Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 705–716
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Improving query precision using semantic expansion

Ahmed Abdelali [a,*], Jim Cowie [a], Hamdy S. Soliman [b]

[a] *Computing Research Laboratory, New Mexico State University, Box 30001/MSC 3CRL, Las Cruces, NM 88003, United States*
[b] *Computer Science Department, New Mexico Institute of Mining and Technology, Socorro, NM 87801-0389, United States*

Received 9 March 2006; received in revised form 6 June 2006; accepted 7 June 2006
Available online 15 December 2006

Abstract

Query Expansion (QE) is one of the most important mechanisms in the information retrieval field. A typical short Internet query will go through a process of refinement to improve its retrieval power. Most of the existing QE techniques suffer from retrieval performance degradation due to imprecise choice of query's additive terms in the QE process. In this paper, we introduce a novel automated QE mechanism. The new expansion process is guided by the semantics relations between the original query and the expanding words, in the context of the utilized corpus. Experimental results of our "controlled" query expansion, using the Arabic TREC-10 data, show a significant enhancement of recall and precision over current existing mechanisms in the field.

Keywords: Query Expansion; Information Retrieval; Precision; Latent Semantic Analysis; Arabic

1. Introduction

In the Web environment, where collections tend to be enormous, it is so important to have robust queries. Typically users submit short queries that do not consider the variety of terms used to describe a topic, resulting in poor recalling power. Moreover, they tend to be too broad to retrieve relevant documents of a specific topic, thus lacking precision. In 2003, Chau, Fang, and Liu Sheng (2005) used the Utah state government website (http://www.utah.gov/), and captured about 2 million queries, over a period of 168 days. In their results, the obtained mean and median of the number of terms in a query are 2.25 and 2, respectively. Spink, Wolfram, Jansen, and Saracevic (2001) have obtained the same median of 2 after analyzing a log of 1,025,910 user queries submitted during a portion of a single day, on the "Excite" search engine. The study showed also that among the 32% of the users who modified their queries, about 29.3% of them added one more term, and 15.5% shortened them by one term. Spink et al. (2001) have concluded that Web users tend to go more often from broad to narrow query formulations, via word addition for more precision.

[*] Corresponding author. Tel.: +1 5056465711.
E-mail addresses: ahmed@crl.nmsu.edu (A. Abdelali), jcowie@crl.nmsu.edu (J. Cowie), hss@nmt.edu (H.S. Soliman).

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.06.007

Instead of the expansion mentioned above, we will introduce an automated process for expanding search queries. The automated process is more efficient since it improves both recall and precision.

In Section 2 of this article we will survey the query expansion issue. The QE problems and some suggested solutions from the literature are discussed in Section 3. In Section 4, we describe our new technique of query expansion which is driven by semantic similarity of involved words. In Sections 5 and 6, we thoroughly evaluate the performance of our system using the TREC-10 data set and discuss the results. Our conclusion and future work are depicted in Section 7.

2. Literature review

One of the most important issues, and still an open ended question, in the Information Retrieval (IR) field is the search for the appropriate wording of a query, for effective information retrieval. Previous work was done to explore methodologies that can enhance the accuracy and the performance of retrieval systems (Cronen-Townsend, Zhou, & Croft, 2004; Efthimiadis, 1996). QE is among the proposed solutions. The overall performance of the QE research is significant in the IR process. Obtained results showed that the approach enhances the precision of IR systems (Imai, Nigel, & Jun'ichi, 1999).

The QE mechanism might be applied via one of the following strategies: manual, interactive, or automatic. For the manual procedure, a user modifies the query by adding or removing words, selectively. QE via the interactive approach, known as "relevance feedback", involves user selected relevant documents to expand the query. The automated query expansion does not involve the user. Upon query submission, few among the top retrieved documents are assumed relevant and therefore used to formulate the new query (pseudo-feedback). Harman (1992) gives a detailed account of relevance feedback and other query reformulation techniques. Additional work in QE can be found in the literature (Efthimiadis, 1996; Harman, 1992; Mitra, Singhal, & Buckley, 1998; Qiu & Frei, 1996; Xu & Croft, 1996).

The application of relevance feedback for query expansion has been experimented within different retrieval systems. The effectiveness of relevance feedback has been demonstrated to improve the recall and the precision of IR systems (Robertson & Sparck-Jones, 1976; Salton & Buckley, 1990). The process of relevance feedback utilizes only the user selected relevant documents from the retrieved list, in response to initial query. Such selected documents will be used to re-weight, expand, and re-formulate a new searching query. Even though the user has the ability to choose the most relevant documents, such relevancy has to exist, in the first place. Lack of documents relevancy would result in less efficient manual expansion. There are many research efforts contributing to the design and improvement of the QE process. Hybrid and automatic approaches for expanding queries are among the current experimental efforts in the IR field. Kekalainen and Jarvelin (1998) evaluated experiments to analyze expansion versus non-expansion performance using a thesaurus. They concluded, in general, that the best performance was obtained using the largest expansion, utilizing all semantic relationships in the thesaurus. Failing cases might be due to either the thesaurus did not provide accurate relations, or that the query was very precise. Grefenstette (1992) designed a system to draw the syntactic relationship between words. The assumption is that the words that appear in the same context share the same trait.

