A Query Expansion Technique Using the EWC Semantic Relatedness Measure

Vitaly Klyuev
University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, 965-8580 Japan
E-mail: vkluev@u-aizu.ac.jp

Yannis Haralambous
Institut Télécom – Télécom Bretagne, Dép. Informatique,
UMR CNRS 3192 Lab-STICC Technopôle Brest Iroise, CS 83818, 29238 Brest Cedex 3, France
E-mail: yannis.haralambous@telecom-bretagne.eu

Keywords: relatedness measure, Wikipedia, WordNet, search engine, query expansion

Received: October 28, 2011

This paper analyses the efficiency of the EWC semantic relatedness measure in an ad-hoc retrieval task. This measure combines the Wikipedia-based Explicit Semantic Analysis (ESA) measure, the WordNet path measure and the mixed collocation index. EWC considers encyclopaedic, ontological, and collocational knowledge about terms. This advantage of EWC is a key factor to find precise terms for automatic query expansion. In the experiments, the open source search engine Terrier is utilised as a tool to index and retrieve data. The proposed technique is tested on the NTCIR data collection. The experiments demonstrated superiority of EWC over ESA.

Povzetek: Članek obravnava razširjanje poizvedb z uporabo semantične povezave med besedami.

1 Introduction

A bag-of-words representation of documents by information retrieval systems results in queries expressed utilising the language of keywords. Users face a vocabulary problem: A keyword language is not adequate to describe the information needs. Statistical analysis of styles of user behaviour showed that the queries submitted to search engines are short (2 to 3 terms, on average) and ambiguous, and users rarely look beyond the first 10 to 20 links retrieved [20].

To improve user queries, search engines provide many tools.

Manuals and instructions are among them. They help a little, although ordinary users do not like reading.

A query suggestion feature is common for general purpose search engines. Its disadvantage is in the way to generate recommended terms. They are based on the first term typed by the user, which is always the first one in all expanded queries [22].

The relevance feedback feature is another instrument to help users. There are at least two steps in the interaction with a search engine: The first one is the submission of the original query, and the second one is the user reaction on the results retrieved in order to provide the system with the user opinion (marking some documents as relevant). This feature is not popular among users because the mechanisms of changing queries are not clear and users cannot control the process [9].

In addition, query suggestion and relevance feedback are used to modify the queries in order to make them more accurate to express the user information needs.

Query expansion is a well-known and popular technique to reformulate the user query in order to reduce the number of non-relevant pages retrieved by information retrieval systems. Another goal of query expansion is to provide the user with additional relevant documents. Automatic query expansion is an important area of information retrieval: Many scientists are involved in designing new methods, techniques, and approaches.

This paper presents authors' technique to automatically expand the user queries. This technique is based on the EWC semantic relatedness measure [21]. This measure takes into account encyclopaedic, ontological, and collocational knowledge about terms. The environment for the experiments includes Terrier as a search engine and NTCIR-1 CLIR data collection for the Japanese-English cross-lingual retrieval task.

The rest of the paper is organised as follows. The next section reviews the approaches to automatic query expansion. Section 3 describes the nature of the measure used. Section 4 provides the necessary details related to this technique to expand the queries. The tools and data utilised in the experiments are presented in Section 5.

[*] This paper is based on V. Klyuev and Y. Haralambous, *Query Expansion: Term Selection using the EWC Semantic Relatedness Measure* published in the proceedings of the 1st International Workshop on Advances in Semantic Information Retrieval (part of the FedCSIS'2011 conference).

The results of the experiments are discussed in Section 6. Comments on the future directions of the investigations are presented there. Concluding remarks can be found in Section 7.

2 Related Work

A comprehensive review of the classical approaches to expand queries can be found in [9]. They propose different ways to obtain semantically (topically) related terms, techniques to evaluate importance of the terms found, mechanisms to define the number of terms to add (expand) the user query, and strategies to evaluate the quality of obtained results.

Generally, the semantics of the terms are clear when they are in the sentences because the meanings of the words are fixed and only one is usually selected from the set of all possible variants. However, in the queries, the terms are separate instances, and their semantics are unclear. This is called the polysemy problem. On the other hand, the same things can be described by using different terms. This is the nature of the synonymy problem. The query should be rich enough to include the possible candidates for expressing user information needs.

The general goal of query expansion is to find a solution for these two aforementioned problems. The classical solution for the synonymy problem is to apply thesauri as instruments to obtain the candidates for expansion. WordNet is widely used for this purpose [19]. Modern techniques suggest Wikipedia as a valuable source to find synonyms [12].

