DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.14

Word-Context Matrix based Query Expansion in Information Retrieval for Turkish Text

Emre Satir
Department of Computer Engineering
Dokuz Eylul University, Izmir, TURKEY
emre.satir@st.cs.deu.edu.tr

Adil Alpkocak
Department of Computer Engineering
Dokuz Eylul University, Izmir, TURKEY
alpkocak@cs.deu.edu.tr

Deniz Kilinc
Faculty of Technology
Celal Bayar University, Manisa, TURKEY
deniz.kilinc@cbu.edu.tr

In this paper, we proposed a Query Expansion (QE) approach on a Turkish Text collection based on word-context matrix with a sliding fixed sized window and Singular Value Decomposition (SVD) method. Our query expansion approach uses semantic relationship of terms to improve the existing query expansion methods available in the literature, namely Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler). We evaluated our approach on Milliyet collection, which is a Turkish IR test bed containing more than 400K documents and 72 queries. The experimentation shows that our approach clearly improves the all three QE methods in terms of major Information Retrieval (IR) performance measures such as MAP, R-precision and P@10.

Keywords: Word-context matrix, SVD, Query expansion, Turkish information retrieval

1. INTRODUCTION

As a result of rapid development of Internet and related technologies, the amount of data is growing day by day. This data has to be organized and retrieved whenever it is needed. Information Retrieval (IR) is a discipline that is related with indexing, retrieving, and structuring documents from any collections. The retrieval methods aim to find the best matching documents according to user query (user need) within a large document collection. In general users may not know how to construct the best query according to their needs and the queries may be inadequate. Query Expansion (QE) is the process of reformulating the basic user query in order to get a better retrieving performance. Different techniques can be conducted to expand a query such as using synonyms of terms, using ontologies to add related terms, or checking spelling errors of terms and correcting them.

In this paper, we proposed a word-context matrix (Turney & Pantel, 2010) based QE on Turkish Text. Although we have applied our technique to Turkish Text, it is applicable to all languages (i.e. it is language independent). We tried to find expanded terms not outside of the collection (like using an ontology) but from the collection itself. First we constructed a huge term co-occurrence matrix from the text collection and applied Singular Value Decomposition (SVD) method in order to obtain a reduced word-context matrix. Then we utilized this matrix to expand queries.

When applied to document similarity, SVD is called Latent Semantic Indexing (LSI), and when applied to word similarity, it's called Latent Semantic Analysis (LSA). (Turney & Pantel, 2010). Researchers, in their study (Landauer & Dumais, 1997) applied SVD to word similarity with using a term-document matrix. But our computations are based on a word co-occurrence matrix with a fixed-sized sliding window. We used these similarity computations to expand queries.

The rest of this paper is organized as follows. In section 2, our proposed method is introduced. Section 3 describes the experimental study, section 4 discusses the experimental results obtained and finally section 5 tells about conclusions and future work.

2. PROPOSED METHOD

The distributional hypothesis in linguistics is that words that occur in similar contexts tend to have similar meanings (Harris, 1954). A word context can be represented with context matrix. In general, in a word–context matrix, the context is given by different context such as blindly separated set of word window, or more morphologically by sentences, paragraphs, chapters, and documents. This context can be an extension for Vector Space Model (VSM) to measuring word similarity. A word may be represented by a vector in which the elements are derived from the occurrences of the word in various contexts. Then, similar row vectors in the word context matrix indicate similar word meanings.

In this study, we firstly constructed a word-context matrix by using S-Space Package that is an open source framework for developing and evaluating word space algorithms. We selected a fixed window size and use a sliding window style in order to count word co-occurrences. In order to compute word-similarities, we performed comparisons between their co-occurrence vectors. There are different types of similarity measures in the literature, herein; we used well-known cosine similarity measure.

If dimensionality of the matrix is reduced before computing semantic similarities, results can be improved (Landauer & Dumais, 1997). A good mathematical way of realizing this is using Singular Value Decomposition (SVD) (Landauer & Dumais, 1997). Actually SVD is a dimension reduction technique that allows to significantly reduce the number of columns so smaller matrix has the advantage that all subsequent similarity computations are much faster.

After creating our word-context matrix, we utilized SVD operation on this matrix for dimensionality reduction. For the first experiment, we used only this matrix to expand queries, and for the second experiment we used a combined method with the QE mechanisms implemented in Terrier IR platform.

For the second experiment (hybrid system), the procedure that we follow is:

- First we run Terrier with the one of the weighting models (Bo1, Bo2 or KL) and got the expanded terms.
- Secondly for every word in the basic query, we run our word co-occurrence based system with the cosine threshold 0.9 (actually we use distance not similarity, i.e. one minus the cosine of the included angle between points). We choose 0.9 because we want to achieve most of the words that close to the seed word. We discard only words that are very far away from the seed word.
- Then look for every word that Terrier found for the expanded terms in our word pool. If a word is included in both results, we added this word to the query (i.e. we filter out some words from QE mechanism implemented in Terrier).
- Finally we re-run Terrier with the original query and filter-outed expanded terms. We use the same weights taken from methods implemented in Terrier for the terms.

