Automatic query expansion and word sense disambiguation with long and short queries using WordNet under vector model

Francisco João Pinto*

Department of Computer Science, University of A Coruña,
Campus de Elviña s/n, A Coruña, 15071, Spain
`fjoao@udc.es`
* Corresponding author

Carme Fernández Pérez-Sanjulián

Department of Galician-Portuguese, French, and Linguistics,University of A Coruña,
Campus da Zapateira s/n, A Coruña, 15071, Spain
`carme@udc.es`

Abstract. This paper describes the experimentation conducted to test the effectiveness of automatic query expansion and word sense disambiguation (WSD) using short and long query of a topic TREC under vector model. We ran different experiments generating queries under vector model using linguistic information extracted from WordNet. Results show that query expansion with short queries and long queries is not able to improve retrieval performance without the selection of the correct meaning of the words but the results are better using short queries.

Keywords: Information retrieval, Query expansion, Word sense disambiguation, WordNet, Vector space model.

1 Introduction

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of web search engines, query expansion involves evaluating a user's input (what words were typed into the search query area, and sometimes other types of data) and expanding the search query to match additional documents. Query expansion involves techniques such as: i) Finding synonyms of words, and searching for the synonyms as well; ii) Finding all the various morphological forms of words by stemming each word in the search query; iii) Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results; iv) Re-weighting the terms in the original query. Query expansion is a technology studied in the field of computer science, particularly within the realm of natural language processing and information retrieval.

Search engines invoke query expansion to increase the quality of user search results. It is assumed that users do not always formulate search queries using the best terms. Best in this case may be because the database does not contain the user entered terms. By stemming a user-entered term, more documents are matched, as the alternate word forms for a user entered term are matched as well, increasing the total recall. This comes at the expense of reducing the precision. By expanding a search query to search for the synonyms of a user entered term, the recall is also increased at the expense of precision. This is due to the nature of the equation of how precision is calculated, in that a larger recall implicitly causes a decrease in precision, given that factors of recall are part of the denominator. It is also inferred that a larger recall negatively impacts overall search result quality, given that many users do not want more results to comb through, regardless of the precision.

The goal of query expansion in this regard is by increasing recall, precision can potentially increase (rather than decrease as mathematically equated), by including in the result set pages which are more relevant (of higher quality), or at least equally relevant. Pages which would not be included in the result set, which have the potential to be more relevant to the user's desired query, are included, and without query expansion would not have, regardless of relevance. At the same time, many of the current commercial search engines use word frequency (tf-idf) to assist in ranking. By ranking the occurrences of both the user entered words and synonyms and alternate morphological forms, documents with a higher density (high frequency and close proximity) tend to migrate higher up in the search results, leading to a higher quality of the search results near the top of the results, despite the larger recall.

This trade-off is one of the defining problems in query expansion, regarding whether it is worthwhile to perform given the questionable effects on precision and recall. Critics state one of the problems is that the dictionaries and thesauri, and the stemming algorithm, are driven by human bias and while this is implicitly handled by the query expansion algorithm, this explicitly affects the results in a non-automated manner (similar to how statisticians can 'lie' with statistics). Other critics point out potential for corporate influence on the dictionaries, promoting advertising of online web pages in the case of web search engines.