3. Corpus for query expansion

The main concern about utilizing QE in IR systems is that the expansion is not always beneficial. The expansion might degrade the performance of some individual queries (Cronen-Townsend et al., 2004). A lower performance might be due to "Query Drift", when the new query has little re-semblance to the original (Mitra et al., 1998; Stenmark, 2005). Consequently, there will be a subject shift in the retrieved documents. Hence, adding more words does not always improve the performance.

Additional QE concerns might be due to the utilized thesaurus, involving its compilation, maintenance, and updating. Hence, it is important to select the appropriate thesaurus for the QE process.

Alternative solutions were suggested using "linguistic corpus", to overcome the above concerns. A linguistic corpus has a large collection of unified, well structured, and balanced texts, usually annotated as well. The coverage of the corpus can be verified, e.g. the amount and the content variety of texts. Recently, it is much easier to compile corpora over the Internet. Processing this huge amount of data can be fully automated. Successful experiments on building corpora from online data have been acknowledged in (Abdelali, Cowie, & Soliman, 2005; Goweder & De Roeck, 2001). Projects at Linguistic Data Consortium (LDC) and The European Language Resources Association (ELRA) have significantly contributed to the availability of such resources. The sentences within a corpus define the relations between their words, e.g. re-occurrences of some words in several sentences will reflect a stronger relation between these words. Among the advantages of using corpora over thesaurus is defining more accurate relations between words. Hence, we would have a better control over the QE. Moreover, the corpora require lesser time and efforts.

The sense that a word takes in one context is defined by its neighboring words. Therefore the statistical information that can be gathered from a corpus will define the senses that a word takes at each occurrence. Using Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1998), each meaning of a word or a sentence is modeled as a vector in the semantic space, where the meaning of a sentence is the sum of its words' vectors. Although this process neglects the sentence syntax, LSA has proven to be a powerful and useful tool. Consider the following examples (Kintsch & Bowles, 2002):

The stock market collapsed.
The bridge collapsed.

The meaning of the word "*collapsed*" depends on its immediate context. In LSA, the vector rep-resenting a predicate combined with the neighboring words' vectors would reflect the actual sense of such predicate. In the first sentence, the meaning of *collapsed* combined with the meaning of "*stock -market*" to create a different context-sensitive interpretation from the meaning of the same word in the second sentence.

If two constructs (words or sentences) have an angle θ between their representing sense vectors, then their semantic similarity measure is a function of $\cos\theta$. Constructs with matching semantic ($\theta = 0$) have a $\cos\theta$ near 1; whereas constructs with orthogonal senses ($\theta = 90$) have $\cos\theta$ near 0.

Stenmark (2005) utilized the LSA but on word-by-word semantic vectors matching between words from the query and the corpus. He did not achieve the expected enhancement of applying his QE approach.

4. System design and description

We implemented a novel QE approach that utilizes the LSA mechanism for more efficient and reliable QE process. In our approach, we construct a total sense vector that represents the en-tire query semantics. Then, instead of basing the QE on matching semantics "word-by-word" between query words and words from the corpus, we use the new total query vector to find the closest matching set of words/documents' vectors in the corpus. Selectively, we expand the query utilizing the obtained set of words. Our approach is motivated by the intuition that searching the corpus with the total query semantic vector would obtain more precise and supportive words/documents to enrich the initial query. As we expected, experimental results of our new system supported our hypothesis.

We used two different settings to experiment with our approach:

1. Expanding using words.
2. Expanding using documents.

In both approaches, we used LSA to compute the corresponding sense vector for every query, after removing the "stop-list" words. Such list contains the most frequent words in the corpus. Stop words are considered irrelevant for searching purposes because they occur very frequently in the language and they tend to slow down the search without improving the results. The stop-list will be also ignored in the process of indexing the documents (see Section 5.2). Using the refined query, we build its corresponding sense vector that is used to obtain, from the corpus, a list of the top n closest words' or documents' vectors to the query in hand. The obtained list will be used to expand the query.

5. System evaluation

5.1. Data

For the purpose of evaluating our new system, we used the LDC Arabic Newswire, a corpus composed of articles from the Agence France Presse (AFP) Arabic Newswire. The LDC corpus size is 869 megabytes which is divided over 383,872 documents. The corpus includes articles from 13 May 1994 to 20 December 2000 with approximately 76 million tokens and 666,094 unique words. In order to facilitate the manipulation of Arabic text, we transliterate the cp-1256 (Arabic Windows) encoded text to Roman codes (Romanization). Due to memory limitation, we could not use the entire corpus to build the se-mantic space;, instead we randomly used only half of it, which accounts for about 37 million words and around 341,000 unique terms. In order to validate the obtained results for the random selection, we executed several runs to check the variations of the results.

