Expansion of Single-word Weak Queries Using Wikipedia as External Data Resource

Kamalika Sanyal
Dept. of Comp. Sc. and Engg.
IIIT Bhubaneswar
Bhubaneswar, India
ksanyal.engg@gmail.com

Nikhil Priyatam
SIE Lab
IIIT Hyderabad
Hyderabad, India
nikhil.priyatam@research.iiit.ac.in

Abstract

Query expansion is an effective technique to improve the performance of weak web search queries. The external data sources like Wikipedia data dump can be taken into confidence to provide reliable related terms for expanding queries. In this work we propose a query expansion approach based on Wikipedia as external knowledge repository and WordNet to detect linguistically important terms. Pseudo Relevance Feedback is used to retrieve top documents from Wikipedia with respect to a query. These top documents serve as a pool of potential expansion terms. Moreover, we focus on travel and lifestyle domain. Hence, intuitively the words absent in WordNet are broadly categorized as Named Entities and certain boosting were given to those terms. We also incorporated some linguistic information extracted from Wikipedia articles within the scoring of the terms. By defining proper term weighting strategies, the query expansion performs effectively. Consequently, we observed that the expansion leads to an overall improvement in the result.

1 Introduction

Query expansion is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. Reformulating mainly means the introduction of new related terms in the seed query. In the context of web search engines, query expansion involves evaluating a users input. A document may not explicitly contain the terms present in the query. Still the document may be relevant with respect to the idea of information need presented by the query. Query reformulation is a common approach in information retrieval to cover the gap between the original information need of the user and the actual query provided.

The methods for tackling this are split into two major classes: global methods and local methods. Local methods adjust a query relative to the documents that initially appear to match the query, e.g. Relevance Feedback (RF), Pseudo Relevance Feedback (PRF). Global methods are techniques for expanding or reformulating query terms independent of the query and results returned from it, so that changes in the query wording will cause the new query to match other semantically similar terms. Global methods include query expansion. Here we propose a method which uses a combination of both local and global method to utilize the strengths of both.

Various approaches have been made in literature to effectively utilize Wikipedia as external data resource. In (Li et al., 2007) retrieval is performed using the MRF model for Term Dependencies following the query formulation. (Xu et al., 2009) does a systematic exploration of the utilization of Wikipedia in pseudo relevance for query dependent expansion. (Pérez-Agüera and Araujo, 2008) tests different approaches to extract the candidate query terms from the top ranked documents returned by the first-pass retrieval. One of them is the co-occurrence approach, based on measures of co-occurrence of the candidate and the original query terms in the retrieved documents. A naive combination of co-occurrence and probabilistic method is developed, with which improved results are obtained. Arguably, the semantic relatedness between terms weakens with the increase in the distance separating them. In (Vectomova and Wang, 2006) a study is conducted to systematically evaluate different distance functions for selecting query expansion terms. A distance factor is proposed in this work that can be effectively combined with the statistical term association measure of Mutual Information for selecting query expansion terms.

If the words of a query cannot provide enough information of the user's need, the retrieval result may be poor (Buckley, 2004). These are called weak queries (Kwok et al., 2005). Regarding Web search of users, almost 48.4% of users submit a single-word query, whereas only 31% users submit queries with three or more words (Spink et al., 2001). Single word queries, with their too much specificity or too much generalness, sometimes do not qualify as an effective web query. To obtain better results, therefore, some automated query expansion is important. We propose a query expansion based approach to bridge the gap between users' information need and the weakness of short queries. This work concentrates on single word queries over travel and lifestyle domain. The main idea is to collect prospective expansion terms from top n pages of Wikipedia which comes out as result of searching the Wikipedia data dump using the original query. Then the collected terms are scored and ranked and highest ranked terms are chosen as expansion terms in order to create a new expanded query.

The remainder of this paper is organized as follows: theoretical background and development of the work are in section 2 and 3. Experimental details and results are discussed in section 4. Finally we conclude with section 5.