The expansion experiments were tested against a subset of **the TREC collection (Text Retrieval Conference (http://www.trec.nist.gov)).** Each initial query in every experiment was generated automatically from the TREC topics. Query words are then selected and expanded using the lexical relations included in WordNet. The selection of the correct meaning of each word was done automatically, because our main interest is to evaluate query expansion and an automatic disambiguation of senses. Thus the effect of expansion can be analyzed with regard to the quality of the disambiguation, which is assumed to be optimum. In addition, we select the words only from the title of the topic. This simulates a common user query with few and generic words. The pioneer work in query expansion using WordNet was made by **Voorhees (1994)**. The results of this work were not good, especially when initial queries are long. In the case of initial short queries, query effectiveness was improved but it was not better than the effectiveness achieved with complete long queries without expansion. In the work of **Mandala et al.(1999)** the relations stored in WordNet are combined with similarity measures based on syntactic dependencies and co-occurrence information. This combination improves the effectiveness of retrieval. The work of **Qiu and Frei (1993)** used an automatically constructed thesaurus and the results were good but the expansion was tested against small document collections. Other successful works used thesaurus adapted with relevance information or were tested against collections in specific domains. We will use only linguistic information and we will test expansion with a long subset of the TREC collection. In these works the retrieval model used is the Vector Space Model (VSM). In this model the queries are represented as vectors. The expansion consists in the simply addition of terms to the vector and found a problem in this direct addition of terms in VSM. Let us consider a query (a, b). If there are three expansion terms a_1, a_2, a_3 related with a and one expansion term b_1 related with b, the expanded query (a, b, a_1, a_2, a_3 ,b_1). In the next section we will examine the vector space model explaining the representation of documents and queries and the model for matching. Next we briefly recall what WordNet is. In section 4 we will present the application developed for formatting queries. Section 5 describes the experiments and their results. The conclusions are presented in the last section.

2 Vector space model

The vector space model assigns non-binary weights to terms in both documents and queries and it provides a frameworks in which partial matching is possible. Documents and queries are

represented as t-dimensional vectors as follows. Documents are represented as vectors $\vec{d}$ = ($W_{1,j}$, $W_{2,j}$, · · ·, $W_{t,j}$) and queries are represented as vectors $\vec{q}$ = ($W_{1,q}$, $W_{2,q}$, · · ·, $W_{t,q}$), where each weight $W_{i,j}$ and $W_{i,q}$ is positive and non-binary. The term weights are used to compute the degree of similarity between each document stored in the document base and the query supplied by the user. Documents are stored in decreasing order of similarity and, thus, the vector space model takes into consideration documents which match the query terms only partially. This implies that the ranked answer set is a lot more precise than answer sets supplied by the boolean model. Since documents and queries are vectors in a dimensional space, measure of correlation between vectors can be applied to get a measure of similarity between documents and queries. One of the most widely used measures computes through the cosine of the angle between the two vectors. Formally,

$$sim(d_{j},q) = \frac{\vec{d}_{j} \bullet \vec{q}}{|\vec{d}_{j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} W_{i_{j}} \times W_{i_{q}}}{\sqrt{\left(\sum_{i=1}^{t} W_{i_{j}}^{2}\right)} \times \sqrt{\left(\sum_{i=1}^{t} W_{i_{q}}^{2}\right)}} = \cos\alpha$$

Where $\bullet$ stands for the inner product between two vectors. The value of sim (dj, q) varies from 0 to 1 and a document might be retrieval even if it matches the query only partially. Many other matching functions have been designed to compute the degree of similarity between a query vector and a document vector. For a detailed study of them we refer to (**Baeza-Yates, 99**).

3 WordNet

WordNet (**Miller et al., 1990**). is a lexical system manually constructed by a group of people leaded by George Miller at the Cognitive Science Laboratory at Princeton University. WordNet is organized in sets of synonyms (*synsets*) with the words with the same meaning. These synsets have different relations between them. The relation of hypernymy/hyponymy (is-a relation) is the principal relation and creates a hierarchic structure. There are also relations of meronymy/holonymy (part-of relation). In addition, WordNet is divided in four taxonomies by the type of word: nouns, verbs, adjectives and adverbs. We only used the taxonomy of nouns because nouns are the most content-bearing words. Expansion terms will be selected from the correct synsets for each noun in the query.