5.2. The retrieval process

As part of the evaluation of our new QE technique, a Unicode Retrieval System Architecture (URSA) search engine (Abdelali, Cowie, Farwell, & Ogden, 2004; Ogden & Davis, 1998) is used to index the entire AFP corpus. A total of 25 Queries from TREC-10 topics is involved to evaluate the system, comparing the results with those of a "baseline-system" with no query expansion. Relevant documents are retrieved for newly regenerated queries (i.e., expanded) via the obtained top *n* closest words or documents to the original query, from the corpus. TRECEVAL software, by TREC, is deployed to evaluate the precision and recall of the retrieval process, as part of assessing the performance of our system.

5.3. Experimental results

In order to assure efficient and precise retrieval process of our QE approach, we define a notion of acceptable *threshold* (lower limit) of similarity between the query and corpus vectors. Such threshold is imposed on the selection of the aforementioned top *n* words/documents to be used in the QE process.

Throughout the expansion process, we choose different values for *n*, e.g., $n = 2, 5, 10, 15, 20, 30$, and 50. For every value of *n*, we expand the query using only those words or documents that possess $\cos\theta \geqslant 0.5$ (i.e., $\theta \leqslant 45$), where θ is the angle between each of their LSA sense vectors and the original query sense vector. The new re-formulated query is run against the indexed corpus; and the retrieved documents is compared to those obtained from baseline system, without any query expansion.

5.3.1. Experimenting with word expansion

Using the threshold cosine similarity measurement, with the total query vector, words are sorted and the top *n* used in the expansion. Table 1 shows that the results of expanding queries using words have retrieved more relevant documents, specifically, we have achieved a 5% increase in number of relevant documents retrieved over the baseline system. Yet, for higher values of *n* (e.g., 30, 50), the number of relevant documents declined.

Tables 2 and 3 provide an overview of the similarities of words to the topics, while the words in Table 2 show weak relation to the topic 19 in Fig. 1. Moreover, topic 20 got a significant boost with the additional words that are much more related. The words highlighted in bold in Table 4 carry more semantic relation to the topic 20.

5.3.2. Experimenting with document expansion

When the expansion uses documents, the new reformulated query is too long to be handled by the retrieval system. Therefore, the expanded query is divided into smaller sub-queries, to be submitted separately. As expected, there will be a huge number of retrieved documents for the generated sub-queries. Hence, we deploy either local or global pruning techniques to trim the obtained set of documents. In the global pruning, we sort all the documents retrieved by the sub-queries, according to their relevancy to the query and get the reasonable

Table 1
TREC-10 retrieval results using words expansion

Words added	0[a]	2	5	10	15	20	30	50
Relevant	4122	4122	4122	4122	4122	4122	4122	4122
Rel_ret	1452	1518	1512	1504	1489	1477	1420	1381
% Change		+4.54	+4.13	+3.58	+2.55	+1.72	−2.20	−4.89
Recall at								
0	0.6533	0.6942	0.6889	0.7088	0.7079	0.6985	0.7098	0.6860
0.1	0.3004	0.3106	0.3154	0.3157	0.3165	0.3147	0.3040	0.2967
0.2	0.2246	0.2386	0.2379	0.2368	0.2346	0.2325	0.2211	0.2100
0.3	0.18	0.1876	0.1874	0.1867	0.1850	0.1834	0.1772	0.1704
0.4	0.1294	0.1349	0.1327	0.1320	0.1300	0.1273	0.1234	0.1233
0.5	0.0973	0.0990	0.0984	0.0981	0.0951	0.0932	0.0918	0.0934
0.6	0.0717	0.0734	0.0724	0.0724	0.0706	0.0704	0.0694	0.0680
0.7	0.0457	0.0482	0.0478	0.0476	0.0466	0.0463	0.0443	0.0439
0.8	0.0287	0.0287	0.0286	0.0284	0.0272	0.0267	0.0257	0.0250
0.9	0.0208	0.0204	0.0200	0.0200	0.0196	0.0192	0.0186	0.0178
1	0.0015	0.0020	0.0019	0.0019	0.0019	0.0019	0.0018	0.0016
Precision at								
5	0.384	0.3920	0.3920	0.4000	0.3920	0.3840	0.3760	0.3600
10	0.368	0.3720	0.3680	0.3720	0.3680	0.3720	0.3800	0.3440
15	0.3307	0.3413	0.3387	0.3493	0.3440	0.3387	0.3413	0.3253
20	0.318	0.3240	0.3220	0.3320	0.3300	0.3280	0.3240	0.3120
30	0.28	0.2920	0.2893	0.2920	0.2947	0.2933	0.2813	0.2707
100	0.1768	0.1860	0.1848	0.1844	0.1820	0.1796	0.1696	0.1652
200	0.1336	0.1400	0.1404	0.1392	0.1378	0.1364	0.1312	0.1250
500	0.0843	0.0875	0.0868	0.0862	0.0852	0.0846	0.0814	0.0800
1000	0.058	0.0607	0.0604	0.0601	0.0596	0.0590	0.0568	0.0552
Significance		0.0002	0.0011	0.0024	0.0073	0.0171	*0.7152*	0.0040

[a] 0: Means no expansion for the query – baseline system.