Many techniques are used to solve the polysemy problem. Approaches described in [13] and [16] are based on the analysis of the query log files of search engines and clicked URLs. Authors of this study [18] utilised WordNet for a deep analysis of the queries submitted to the information retrieval system in order to find the concepts and then obtain the candidate terms for expansion. The involvement of users is the feature of the approach discussed in [17]. They should select the correct ontology for each query submitted to expand the query.

The authors of this study [14] also pointed out that the information exploited by different approaches differs, and combining the different query expansion approaches is more efficient than the use of any of them separately. They investigated techniques to rank the terms extracted from the retrieved documents. One is based on the measures of occurrence of the candidate and query terms in the retrieved documents. The other one utilises the differences between the probability distribution of terms in the collection and in the top ranked documents retrieved by the system. A similar idea is discussed in [15].

The authors of [10] combined the concept-based retrieval, based on explicit semantic analysis (ESA), with keyword-based retrieval. At the first step, they use keyword-based retrieval to obtain the candidates for query expansion. Then, they tune queries applying ESA. After that, they perform the final retrieval in the space of concepts.

It is difficult to compare the aforementioned approaches, because different data sets were used to evaluate them. In many cases, it is not clear wherever the test queries cover a wide range of data set topics. The performance evaluation is done automatically for some approaches, whereas for others, the authors involve the users to judge the quality of retrieval.

3 Measure Description

In study [21], the new measure of words relatedness is introduced. It combines the ESA measure μ_{ESA} [10], the ontological WordNet path measure μ_{WNP}, and the collocation index C_ξ. This measure is called EWC (ESA plus WordNet, plus collocations) and is defined as follows:

$$\mu_{EWC}(w_1, w_2) = \mu_{ESA}(w_1, w_2) * \alpha$$
$$\alpha = (1 + \lambda_\sigma(\mu_{WNP}(w_1, w_2))) * \gamma$$
$$\gamma = (1 + \lambda'_\sigma(C_\xi(w_1, w_2)))$$

where λ_σ weights the WordNet path measure (WNP) with respect to ESA, and λ'_σ weights the mixed collocation index with respect to ESA. This index is defined as follows:

$$C_\xi = \frac{2 * f(w_1, w_2)}{f(w_1) + f(w_2)} + \xi \frac{2 * f(w_2, w_1)}{f(w_1) + f(w_2)}$$

where $f(w_1, w_2)$, $f(w_2, w_1)$ are the frequency of the collocations of $w_1 w_2$ and $w_2 w_1$ in the corpus, and $f(w_i)$ is the frequency of word w_i. The values for constants λ_σ, λ'_σ, and ξ are set to 5.16, 48.7, and 0.55, respectively on the basis of empirical tests.

Study [21] demonstrated the superiority of this measure over ESA on the WS-353 test set.

The current implementation of EWC does not take into account Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). As a result, the proposed technique cannot distinct multiple term items from collocations and give them the highest score.

4 Technique to Expand Queries

Assume that Z is a pool of term-candidates for query expansion. The formulas below present the method to select terms to expand queries. N is a number of original query terms, and j is an index of them. Values for the

WordNet component and collocation component should be above zero in order to choose related terms. Thresholds t_2 for EWC values and t_1 are parameters adjusted in the experiments. For every word $w_i \in Z$, the weight is calculated. Word w_i is selected for expansion if its weight is equal to 1.

$$weight(w_i) = \begin{cases} 1, if \sum_{j=0}^{N} \frac{score(w_1, w_2)}{N} > t_1 \\ 0, otherwise \end{cases}$$

$$score(w_1, w_2) = \begin{cases} 1, if \mu_{WNP} > 0; C_{\xi} > 0; \mu_{EWC} > t_2 \\ 0, otherwise \end{cases}$$

This approach can be interpreted as follows: A term is selected from the list of term-candidates, if the similarity score between this term and the majority of original query terms is higher than a given threshold t_1. The term-candidate should have non-zero values for μ_{WNP} and C_{ς} components.

5 Tools and Data Sets Used

The open source search engine Terrier [1] was used as a tool to index and retrieve data. It provides the different retrieval approaches. TF-IDF and Okapi's B25 schemas [6, 9] are among them.

As a data set for experiments, the NTCIR CLIR data collection [2] was used. It consists of 187,000 articles in English. These articles are summaries of papers presented at scientific conferences hosted by Japanese academic societies. The collection covers a variety of topics such as chemistry, electrical engineering, computer science, linguistics, and library science. The size of the collection is approximately 275.5 MB. A total of 83 topics are in Japanese. A structure of the dataset and topics is similar to that of TREC [3].