3. EXPERIMENTAL STUDY

In this study, our aim is to generate a new query expansion method based on word-context matrix and SVD, and to evaluate it experimentally. We used The Terrier IR Platform that is developed at the School of Computing Science, University of Glasgow. It is open source and written in Java. It is efficient and effective search engine which implements indexing and retrieving operations for large-scale collection of documents.

3.1 Data Set

In the study, we used Bilkent Milliyet Collection (Can, et al., 2008) that contains 408,305 documents (news articles and columns of five years, 2001 to 2005) from Turkish newspaper Milliyet. Each document includes approximately 234 words on the average. The collection contains about 95.5 million words. There are also 72 ad-hoc queries that are evaluated by 33 assessors. The query file includes Topic, Description and Narrative fields, but we only use Topic field in this study named as short query at (Can, et al., 2008). Table 1 presents first 5 short queries and their English translations.

Table 1: *First 5 short queries*

Query No	Topic (in Turkish)	Topic (English translation)
1	*Kuş Gribi*	*Bird Flu*
2	*Kıbrıs Sorunu*	*Cyprus Issue*
3	*Üniversiteye giriş sınavı*	*The university entrance exam*
4	*Tsunami*	*Tsunami*
5	*Mavi Akım Doğalgaz Projesi*	*Blue Stream Natural Gas Project*

3.2 Pre-processing and Indexing

Pre-processing is an important step before indexing. Tokenization, stop-word elimination and stemming are the most widely used pre-processing methods. In this study, we pre-processed the related tag contents. First, we converted all uppercase letters to lowercase equivalents and then we converted Turkish special characters to their Latin alphabet counterparts (ç→c, ğ→g, ı→i, ö → o, ş→s , ü→u). Also the same two operations are applied to the all fields of the query file. For stemming, we used fixed prefix stemming (Can, et al., 2008) and we selected the first 5 characters of a term as its stem. We employed a semi-automatically generated stop-word list contains 147 words that taken from (Can, et al., 2008). After pre-processing step, we indexed headline and text fields of the documents by using Terrier. We performed TF×IDF weighting model (Manning, et al., 2009) for our experiments.

3.3 Preparing the word co-occurrence matrix

As mentioned before we used S-Space package to construct the matrix. Before creating the matrix, we made some operations on Milliyet collection. In brief, stop-words removal operation is performed headline and text fields contents, the first 5 characters of a term is selected as its stem and Turkish special characters are converted to their Latin alphabet counterparts.

After these operations, we processed the converted collection with the S-Space package system. Total vocabulary (unique term size) is 139,916. For window size we used 5. SVD is a computationally exhausting operation so we decided not to use all terms in the vocabulary. We set the minimum frequency to 20, so we restricted our vocabulary to all terms occurring at least 20 times in the collection. So our new vocabulary size is 34,839 which led to a matrix of size 34,839x34,839. Finally we made SVD operation on this matrix with the reduction parameter of 400. So this gave us final matrix of size 34,839x400.

4. EXPERIMENTAL RESULTS

We utilized trec_eval IR evaluation tool for evaluation purposes. First we performed our baseline evaluations without any query expansion. Baseline results can be seen from the Table 2.

4.1 QE Methods Implemented in Terrier Results

Pseudo relevance feedback (Blind relevance feedback) uses a method like automatic local analysis. (Manning, et al., 2009). The user doesn't give feedback to the system, because the system automatically reformulates the query. First the system does the normal retrieval according to initial query and after finding documents the system assumes that the top *k* ranked documents are relevant, and finally to do relevance feedback under this assumption. Generally the procedure is:

- Take *k* documents from the initial retrieving as relevant results.
- Select top *t* terms from these documents using for example TF×IDF weights.
- Add these terms to query, do the retrieving process again and finally return new results.

Terrier has a built-in query expansion functionality that utilizes pseudo relevance feedback. In this study, we experimented with three query expansion models of Terrier: Bo1 (Bose-Einstein 1), Bo2 (Bose-Einstein 2) and KL (Kullback-Leibler). (Amati, 2003) Each model used for expanding the query with the most informative terms of the top-ranked documents. We experimented with only default values, which are 10 for the number of terms to expand a query with, and 3 for the number of top-ranked documents from which these terms are extracted.