Table 2
Top 20 closest words to topic 19

Word	Gloss	Freq.	Cosine
جديدة	New	30,374	0.682221
الجديدة	The new	17,086	0.463125
الاسكندرية	Alexandria	3348	0.452056
هناك	There	27,587	0.446183
واهديها	Give it	2	0.404811
ادشنها	Inaugurate it	1	0.402826
منارتها	Its minaret	1	0.402826
الفلورسنتي	Florescent	1	0.402826
الدينا	World	2	0.402826
هيروغيليفية	Hieroglyphic	1	0.402826
الارتكاس	Reflex	1	0.402826
ولاتينية	Latin	3	0.390386
ونفعل	And do	11	0.387532
البناء	The building	177	0.387407
دارول	Darolles	11	0.385701
وتمديدات	And extensions	8	0.382602
ردم	Fill up	134	0.381007
تفعلها	Do it	7	0.374624
لمنارة	For minaret	5	0.374080
سوربيلي	Sorbilli	1	0.373508

Table 3
Top 20 closest words to topic 20

Word	Gloss	Freq.	Cosine
القاهرة	Cairo	53,123	0.815343
المصرية	The Egyptian	18,737	0.770064
الجيزة	Giza	490	0.730608
مصرية	Egyptian	1695	0.726129
الاهرام	The pyramids	2172	0.725369
مصري	Egyptian	2861	0.708850
والاسكندرية	And Alexandria	135	0.703337
المصريين	The Egyptians	3531	0.694185
مصر	Egypt	43,938	0.674919
الاقصر	Luxor	1472	0.659686
مصريا	Egyptian	654	0.655602
بالقاهرة	In Cairo	239	0.641676
البلتاجي	El Beltagui	185	0.639177
مصريون	Egyptians	291	0.638932
اهرامات	Pyramids	144	0.638750
الاهرامات	The Pyramids	176	0.629453
المصريون	The Egyptians	721	0.627789
اهرام	Pyramids	71	0.627404
لمصر	For Egypt	1506	0.614105
زيات	Zayat	6	0.610130

number (say D) of top ranked documents. Localized pruning selects the top D/m documents from each of the m sub-queries retrieved documents. For the best optimal pruning process, we are exploring solutions from the "data fusion" research field (Vogt, 2000).

Compared to word expansion, Table 4 shows better results of document expansion. Specifically, we obtained an additional 60% in-crease in the number of relevant documents. Tables 5 and 6 provides an overview of the similarities of documents to the topics 19 and 20 respectively.

For example in Table 7, at a Recall point of 0.7, document expansion approach achieved an additional 98.91% (0.0909 versus 0.0457 of the baseline system) higher precision when compared to 4.60% (similarly 0.0478 versus 0.0457) precision by word expansion.

Also, as shown in Table 7, the precision at higher number are better than the baseline system. Moreover, we obtained good results with respect to relevant documents. Yet, we are still concerned about the lack of better precision at lower recall points. A possible explanation might be due to the aforementioned pruning strategies we used.