A straightforward approach was applied to translate queries into English: Google's translation service [4] generated queries in English. This method was selected because on-line dictionaries do not work well with terms in katakana and specific terminology [7]. Katakana is one of four sets of characters used in Japanese writing. It is primarily applied for the transcription of foreign language words into Japanese.

A Porter Stemmer algorithm was applied to the documents and queries, and a standard stop word list provided by Terrier was also utilised. Only the title fields were considered as a source of the queries. They are relatively short: each query consists of a few keywords. The authors of the study reported in [5] experimented with Terrier applying the same conditions to the TREC data.

To measure the term similarities, an experimental tool described in [21] was utilised.

6 Results of Experiments

The authors implemented the proposed technique to expand queries as follows.

To obtain the candidates for query expansion, a query expansion functionality offered by Terrier was adopted. It extracts the most informative terms (in this case 10) of the top-ranked documents (in this case 3) by using a particular DFR (divergence from randomness) term weighting model [8].

Table 1 provides the list of original queries (topics 1, 12, and 24), candidate terms for expansion (arranged by the decreasing score calculated by Terrier), and the final sets of terms used to expand queries (they are in bold).

Table 1: Original and Expanded Queries for Topics: 1, 12 and 24.

Topic	Original query	Terms for expansion
1	Robot	**Robot** person **human** multi comput sice design will confer paper
12	Mining methods	Mine method rule data databas associ discoveri **larg** tadashi solv amount
24	Machine translation system	Machin translat **system** exampl **base** **masahiro** **method** nation convert **problem**

Table 2: Thresholds Tuning: Topics 1 to 30.

t1	t2	EWC: Average Precision R-Precision			ESA/ BM25
		InL2	TF-IDF	BM25	
0.5	0.1	0.2940 0.3216	0.3031 0.3314	0.3072 0.3324	**0.3101** **0.3347**
0.65	0.09	0.2940 0.3300	0.2936 **0.3332**	0.2955 0.3278	**0.2961** 0.3276
0.67	0.07	**0.2973** **0.2977**	0.2954 0.2963	0.2960 0.2916	0.2959 0.2916
0.67	0.08	0.3101 0.3277	0.3151 0.3268	0.3140 0.3265	**0.3172** **0.3315**
0.67	0.09	0.3073 0.3256	0.3105 **0.3373**	**0.3106** 0.3352	0.3080 0.3373
0.67	0.1	0.3030 0.3295	**0.3103** **0.3350**	0.3099 0.3349	0.3099 0.3318
0.67	0.11	0.3049 0.3239	**0.3121** **0.3309**	0.3110 0.3292	0.3102 0.3302
0.67	0.12	0.3049 0.3239	**0.3121** **0.3309**	0.3110 0.3292	0.3092 0.3284
0.67	0.13	0.3049 0.3125	**0.3114** 0.3245	0.3099 0.3248	0.3111 **0.3282**
0.67	0.15	0.3033 0.3143	**0.3115** 0.3267	0.3111 **0.3282**	0.3110 0.3282
0.69	0.09	0.3073 0.3256	0.3105 **0.3373**	**0.3106** 0.3352	0.3080 0.3373
0.75	0.1	0.3030 0.3295	0.3103 0.3330	**0.3099** **0.3349**	0.3099 0.3318
System	system	0.2980 0.2995	0.3034 0.3166	0.3017 0.3163	

Candidate terms for expansion are presented in stemmed form after applying the aforementioned Porter algorithm. One to five terms were selected by this method. As one can see from this table, this technique does not usually select the top-ranked terms as candidates for expansion from the Terrier engine point of view.

As mentioned in Section 5, a total of 83 topics are available to retrieve documents from the collection. The original goal of topics 0001 to 0030 is to tune the parameters of the retrieval system. Relevance judgments for some of them are known in advance. Topics 0031 to 0083 were used in official runs at the NTCIR 1 Workshop. Organisers found that the number of relevant documents for 13 topics of the 53 contained less than five relevant documents per topic in cross-lingual retrieval. Hence, they discarded these topics from evaluation [25]. The full set was used in these experiments because the goal is to compare the performance of different methods implemented in the same environment.

Table 3: Evaluation Results for Topics 31 to 83.