2 Theoretical Background

The methods for tackling the problem of query reformulation are broadly split into two major classes: global methods and local methods. Relevance Feedback is a local method which is found to be one of the most powerful methods for improving IR performance. It is an iterative process, best modeled as a continuous loop and is a user-centered approach. Here an interactive user of the system explicitly marks a few top ranked documents as relevant or irrelevant to their information need, then the query is reformulated based on this information. Pseudo relevance feedback is a form of relevance feedback where no user interaction is involved. It is assumed that n top ranked documents are relevant and a new query is learned from these pseudo-relevant documents to improve the performance of the system. Global methods are techniques for expanding or reformulating query terms independent of the query and results returned from it, so that changes in the query wording will cause the new query to match other semantically similar terms.

2.1 Combination of Local and Global Methods

Here we have taken an optimal combination of global and local methods into consideration. According to (Xu et al., 2009), Wikipedia can be seen as a large, manually edited document collection which could be exploited to improve document retrieval effectiveness within pseudo relevance feedback.

Unlike (Li et al., 2007), in which pseudo relevance feedback, WikiText retrieval and ranking are used separately and compared with each other, the methods are fused here. Firstly, the concept of pseudo relevance feedback is used. The indexed Wikipedia data dump is searched using the original query word. It is assumed that the top n documents retrieved as results of the original query are the most relevant ones, and they are taken into consideration. The words contained in those documents only are chosen as prospective expansion terms or candidate terms. So this actually applies pseudo relevance feedback on an external corpus, a global resource, thus combining a local query expansion method to a global one (Müller and Gurevych, 2008).

2.2 Using Wikipedia

As per (Li et al., 2007; Xu et al., 2009; Müller and Gurevych, 2008; Singhal, 2001), it is evident that using external data resources is not an uncommon practice in Information Retrieval. Using these external data repositories, boosting, and query expansion can be done. There are many publicly available databases that contain various kinds of information. Wikipedia is a multilingual, web-based, free-content encyclopedia. Wikipedia also has detailed guidelines governing its content quality. The exponential growth and the reliability of Wikipedia make it a potentially valuable resource for IR (Xu et al., 2009). These characteristics make it an ideal repository as background knowledge (Li et al., 2007).

2.3 Boosting Terms with the help of WordNet

WordNet[1] is a large lexical database of English which provides traditional lexicographic information based on modern computing. In this work WordNet 3.0 is used. This version contains more than 155,287 different word forms and more than 117,659 different word senses. It contains semantic relationships between words in English language.

But in this work, instead of using this relationship, the non-relationship is exploited in the proposed work. Intuitively, it is being considered that in the domain of travel and lifestyle, the words which fall broadly in the category of Named Entities are better expansion terms for query rather than those which are not. For example, "Hyderabad Charminar" will be a better search query than "Hyderabad street" because "Charminar" is a Named Entity and uniquely identifiable as the main attraction of the mentioned city. The expansion term "street" is not able to give such uniqueness in the search.

3 Automated Query Expansion using Wikipedia and WordNet

Here a detailed description of the methodology used is given. PRF is used to fetch top k documents from Wikipedia and these documents are used as the pool of prospective candidate terms. The candidate terms are scored within the external Wikipedia collection using Bo1 term weighting model and vicinity values, which are essential to measure the importance of a term with respect to a complete collection. Figure 1 shows the steps of this approach.

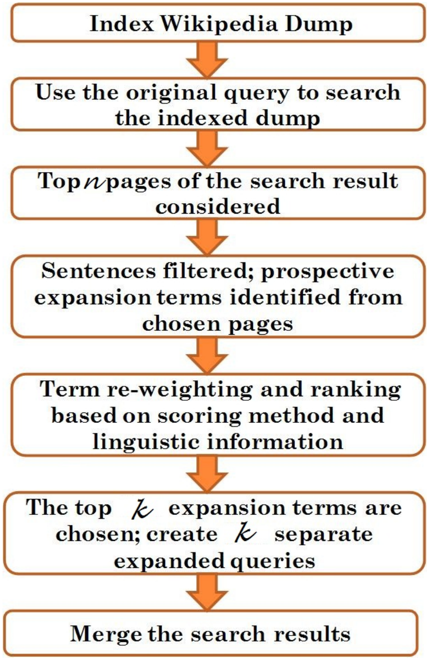
3.1 Indexing Wikipedia dump

The Wikipedia dump is a collection of all recent articles in an annotated form. It is a huge XML file, which contains the articles from Wikipedia. The documents are indexed based on the content of the documents present in that collection.