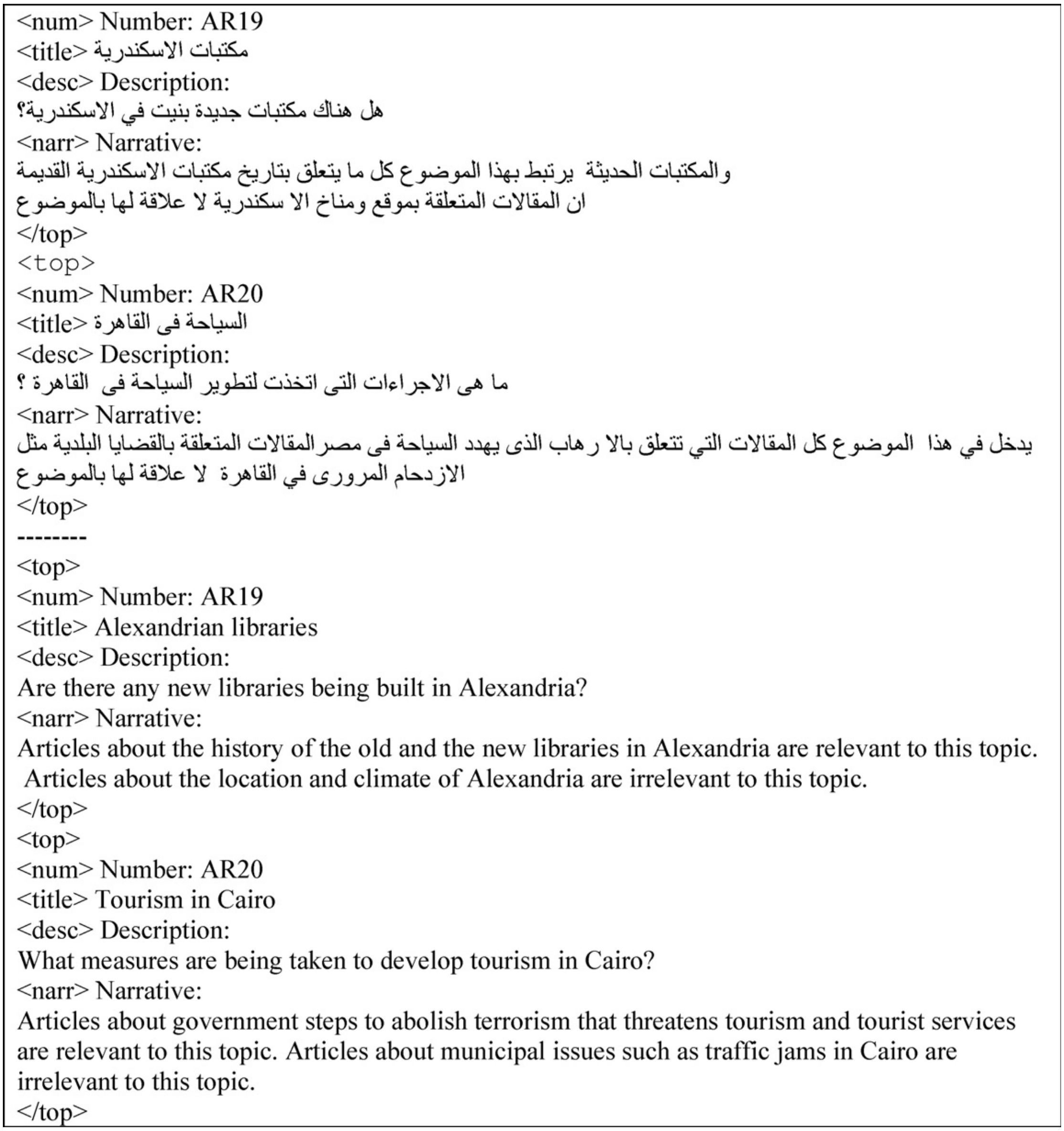
5.3.3. Re-visiting word expansion

The results of using document expansion, especially in recalled documents (Table 6), seem very encouraging. Hence, motivated by the document expansion results, we deployed the same approach of using the query division mechanism in word expansion. The major drawback of the sub-queries approach is the extra time (overhead) of manufacturing and processing m sub-queries for every original query. The new approach has been evaluated using the same sets of data and topics. The results showed significant increase (between 45% top 110%) in the number of relevant retrieved document; with the obtained precision between 5% and 10% (see Table 8).

6. Discussion

Fig. 2 histograms display the achievements of the extended queries comparing to the baseline system, an average of an additional 50 relevant documents has been retrieved.

Among the 25 queries used for the evaluation, four queries were not improved – two among them got worse results than the baseline system (see Fig. 2). Such degradation is insignificant compared to the positive results achieved with other queries.



```
<num> Number: AR19
<title> مكتبات الاسكندرية
<desc> Description:
هل هناك مكتبات جديدة بنيت في الاسكندرية؟
<narr> Narrative:
والمكتبات الحديثة  يرتبط بهذا الموضوع كل ما يتعلق بتاريخ مكتبات الاسكندرية القديمة
ان المقالات المتعلقة بموقع ومناخ الا سكندرية لا علاقة لها بالموضوع
</top>
<top>
<num> Number: AR20
<title> السياحة فى القاهرة
<desc> Description:
ما هى الاجراءات التى اتخذت لتطوير السياحة فى  القاهرة ؟
<narr> Narrative:
يدخل في هذا  الموضوع كل المقالات التي تتعلق بالا رهاب الذى يهدد السياحة فى مصرالمقالات المتعلقة بالقضايا البلدية مثل
الازدحام المرورى في القاهرة  لا علاقة لها بالموضوع
</top>
--------
<top>
<num> Number: AR19
<title> Alexandrian libraries
<desc> Description:
Are there any new libraries being built in Alexandria?
<narr> Narrative:
Articles about the history of the old and the new libraries in Alexandria are relevant to this topic.
 Articles about the location and climate of Alexandria are irrelevant to this topic.
</top>
<top>
<num> Number: AR20
<title> Tourism in Cairo
<desc> Description:
What measures are being taken to develop tourism in Cairo?
<narr> Narrative:
Articles about government steps to abolish terrorism that threatens tourism and tourist services
are relevant to this topic. Articles about municipal issues such as traffic jams in Cairo are
irrelevant to this topic.
</top>
```

Fig. 1. TREC-10 topic 19 and 20.

From the obtained results, the failure of the two queries might be related to their length, since the length fairly correlates negatively with the performance of the expansion (see Table 8). Short queries tend to perform better than the long ones. Long queries are hard to expand because the length reflects the semantic richness of the query, hence adding more words would slightly increase the query sense Table 9.

The analysis of our results indicates consistent improvement of the word expansion precision, compared to the baseline system. Moreover, our strategy of controlled query expansion, using the expansion threshold, promises a much better IR performance than existing LSA based competitive techniques.

In summary, our technique shows a significant increase in recall and precision due to the use of small, but semantically related numbers of words/documents in the expansion process. The uncontrolled (not well thought out) addition of words to the query might hurt the retrieval process, and the expansion will not be beneficial.