Recall level	Precision		
	EWC	ESA	System
at 0.00	0.9594	0.9702	0.9676
at 0.10	0.7911	0.7852	0.7972
at 0.20	0.7007	0.6904	0.5859
at 0.30	0.5288	0.5245	0.5001
at 0.40	0.3642	0.3586	0.3599
at 0.50	0.2878	0.2856	0.3011
at 0.60	0.176	0.1767	0.181
at 0.70	0.1597	0.1571	0.1394
at 0.80	0.1061	0.101	0.0862
at 0.90	0.0557	0.0533	0.0354
at 1.00	0.0404	0.0405	0.0242

In the evaluations, the partially relevant documents were considered as irrelevant. To archive this, the corresponding file with the answers provided by NTCIR Workshop organisers was applied when evaluating the retrieval results.

Table 2 summarises the results of retrieval to tune thresholds t_1 and t_2. The test queries were generated from topics 0001 to 0030. It is important to note that when the queries are expanded with all the terms proposed by Terrier, the retrieval results drop to zero. The retrieval utilising the TF-IDF schema produced better results for the original queries (without expansion) compared to the BM25 and InL2 models [1]. The line *system* shows this result. The first number in the cells is the value of average precision, and the second one is the value of R-precision. The performance of retrieval with expanded queries utilising the ESA and EWC approaches for the threshold values (t_1 equals 0.67 and t_2 ranges from 0.08 to 0.15) is better compared to the variant without expansion. For the EWC measure, the maximum of the retrieval performance is reached when the values of thresholds t_1 and t_2 are set to 0.67 and 0.12. For ESA, the optimal threshold values are 0.67 and 0.08. The performance of ESA is higher than EWC.

Table 3 shows precision/recall evaluation results across 53 queries.

Table 4 summarises the results of retrieval for topics 31 to 83. The threshold values were set to the optimal parameters (see Table 2). Six runs were executed. The EWC measure demonstrated better performance (see values in bold) over ESA in both cases (average

Table 4: Retrieval Results: Topics 31 to 83.

t1	t2	EWC: Average Precision R-Precision		ESA: Average Precision R-Precision	
		BM25	TF-IDF	BM25	TF-IDF
0.67	0.08	**0.2363** **0.2416**		0.2349 0.2390	
0.67	0.12		0.2200 0.2317		0.2161 0.2284
System	system	0.2225 0.2350	**0.2218** **0.2359**		

Table 5: Differences in the Expanded Queries for Topics 31 to 83.

Topic	Original Queries	Expansion Terms	
		EWC	ESA
32	Network Collaboration	design technolog web world tool focu wide	design technolog world tool focu wide
37	buffer control	buffer control memori input	memori input
38	TCP / IP Throughput Performance Communications	network	
46	Reset period algorithm	system	
56	Information Lifecycle artifacts knowledge sharing	knowledg design product life	knowledg
62	Lifelong learning and volunteer	learn educ	educ
69	Computer-aided teaching	aid teach educ instruct person system	aid educ person system
75	Simulation exercise	system work model	system work
81	Sex differences in brain	differ mean	differ
82	Antimalarial drugs	antimalari drug	antimalari

precision and R-Precision). The line *system* shows the retrieval results without expansion for TF-IDF and BM25 schemes

Table 5 presents the differences in the expanded queries for topics 31 to 83 for EWC and ESA metrics. Expansion terms are presented in stemmed form in this table. One can see that some expansion terms are taken from original queries. This is the case for topics 37, 56, 62, 69, 81, and 82. Such a selection increases the importance of the respective search terms at the searching process.

Among 53 topics, there are only the differences in topics 32, 37, 38, 46, 56, 62, 69, 75, 81, and 82. For the remaining topics, both approaches (EWC and ESA) generated the same queries. In other words, only these 10 queries contributed in improvements of the performance of the search.

Original queries for topics 38 and 46 were not expanded when the ESA approach was applied.

As we mentioned in Section 3, the implementation of EWC discards Wikipedia articles with titles consisting of multiple terms (they are dimensions in the Wikipedia space). Multiple word terms cannot be recognized and scored accurately. These terms are widely used in scientific terminology. Recognition of such terms seems to be the most promising direction to enhance EWC.

To summarize, one can conclude that the EWC measure provides little benefit over ESA, as the results of the retrieval are better. Ontological knowledge combined with collocational knowledge helps in the selection of expansion terms.

7 Conclusion

This study tested the semantic relatedness measure when selecting the terms to expand queries. Key components of this measure are the ESA measure, the WordNet path measure, and the mixed collocation index. Results produced by the Terrier search engine were a base line in the experiments.

