Arrow's Impossibility Theorem. In: Social Choice Theory, Kelly, J.S. (Ed.). Springer, Berlin, Germany, ISBN: 978-3-662-09927-8, pp: 80-87.

Manning, C., P. Raghavan and H. Schutze, 2010. Introduction to information retrieval. Nat. Lang. Eng., 16: 100-103.

Miao, J., J.X. Huang and Z. Ye, 2012. Proximity-based rocchio's model for pseudo relevance. Proceedings of the 35th ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR'12), August 12-16, 2012, ACM, Portland, Oregon, USA., pp: 535-544.

Pal, D., M. Mitra and K. Datta, 2013. Query expansion using term distribution and term association. Inf. Retrieval, Vol. 1,

Rocchio, J.J., 1971. Relevance Feedback in Information Retrieval. In: The SMART Retrieval System: Experiments in Automatic Document Processing, Salton, G. (Ed.). Prentice-Hall, Upper Saddle River, New Jersey, USA., pp: 313-323.

Singh, J. and A. Sharan, 2016. Relevance feedback-based query expansion model using ranks combining and Word2Vec approach. IETE. J. Res., 62: 591-604.

Singh, J., 2017. Ranks aggregation and semantic genetic approach based hybrid model for query expansion. Int. J. Comput. Intell. Syst., 10: 34-55.