Table 4
TREC-10 retrieval results using documents expansion

Docs added	0	2	5	10	15	20	30	50
Relevant	4122	4122	4122	4122	4122	4122	4122	4122
Rel_ret	1452	2008	2422	2396	2324	2331	2338	2248
% Change		+38.29	+66.80	+65.01	+60.06	+60.54	+61.02	+54.82
Recall at								
0	0.6533	0.5735	0.5509	0.5431	0.2773	0.3052	0.3032	0.2994
0.1	0.3004	0.2050	0.1815	0.1534	0.1569	0.1534	0.1546	0.1469
0.2	0.2246	0.1731	0.1598	0.1357	0.1375	0.1429	0.1439	0.1351
0.3	0.18	0.1464	0.1314	0.1093	0.1179	0.1112	0.1231	0.1115
0.4	0.1294	0.1165	0.1147	0.0959	0.1001	0.0997	0.0991	0.0953
0.5	0.0973	0.0965	0.1032	0.0855	0.0869	0.0887	0.0807	0.0859
0.6	0.0717	0.0824	0.0950	0.0824	0.0831	0.0775	0.0799	0.0779
0.7	0.0457	0.0680	0.0909	0.0792	0.0718	0.0752	0.0789	0.0769
0.8	0.0287	0.0501	0.0684	0.0627	0.0710	0.0729	0.0769	0.0764
0.9	0.0208	0.0338	0.0667	0.0596	0.0688	0.0713	0.0752	0.0757
1	0.0015	0.0146	0.0643	0.0595	0.0664	0.0705	0.0722	0.0742
Precision at								
5	0.384	0.2320	0.1440	0.1440	0.1280	0.1600	0.1680	0.1600
10	0.368	0.2360	0.1720	0.1080	0.1160	0.1200	0.1240	0.1440
15	0.3307	0.2213	0.1867	0.1307	0.1120	0.1147	0.1227	0.1307
20	0.318	0.2200	0.1880	0.1380	0.1180	0.1220	0.1200	0.1340
30	0.28	0.2080	0.1853	0.1440	0.1200	0.1240	0.1173	0.1253
100	0.1768	0.1496	0.1520	0.1388	0.1180	0.1204	0.1236	0.1224
200	0.1336	0.1108	0.1218	0.1198	0.1122	0.1168	0.1196	0.1180
500	0.0843	0.0798	0.0991	0.0929	0.1006	0.1031	0.1054	0.1045
1000	0.058	0.0802	0.0967	0.0957	0.0929	0.0932	0.0934	0.0898
Significance		0.0115	0.0274	0.0172	0.0130	0.0139	0.0146	0.0137

Table 5
Top 20 closest documents to topic 19

Document ID	Cosine
47264	0.405001
125889	0.394175
80502	0.385453
40142	0.374702
55825	0.372586
70809	0.370445
165297	0.368395
33764	0.366126
70931	0.357538
64556	0.355980
100727	0.352582
42111	0.352029
172371	0.350262
178998	0.349964
80473	0.348826
95613	0.348401
13293	0.347277
40152	0.346684
157325	0.346358
25351	0.346158

All the results of the above experiments were statistically significant using Student's t-test at a p-value of <0.05.

Table 6
Top 20 closest documents to topic 20

Document ID	Cosine
118324	0.683647
114450	0.641621
138892	0.634451
99204	0.633222
74029	0.623068
62904	0.621790
72913	0.615625
128822	0.609677
117669	0.607929
69583	0.596922
130671	0.593362
17952	0.591798
86686	0.584316
103063	0.579152
188926	0.578326
188925	0.578326
44386	0.576235
118465	0.575904
96158	0.573629
87406	0.572638

Table 7
Recall and precision performance for expansion using 5 words and documents

	5 words	5 documents
Recall at		
0	5.45%	−15.67%
0.1	4.99%	−39.58%
0.2	5.92%	−28.85%
0.3	4.11%	−27.00%
0.4	2.55%	−11.36%
0.5	1.13%	6.06%
0.6	0.98%	32.50%
0.7	4.60%	98.91%
0.8	−0.35%	138.33%
0.9	−3.85%	220.67%
1	26.67%	4186.67%
Precision at		
5	2.08%	−62.50%
10	0.00%	−53.26%
15	2.42%	−43.54%
20	1.26%	−40.88%
30	3.32%	−33.82%
100	4.52%	−14.03%
200	5.09%	−8.83%
500	2.97%	17.56%
1000	4.14%	66.72%

7. Conclusion and future work

This paper demonstrates that the QE process can play a significant role in the performance enhancement of IR systems. The most critical factor in the process is the selection of words/documents to expand the query. The wrong choice of expansion constructs might harm the retrieval process by drifting it away from the optimal correct answers.