Term candidates for the expansion were also generated by Terrier. The proposed techniques were applied to the ad-hoc retrieval task. As a data set, the NTCIR-1 CLIR Test collection was used. The initial English queries were obtained automatically applying Google translate. The queries were expanded by applying the Wikipedia-based Explicit Semantic Analysis measure, and the DFR mechanism, and the semantic relatedness measure. The retrieval results showed superiority of the last one over ESA and DFR.

A promising new direction to enhance the EWC measure is to take into account Wikipedia articles with titles consisting of multiple terms in order to get knowledge about scientific terminology and general purpose terminology. The expected outcome of this is the more precise term selection to expand the user queries submitted to search engines.

References

[1] Terrier (2011). [Online document], http://terrier.net

[2] NTCIR-1 CLIR data collection (1999). [Online document], http://research.nii.ac.jp/ntcir/data/data-en.html

[3] TREC (2011). [Online document], http://trec.nist.gov/

[4] Google Translate (2011). [Online document], http://translate.google.com/

[5] Ben He and Iadh Ounis (2009). Studying Query Expansion Effectiveness. In Proc. *The 31st European Conference on Information Retrieval (ECIR09)*. Toulouse, France.

[6] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne (1995). Okapi at TREC-4. In Proc. *TREC 4*.

[7] Aitao Chen, Fredric C. Gey, Kazuaki Kishida, Hailing Jiang and Qun Liang (1999). Comparing Multiple Methods for Japanese and Japanese-English Text Retrieval, NTCIR Workshop 1. In Proc. *The First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition,* Tokyo.

[8] G. Amati and C.J. Van Rijsbergen (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

[9] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (2008). *Introduction to Information Retrieval*, Cambridge University Press.

[10] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich (2011). Concept-Based Information Retrieval using Explicit Semantic Analysis, *ACM Transactions on Information Systems*, 29(2).

[11] Philipp Sorg and Philipp Cimiano (2009). An Experimental Comparison of Explicit Semantic Analysis Implementations for Cross-Language Retrieval. In Proc. *The International Conference on Applications of Natural Language to Information Systems (NLDB)*, Saarbrücken.

[12] Yinghao Li, Wing Pong Robert Luk, Kei Shiu Edward Ho, and Fu Lai Korris Chung (2007). Improving weak ad-hoc queries using Wikipedia as external corpus. In Proc. *The 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, pp 797 - 798.

[13] Hamada M.Zahera, Gamal F. El Hady, and Waiel.F Abd El-Wahed (2010). Query Recommendation for Improving Search Engine Results. In Proc. *The World Congress on Engineering and Computer Science 2010* Vol I, WCECS 2010, San Francisco, USA.

[14] Jose R. Perez-Aguera1 and Lourdes Araujo (2008). Comparing and Combining Methods for Automatic Query Expansion. *Advances in Natural Language Processing and Applications Research in Computing Science 33*, pp. 177-188.

[15] Ming-hung Hsu, Ming-feng Tsai, and Hsin-hsi Chen (2008). Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach. In Proc. *Asia Information Retrieval Symposium*, pp. 213-224.

[16] Hang Cui1, Ji-Rong Wen, Jian-Yun Nie3, and Wei-Ying Ma (2002). Probabilistic query expansion using query logs. In Proc. *The 11th international conference on World Wide Web*, ACM New York, NY, USA.

[17] J. Malecka and V. Rozinajova (2006). An Approach to Semantic Query Expansion. In Proc. *Tools for Acquisition, Organization and Presenting of Information and Knowledge, Research Project Workshop*, Bystra dolina, Tatry, pp. 148-153.

[18] Jiuling Zhang, Beixing Deng, and Xing Li (2009). Concept Based Query Expansion Using WordNet. In Proc. *The 2009 International e-Conference on Advanced Science and Technology*, pp. 52-55.

[19] M. Ellen Voorhees (1994). Query expansion using lexical-semantic relations. In Proc. *The 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*.

[20] D. Gayo-Avello Brenes (2009). Stratifies analysis of AOL query log. *Information Sciences*, 179, pp. 1844-1858.

[21] Yannis Haralambous and Vitaly Klyuev (2011). A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge. In the Proc. *IJCNLP*.

[22] Vitaly Klyuev, Ai Yokoyama (2010). Web Query Expansion: A Strategy Utilizing Japanese WordNet. *Journal of Convergence*, V. 1, Number 1.

[23] Space ALC (2011). [Online document], http://www.alc.co.jp/

[24] Mecab (2011). [Online document], http://mecab.sourceforge.net/

[25] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato and Soichiro Hidaka (1999). Overview of IR tasks. In Proc. *The First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, Tokyo.