4 Query formulation

Queries are generated starting from the TREC topics, selecting the expansion terms and fetching lexical information from WordNet. We developed an application that let us to test different options of expansion. In the definition of an experiment we can choose among different options as if it can see the continuation. We consider one topic TREC with following fields T, D and N where T is the title, D is the description and N is the narrative, that if can represent for T, D, N, only considering title T if it gets the following queries: i) Original query: given a query q_1 = (A_0, B_0, C_0), the vector of the query not expanded only with the title, where A_0, B_0, and C_0 are pertaining terms to the related vector, in which A0 and C_0 are nouns. ii)Expanded query: Also let us assume that from the expansion choosing only the nouns that appear in the literal description of the topic using the linguistic resource WordNet we get the following terms: A_1, A_2 and A_3 as we synonyms related with A_0 and C_1, C_2, C_3 and C_4 as we synonyms related with C_0, then the new vector of the

expanded consultation will be in the following way: $q_2 = (A_0, B_0, C_0, A_1, A_2, A_3, C_1, C_2, C_3, C_4)$.
iii) Expanded query and election of the correct meanings: Let us assume now that in the query: q_2, if selected the terms A_1, C_1 y C_2 as correct, the query will be in the following way: $q_3 = (A_0, B_0, C_0, A_1, C_1, C_2)$ with this if intends that if in one it query an original term has diverse meanings, assumes that those that are different of the correct one do not influence in the effectiveness of the information retrieval. This process is analogous with any syntactic category and with any field of topic.

The processing of the queries after expansion and election of the correct meanings, are made using the vector space model. In this model, the documents of the collection and the queries are represented as vectors of terms inside of a vector space. A gotten time expanded query and the query with the election of the meaning correct of the words, the selected field is indexed using the routines standards of the information retrieval System Lemur , computes one ranking of documents for each query. For each query the documents a are stored in decreasing order of similarity. The value of similarity of a document d with the query q is given using the inner product between the two vectors, and later one time gotten the results of the query expanded and of the query expanded and the election of the correct meanings, on the different contemplated experiments will be proceeded to the evaluation and the comparison between them with the ones of the query without expanding.

5   Experiments and results

We used an information retrieval system called lemur **(http://www.lemurproject.org)** and a subset of documents from the reference TREC-8 text collection for evaluating the effects of query expansion in our experiments. More precisely, in all our experiments those topics numbered between 401 and 450 were taken, so that a collection composed of fifty test queries was used in each experiment. The Small Web (WT2g) was the collection of documents used from TREC-8. WT2g contains around 250,000 documents. In addition, the measurements of recall and precision were used to evaluate the results obtained in the retrieval process. The main goal of our experiments was to prove if the expansion of the original queries using all fields of topics and the selection of the correct meaning lead to some improvements in the effectiveness obtained during the information retrieval process.

5.1 Original query

In this section we focus in the effects of using the original query in an information retrieval system. In the Table 1 , we present the values of precision obtained by the original query using short and long query. In these experiments, the query expansion and word sense disambiguation (WSD) was not applied. As expected, the best results were obtained when the long query were performed. To sum up, the average precision (non-interpolated) value can be obtained. In the practice the average precision (non-interpolated) value using the title is 0, 0839, average precision (non-interpolated) value using the description is 0, 1495, average precision (non-interpolated) value using narrative is 0, 2103, and the average precision (non-interpolated) value using the full query is 0, 2431.

Short and long queries	Title	Description	Narrative	Full
Average precision(non-interpolated) value	0,0839	0, 1495	0, 2103	0,2431
Set total number of retrieved docs	1112	1404	1578	1677

Tabla 1: Results of original query

5.2 Query expansion without disambiguation

In this section we focus in the effects of using query expansion in an information retrieval system. In the Table 2, we present the values of precision obtained by the query expansion and those obtained by applying query expansion of nouns. In these experiments, word sense disambiguation (WSD) was not applied. As expected, the best results were obtained when the short query was performed, since it leads to higher values of precision. To sum up, the average precision (non-interpolated) value can be obtained. In the practice the average precision (non-interpolated) value using the title is 0, 1655, average precision (non-interpolated) value using the description is 0, 0964, average precision (non-interpolated) value using narrative is 0, 0628, and the average precision (non-interpolated) value using the full query is 0, 0445.