Table 8
TREC-10 retrieval results using the new word expansion approach

Words	0*	2	5	10	15	20	30	50
Relevant	4122	4122	4122	4122	4122	4122	4122	4122
Rel_ret	1452	2109	2436	2567	2668	2773	2914	3063
% Change		45.25	67.77	76.79	83.75	90.98	100.69	110.95
Recall at								
0	0.6533	0.6744	0.6837	0.7104	0.6865	0.6865	0.6696	0.6786
0.1	0.3004	0.3286	0.3784	0.3706	0.3843	0.3779	0.3777	0.3882
0.2	0.2246	0.2744	0.3273	0.3138	0.3176	0.3181	0.3161	0.3134
0.3	0.18	0.2397	0.2792	0.2635	0.2652	0.2632	0.2649	0.2692
0.4	0.1294	0.2152	0.2337	0.2343	0.2452	0.2445	0.2452	0.2493
0.5	0.0973	0.1865	0.2155	0.214	0.2182	0.2152	0.2228	0.2224
0.6	0.0717	0.1575	0.1808	0.1891	0.195	0.1956	0.2003	0.2003
0.7	0.0457	0.1298	0.1614	0.1734	0.1828	0.1806	0.1854	0.1873
0.8	0.0287	0.1092	0.1473	0.1541	0.1621	0.1647	0.1684	0.1734
0.9	0.0208	0.0936	0.1202	0.1468	0.1457	0.1539	0.1588	0.1618
1	0.0015	0.0838	0.111	0.1125	0.1236	0.1357	0.1395	0.1594
Precision at								
5	0.384	0.408	0.44	0.464	0.488	0.488	0.464	0.424
10	0.368	0.4	0.372	0.392	0.396	0.404	0.412	0.396
15	0.3307	0.3947	0.3707	0.3787	0.3893	0.3947	0.3973	0.384
20	0.318	0.364	0.358	0.364	0.386	0.39	0.392	0.38
30	0.28	0.3333	0.3387	0.332	0.3547	0.356	0.3653	0.356
100	0.1768	0.2336	0.2512	0.2464	0.2608	0.2608	0.2652	0.2712
200	0.1336	0.1794	0.2016	0.1986	0.2086	0.2112	0.2144	0.218
500	0.0843	0.1236	0.1354	0.1404	0.1481	0.1514	0.1564	0.1612
1000	0.058	0.0844	0.0974	0.1026	0.1067	0.1108	0.1165	0.1224
Significance		0.00001	0.00006	0.00000	0.0000	0.00000	0.00000	0.00001

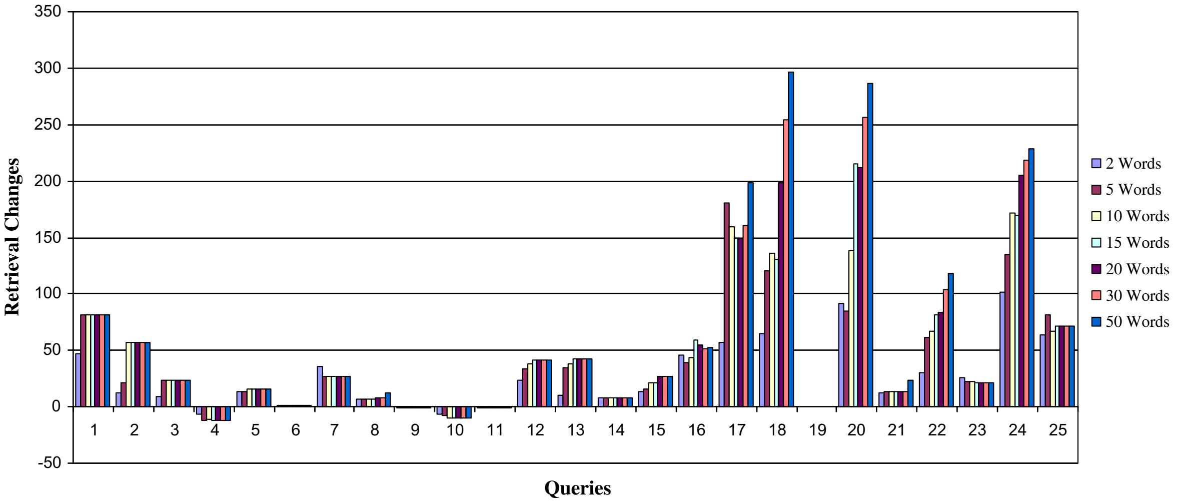


Fig. 2. Retrieved documents changes comparing to the baseline system using the new expansion approach.

In order to avoid the harmful expansion, expanding words/documents might be selected under the guidance of information inferred from corpora. We designed a mechanism that can automatically select corpus words

Table 9
Retrieval performance and length correlation by topic using word expansion