3.2 Identifying Candidate Terms for Query Expansion

Let q_0 be the original query. The documents from the Wikipedia dump are retrieved using search method in secondary index of the Wikipedia Dump.

Figure 1: Steps of the proposed system.


Index Wikipedia Dump
Use the original query to search the indexed dump
Top n pages of the search result considered
Sentences filtered; prospective expansion terms identified from chosen pages
Term re-weighting and ranking based on scoring method and linguistic information
The top k expansion terms are chosen; create k separate expanded queries
Merge the search results

Next, the top n documents are retrieved, which are the URLs of top n articles of Wikipedia. These articles are considered as the pool of prospective expansion terms. Furthermore, the terms in these documents are filtered based on the following criteria:

1. it should not be a stop word.
2. it should be in the same sentence which contains q_0 in it.
3. it should not be a non-ASCII.
4. it should not be a numeric.

The terms which are selected after fulfilling the criteria as stated above are chosen as the candidate terms.

3.3 Assigning Scores to Candidate Terms

The Candidate Terms are scored and ranked according to the following two scoring factors: Bo1 term weighting model and Vicinity. Then certain terms are selected and re-weighted based on WordNet.

The Bo1 scoring function is based on Bose Einstein statistics. The Divergence From Randomness (DFR) term weighting model infers the informative characteristics of a term by the divergence between its distribution in the top ranked documents and a random distribution. Bo1 model is the most effective DFR term weighting model (Pérez-Agüera and Araujo, 2008). In Bo1, the informativeness $w(t)$ of a term is given by the following equation:

[1] http://wordnet.princeton.edu

$$w(t) = tf_t.log_2 \frac{1 + P_n}{P_n} + log_2(1 + P_n) \quad (1)$$

Here, tf_t is the term frequency of the term in the pseudo-relevant document set, commonly referred to as *within document term frequency*. P_n is given by $\frac{F}{N}$. F is the term frequency of the term t in the whole collection and N is the number of documents in the collection.

In this work this model is used to score the candidate terms with respect to their importance in the Wikipedia dump used to collect these terms.

Vicinity values are the deciding factor for determining importance of a term with respect to the original query word. The idea is based on the *association hypothesis* (Pérez-Agüera and Araujo, 2008). The hypothesis states that if an index term is good at discriminating relevant from non-relevant documents then any closely associated index term is likely to be good at this. This is true when the co-occurrence analysis is done on the whole collection but if we apply it only on the top ranked documents discrimination does occur. The semantic relatedness between terms weakens with the increase in the distance separating them (Vectomova and Wang, 2006).

Hence the proximity of the terms with q_0 is given certain importance. If a term is present in the same sentence and nearby q_0, the distance $d(t; q_0)$ between the term t and q_0 is calculated. The vicinity $V(t; q_0)$ is calculated as:

$$V(t; q_0) = \frac{1}{d(t; q_0)} \quad (2)$$

If the sentence contains the original query term q_0 more than once, the minimum distance and therefore the maximum vicinity of the t and q_0 is considered. For the sake of completeness one can define $V(q_0, q_0) = 0$.

The goal of re-weighting is to pull some words up in ranked list which are important with respect to travel and lifestyle domain. Intuitively it is being considered that the words which are not present in WordNet can be classified broadly as Named Entities and they have higher importance in travel and lifestyle domain. Here, the words not present in WordNet are given a certain boost by multiplying their score with a boosting factor B. $B(t)$ is defined as:

$$B(t) = \begin{cases} b & \text{if } t \text{ is in WordNet} \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

Here b is a positive real number.

Taking all the parameters in consideration, the final score of each candidate term t, denoted by $S(t)$ is calculated. First score is given by Bo1 term weighting model. $scoreBo1(t)$ is obtained by equation 1. Then the vicinity score $V(t)$ is multiplied with $scoreBo1(t)$ to give extra weight to those candidate terms appearing near the original query word. Lastly, the boosting factor $B(t)$ is calculated based on WordNet and multiplied to obtain the final score $S(t)$:

$$S(t) = scoreBo1(t).V(t).B(t) \quad (4)$$

The candidate terms are assigned score and are ranked in decreasing order of scores. Now, the top n terms are selected as expansion terms.