Terrier utilizes a particular Divergence From Randomness (DFR) term weighting model to weight terms in the top-returned documents Bo1 model uses the Bose-Einstein statistics (Amati, 2003):

$$w(t) = tf_x \, log_2 \left(\frac{1+P_n}{P_n}\right) + log(1+P_n)$$

where tf_x is the frequency of the query term in the top-ranked documents. P_n is given by *F/N* where *F* is the frequency of the query term in the collection and *N* is the number of documents in the collection.

The second useful approximation of the Bose-Einstein statistics is generated by the Stirling formula (Amati, 2003).

Kullback-Liebler divergence computes the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained for a first pass retrieval using the original user query. (Cover & Thomas, 1991) For the term t this divergence is :

$$KLD_{(PR,PC)}(t) = \; P_R(t) \, log \frac{P_R(t)}{P_C(t)}$$

where $P_R(t)$ is the probability of the term *t* in the top ranked documents, and $P_C(t)$ is the probability of the term *t* in the whole collection.

QE methods implemented in Terrier results can be seen from the Table 2.

4.2 Proposed QE Results

4.2.1 First Experiment

In the first experiment, we searched each query term in our word-context matrix and calculate cosine similarity ratio with all other words. If this value is less than 0.1 (because we use one minus the cosine in our system) we added this word to the query. We applied the same weights with main query words, and received MAP value of 0.3093.

4.2.2 Second Experiment

As shown in the first experiment, the MAP result is below our baseline. This may be caused by reason that we didn't use any weighting method for the query terms or some expansion terms caused query drifting. In the second experiment, we attempted to develop a hybrid system. We first performed Terrier Query Expansion and got the expanded terms. Then, we used our word-context matrix and we compared these two systems words. If a word is included in both result, we added this word to the query. Also we used the weight taken from QE methods implemented in Terrier for the terms this time. Table 2 summarizes the results that we obtained from experimentation. The second column in each measure, shown in boldface font, indicates the results of our approach for each method. These results show that our approach clearly improves the performance of all three QE methods we tested.

Table 2: *Results of experiments for different methods.*

Method	**MAP**		**R-prec**		**P@10**	
Baseline	0.343		0.361		0.567	
Bo1	0.365	**0.377**	0.380	**0.393**	0.576	**0.589**
Bo2	0.361	**0.371**	0.374	**0.386**	0.565	**0.589**
KL	0.366	**0.377**	0.381	**0.393**	0.574	**0.586**

5. CONLCUSIONS AND FUTURE WORK

In this study, we proposed a Query Expansion (QE) approach on a Turkish Text collection based on word-context matrix with a sliding fixed sized window and Singular Value Decomposition (SVD) method. In order to evaluate the proposed approach, we performed a baseline experiment. Then, we conducted QE methods implemented in Terrier and obtained Mean Average Precision (MAP) results. Finally, we evaluated our proposed Query Expansion approaches.

In the first experiment we expanded the queries with only our system and we couldn't get a good result. But in the second experiment we used a hybrid method (our algorithm plus QE methods implemented in Terrier) and this time we could outperform the baseline and QE methods implemented in Terrier.

We can conclude that co-occurrence information succeeds on synonym and automatically synonym can help us in Query Expansion. Maybe our co-occurrence matrix based technique can be combined to QE mechanism implemented in Terrier in the future.

Currently, we are working on fine-tuning and how to get higher MAP scores from the queries that we couldn't enhance.

For now we work only with single words. But plenty of queries have word phrases. Handling word phrases with the single words, can improve results fairly. To do this, of course a mechanism must be developed to catch word phrases from the query words.

Finally, we don't change term weightings after running Terrier. In other words we use them with the same states. Maybe a work can be carried out to adjust weights. i.e., the importance of words in the query must be determined.

6. ACKNOWLEDGEMENTS

We thank ELIAS for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

7. REFERENCES

Amati, G. (2003). Probability Models for Information Retrieval based on Divergence from Randomness. PhD thesis, University of Glasgow.

Can, F, Kocberber, S. Balcik, E., Kaynak,C. Ocalan, H.C., Vursavas, O.M. (February 2008), Information retrieval on Turkish texts. Journal Of The American Society For Information Science And Technology, vol. 59, no.3, pp.407-421.

Cover, T.M., Thomas, J.A. (1991) Elements of Information Theory. Wiley-Interscience, New York, USA.

Harris, Z. (1954). Distributional structure. Word, 10(23), 146–162.

Landauer, T.K. Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.

Manning, C. D., Raghavan, P., Schütze, H. (2009) An Introduction to Information Retrieval. Cambridge University Press, Cambridge, England.

S-Space Package. https://code.google.com/p/airhead-research/ (12 April 2015)

The Terrier IR Platform. http://terrier.org/docs/v3.6/ (15 September 2012)

trec_eval. http://trec.nist.gov/trec_eval/ (26 September 2012)

Turney, P. D., Pantel, P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research. 37, 141-188