Short and long queries	Title	Description	Narrative	Full
Average precision(non-interpolated) value	0,1655	0, 0964	0, 0628	0,0445
Set total number of retrieved docs	1696	1507	1455	1688

Tabla 2: Results of query expansion

5.3 Query expansion with desambiguation

Our dictionary-based disambiguation method makes use of WordNet. Basically, it is based on **(Pinto, 2007a; 2007b)** and contains the following steps: *i)* Extraction from WordNet of the sets of synonyms for all the words being disambiguated, *ii)* determining the coincidence between the context of the words being disambiguated and the sets of synonyms, and *iii)* selection of the correct meaning for the words in a text, where the context is given in an automatic form. In the Table 3, we present the results obtained when query expansion is combined with WSD. By comparing these results against those presented in the Table 2, it can be seen that the use of word sense disambiguation leads to better effectiveness values for all the syntactic categories over which query expansion was performed. More precisely, by including the disambiguation of nouns into the query expansion process: i)The average precision (non-interpolated) obtained using the title is around 0,2030, ii) The average precision (non-interpolated) obtained using the description is around 0,1672, iii) The average precision (non-interpolated) obtained using the narrative is

around 0,1577 and The average precision (non-interpolated) obtained using the full query is around 0,1345.

Short and long queries	Title	Description	Narrative	Full
Average precision(non-interpolated) value	0, 2030	0, 1672	0, 1577	0,1345
Set total number of retrieved docs	1495	1371	1344	1298

Tabla 2: Results of query expansion and WSD

Conclusions

Applying to the retrieval to the original queries the results they are better using the long queries because they have more terms. In this case is important to improve the results of the short queries adding similar terms (query expansion), but some times the similar terms that are added can not be wished, for that reason it is important to select the meaning correct (WSD). Therefore query expansion is important in short queries because when adding more terms similar to long queries only bring about noise. The word sense disambiguation is a non trivial task within an information retrieval system. Many systems try to select the most appropriate sense for a polysemous word using statistics and/or automatic learning. Despite such systems, our proposal uses dictionary-based disambiguation. It uses disambiguation during the query expansion process, yielding interesting improvements in the effectiveness obtained. In our experiments (using vector space model), the new approach yielded improved effectiveness when disambiguation was applied using WordNet. In practice, the best results where obtained when disambiguation was applied using short query. The results obtained regarding both the expanded queries and the word sense disambiguation permit us to extract some interesting conclusions: By using different expansion techniques and WSD the results had slightly improved with respect to the short queries. Experiments show that by using word sense disambiguation more advantages from the query expansion with WordNet can be exploited. In practice, our results in a long text collection support the results obtained in previous research works that reported good effectiveness values when query expansion was used in conjunction with word sense disambiguation over small text collections.

Reference

Baeza-Yates, R.A. and Ribeiro-Neto, B. (1999) 'Modern Information Retrieval'. Addison Wesley.

Mandala, R., Tokunaga, T., and Tanaka, H. (1999). 'Combining multiple evidence from different types of thesaurus', *Proceedings of the 22th ACM-SIGIR Conference*, pp.191-197.

Miller, A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J (1990). 'Introduction to WordNet: An on- line lexical database'. *International Journal of Lexicography*. Vol.3, No.4, pp.235–312.

Pinto, F. J. (2007a). 'Uso del Recurso Linguístico WordNet en la Expansión de Consultas con un Modelo de Usuario de Recuperación de Información'. *Proceedings of the 2nd CEDI*. In Spanish.

Pinto, F. J. (2007b). 'Evaluación del Sistema de Recuperación de Información Lemur con Distintos Tipos de Indexación Automática'. *Proceedings of the 12th Conferencia de la AEPIA (Zoco'07/CAEPIA)*. In Spanish.

Qiu, Y. and Frei, H. (1993). 'Concept-based query expansion', *Proceedings of the 16th ACM-SIGIR Conference*, pp. 160–169.

Voorhees, E. M. (1994). 'Query Expansion using Lexical-Semantic Relations', *Proceedings of the 17th ACM-SIGIR Conference*, pp. 61–69.