Topic	Length	0*	2	5	10	15	20	30	50
1	66	47	82	82	82	82	82	82	47
2	69	12	21	57	57	57	57	57	12
3	48	9	24	24	24	24	24	24	9
4	73	−7	−12	−11	−12	−12	−12	−12	−7
5	89	13	13	16	16	16	16	16	13
6	166	1	1	1	1	1	1	1	1
7	73	36	27	27	27	27	27	27	36
8	80	7	7	7	7	8	8	12	7
9	114	−1	−1	−1	−1	−1	−1	−1	−1
10	41	−6	−8	−10	−10	−10	−10	−10	−6
11	56	−1	−1	−1	−1	−1	−1	−1	−1
12	40	23	34	38	41	41	41	41	23
13	26	10	35	38	43	43	43	43	10
14	57	8	8	8	8	8	8	8	8
15	46	14	16	21	21	27	27	27	14
16	31	46	39	44	59	55	51	53	46
17	43	57	181	159	149	150	161	198	57
18	45	65	120	136	131	198	254	296	65
19	33	0	0	0	0	0	0	0	0
20	46	91	85	138	215	212	256	286	91
21	48	12	14	14	14	14	14	24	12
22	58	30	61	67	82	84	104	118	30
23	43	26	22	22	21	21	21	21	26
24	41	101	135	172	170	205	219	229	101
25	53	64	81	67	72	72	72	72	64
Total		1452	2109	2436	2567	2668	2773	2914	3063
Correlation			−0.319	−0.32	−0.319	−0.315	−0.298	−0.293	−0.32

that are semantically related to the query, in the expansion process. The major advantage of our approach is its well thought out mathematical approach in selecting the query expanding words/documents from the corpus. Only corpus words with the largest sense vector similarity (within some "cosine" threshold) to the sense vector of the "entire" initial query will be selected for the expansion process. Such algorithmic approach will guarantee the usefulness of the expansion, rather than inefficient traditional blind QE. Moreover, it assures the topic consistency and helps in a stable QE process that would not result in topic shift or query drift.

The obtained results of our newly automated expansion are very promising. On the other hand, we noticed that some of words added to the query were not frequent in the collection; hence their effect on the retrieval scores is not significant.

In our future work we will investigate a possible solution to the above problem. We might try the inclusion of the frequency of occurrences factor of words/documents in the corpus, as additional criteria, in the QE selection process. We will also experiment with boosting the recall via expanding the query using only more frequent words. The question that we will explore is "Would the use of this category of scarce words lead to the discovery of otherwise unseen documents?" The answer to the above question might be particularly important if there are very few relevant documents in the collection.

References

Abdelali, A., Cowie, J., Farwell, D., & Ogden, W. C. (2004). UCLIR: a multilingual information retrieval tool inteligencia artificial. *Revista Iberoamericana de Inteligencia Artificial, 8*(22), 103–110.

Abdelali, A., Cowie, J., & Soliman, H. (2005). Building a modern standard Arabic corpus. Workshop on Computational Modeling of Lexical Acquisition. The Split Meeting. Split, 25th to 28th of July 2005.

Chau, M., Fang, X., & Liu Sheng, R. O. (2005). Analysis of the query logs of a Web site search engine. *Journal of the American Society for In-formation Science, 56*(13), 1363–1376.

Cronen-Townsend, S., Zhou, Y., & Croft, W.B. (2004). A framework for selective query expansion. A poster presentation. In *Proceedings of CIKM'04* (pp. 236–237).

Efthimiadis, E. N. (1996). Query expansion. In M. E. Williams (Ed.). *Annual review of information science and technology* (Vol. 31, pp. 121–187). Medford NJ: Information Today Inc.

Goweder, A., & De Roeck, A. (2001). Assessment of a significant Arabic corpus. Presented at the Arabic NLP Workshop at ACL/EACL 2001, Toulouse, France, 2001.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In N. Belkin, P. Ingwersen, & A. M. Pesjtersen (Eds.), *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval, Copenhagen, Denmark* (pp. 89–97). New York: ACM Press.

Harman, D. (1992). Relevance feedback and other query modification techniques. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithm* (pp. 241–263). Engle-wood Cliffs, NJ: Prentice-Hall.

Imai, H., Nigel, C., & Jun'ichi, T. (1999). A combined query expansion approach for information retrieval. In *Proceedings of Genome Informatics*. Tokyo, Japan: Universal Academy Press Inc.

Kekalainen, J., & Jarvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 130–137).

Kintsch, W., & Bowles, A. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 17*, 249–262.

Landauer, T. K., Foltz, P., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259–284.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 206–214).

Ogden, W., & Davis, M. (1998). Design, Implementation and User's Guide to URSA, the UNICODE Retrieval System Architecture.

Qiu, Y., & Frei, H. (1993). Concept based query expansion. In *Proceedings of the sixteenth annual international ACM SIGIR conference on research and development in information retrieval* (pp. 160–169).

Robertson, S., & Sparck-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288–297.

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

Stenmark, D. (2005). Query expansion on a corporate intranet: Using LSI to increase precision in explorative search. In *Proceedings of the 38th Hawaii international conference on system sciences – 2005*.

Vogt, C. (2000). How much more is better? characterizing the effects of adding more IR systems to a combination. In *Proceedings of the computer assisted information retrieval international conference (RIAO)* (pp. 457–475).

Xu, J., & Croft, B. (1996). Query expansion using local and global document analysis. In H. Frie, D. Harman, P. Schauble, & R. Witkinson (Eds.), *Proceedings of the nineteenth international ACM-SIGIR conference on research and development in information retrieval* (pp. 4–11). Zurich, Switzerland: ACM Press.