3.4 Creating new expanded queries

The top k expansion terms are used to create new queries. Each expansion term t_i where $i \in (1, k)$, is added with the original query q_0 and a new query is created. Hence the new queries are $(q_0 + t_1), (q_0 + t_2) \ldots (q_0 + t_k)$.

3.5 Merging the results

The expanded queries are used to search the web using Google[2]. The results obtained are merged and ranked as per the rank of the expansion terms. Suppose scores of t_1, t_2, t_3 are in descending order, the results of $q_0 + t_1$ is taken first, result of $q_0 + t_2$ is then appended with that and so on. First r results for each $(q_0 + t_i)$ are considered for merging. For evaluating the merged results presicion at 10^{th} document (`P@10`) and average precision (`AP`) are calculated.

4 Experimental Design and Evaluation

The English language Wikipedia article collection is available as an open source, online. In this work, Wikipedia's top articles are used as per the notion of PRF. The scoring of the prospective terms are also done by taking Wikipedia as the collection. The experiments are performed for indexing Wikipedia repository, ranking of the expansion terms according to scores and after that the effectiveness of this approach of query expansion is experimentally evaluated.

4.1 Indexing

The Wikipedia external data is collected from WikiDump[3]. The dump contains approximately

[2] http://google.co.in

[3] http://dumps.wikimedia.org/enwiki/

q_0	t_1	t_2	t_3	t_4	t_5
Train	ICE3	Alternatively	Look	Highspeed	M18
Weather	Largescale	Geoengineering	Today	Oceanographic	Dry
Rhinoceros	Victoriaceros	Narrownosed	Twohorned	Sondaicus	Mercks
Bhubaneswar	Boomtown	DPR	Rasgulla	Mandap	Stinks
Shimla	Niwas	Ninehole	Juga	Busstand	Almora
Valley	Subsidence	Hydrologically	LeHigh	Calanca	Scotty

Table 1: Some original query terms and their top 5 expansion terms

4 million documents or articles, in XML format. It occupies almost 6 GB of space when uncompressed in the hard disk. Apache Solr 3.6.2[4] is used to index Wikipedia. Wikipedia is dumped and indexed separately.

4.2 Search and URL List creation and Candidate Term extraction

After indexing, the search is done on the index file using the original query. The top n articles are chosen as the top documents to contain the candidate terms. Here n is chosen as 5 for the experiments.

4.3 Scoring the candidate terms

The candidate terms are scored and ranked in decreasing order of their scores. We have taken top k scoring terms as expansion terms. Here k is taken as 5. The boosting factor $B(t)$ is equal to the default 1.0 if term t is present in WordNet and is equal to 2.0 if t is absent from WordNet. Table 1 shows some original query terms and the top ranked candidate terms.

4.4 Evaluation of results

For experimental purpose, 20 single-word web search queries are considered. The results obtained for expanded queries are merged as explained in 3.5. In this work number of pages considered per query r is taken as 5. For evaluation manual relevance judgement of the results origianl query and of the merged results for expanded queries are done by a set of beta users. Final `P@10` and `AP` reported here is the average of those values calculated based on the relevance judgements of each beta user. Table 2 show the statistics for `P@10` and `AP` of original query terms and the expanded query terms.

It is evident from the results that our approach increased the performance over the single-word queries. Among the 20 web search queries the word “safari” had shown most improvement when expanded. This observation established the fact that sometimes some seemingly unrelated words can have certain relationship with other words, such as “struik” with “safari” and “merck” with “rhinoceros”, which are obviously the result of consulting one of the best knowledge repositories around, Wikipedia.

On the other hand it has been observed that in some cases, for example for the widely known city names like “Kolkata”, “Mumbai” expanded versions performed comparatively poorly. For these important cities, often the non-travel and lifestyle oriented pages, like political or law enforcement or job-oriented pages are coming up as top results because of their importance with respect to those cities. For example, the top expansion terms like “NCP” are not relevant for original query “Mumbai” with respect to travel and lifestyle domain, so is “Telegraph” for original query “Kolkata”.

5 Conclusion and Future Work

The above-mentioned observations shows that for most of the query words, after expanding using top five expansion terms, the average precision has increased significantly. The terms inserted in the query are mostly good terms in the Bo1 and Vicinity scoring sense. The concept of Named Entities, in a broad sense though, have also performed fair with respect to travel and lifestyle. So as a summary it can be stated that the query expansion approach based on both local and global methods presented here succeeds to outperform the original single-word weak queries.

The documents can be considered as multi-dimensional vectors in order to find the evaluation of diversity of the result. The unique words in the top n retrieved documents for the original query and those for expanded queries can be considered as elements of the vector. The centroid of the vectors of original query as well as for expanded queries can be calculated. If the average distance

[4] `http://lucene.apache.org/core/`

Original query term (q_0)	AP of results of q_0	AP of merged results of expanded queries	P@10 of results of q_0	P@10 of merged results of expanded queries
Rhinoceros	0.8119	1.0000	0.9000	1.0000
Bhubaneswar	0.7651	0.7807	0.8000	0.9000
Backpacking	0.9922	1.0000	1.0000	1.0000
Train	0.8644	0.8949	0.8000	0.8000
Brahmaputra	0.7152	0.3542	0.6000	0.4000
Flight	0.8286	0.9126	0.8000	0.9000
Kolkata	0.7356	0.6703	0.6000	0.4000
Restaurant	0.9392	0.8413	1.0000	1.0000
Delhi	0.7307	0.7985	0.6000	0.6000
Hyderabad	0.6734	0.8191	0.6000	0.6000
Valley	0.8015	0.7485	0.7000	0.7000
Cycling	0.8607	0.8626	0.9000	0.9000
Haridwar	0.8569	0.9119	0.8000	0.7000
Safari	**0.2057**	**0.7782**	**0.1000**	**0.8000**
Kanyakumari	0.9528	0.8305	1.0000	0.9000
Shimla	0.9040	1.0000	0.9000	1.0000
Mumbai	0.7251	0.5775	0.6000	0.6000
Weather	0.9787	1.0000	1.0000	1.0000
Seaside	0.5786	0.8315	0.5000	0.9000
Chicago	0.5601	0.5931	0.5000	0.5000

Table 2: AP and P@10 measures of original and expanded query terms

of the centroid and each document in the expanded queries' result is greater than that of the original query, it can be established that the diversities of the results of the expanded queries are more than the results of original query.

Tuning of different parameters and using different standard search engines may provide further improvements. Moreover, using WordNet to identify Named Entities is a very broad approach. Alternate NER models like Stanford NER[5] can be used to refine the Named Entity identification.

References

C. Buckley. 2004. *Why current IR engines fail* Proceedings of the 27th ACM SIGIR, Pages 584–585.

K.L. Kwok, L. Grunfeld and P. Deng. 2005. *Improving weak ad-hoc retrieval by web assistance and data fusion* Proceedings of Asia Information Retrieval Symposium Pages 17–30.

A. Spink, D. Wolfram, Major B. J. Jansen and T. Saracevic. 2001. *Searching the web: The public and their queries* Journal of the American Society for Information Science and Technology, 52(3):226–234.

Y. Li, R.W.P. Luk, E.K.S. Ho and F.L. Chung. 2007. *Improving Weak Ad-Hoc Queries using Wikipedia as External Corpus* Proceedings of the 30th ACM SIGIR, Pages 797–798.

Y. Xu, G.J.F. Jones and B. Wang. 2009. *Query dependent pseudo-relevance feedback based on Wikipedia* Proceedings of the 32nd ACM SIGIR, Pages 19–23.

J.R. Pérez-Agüera and L. Araujo. 2008. *Comparing and Combining Methods for Automatic Query Expansion* Proceedings in Natural Language Processing and Applications, Research in Computing Science 33, Pages 177–188.

O. Vectomova and Y. Wang. 2006. *A Study of the Effect of Term Proximity on Query Expansion* Journal of Information Science, 32(4):324–333.

C. Müller and I. Gurevych. 2008. *Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval* Proceedings of the 9th Crosslanguage evaluation forum conference on evaluating systems for multilingual and multimodal information access, Pages 219–226.

A. Singhal. 2001. *Modern Information Retrieval: A Brief Overview* IEEE Data Eng. Bull., Vol. 24, Pages 35–43.

[5] http://nlp.stanford.edu/software/CRF-NER.shtml