Query Phrase Expansion Using Wikipedia in Patent Class Search

Conference Paper · December 2011

DOI: 10.1007/978-3-642-25631-8_11 · Source: DBLP

Bashar Al-Shboul
University of Jordan
35 PUBLICATIONS **341** CITATIONS
SEE PROFILE

Sung-Hyon Myaeng
Korea Advanced Institute of Science and Technology
212 PUBLICATIONS **3,095** CITATIONS
SEE PROFILE

Query Phrase Expansion Using Wikipedia in Patent Class Search

Bashar Al-Shboul and Sung-Hyon Myaeng

Korea Advanced Institute of Science and Technology
373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, South Korea
{bashar,myaeng}@kaist.ac.kr

Abstract. Relevance Feedback methods generally suffer from topic drift caused by words ambiguity and synonymous uses of words. As a way to alleviate the inherent problem, we propose a novel query phrase expansion approach utilizing semantic annotations in Wikipedia pages, trying to enrich queries with context disambiguating phrases. Focusing on the patent domain, especially on patent search where patents are classified into a hierarchy of categories, we attempt to understand the roles of phrases and words in query expansion in determining the relevance of documents and examine their contributions to alleviating the query drift problem. Our approach is compared against Relevance Model, a state-of-the-art, to show its superiority in terms of MAP on all levels of the classification hierarchy.

Keywords: Pseudo-Relevance Feedback, Patent Information Retrieval, Wikipedia Categories, Query Expansion, Phrase-based Query Expansion.

1 Introduction

Query Expansion (QE) is one of the Information Retrieval (IR) techniques to enhance effectiveness of document retrieval. It is generally used to disambiguate the context of a user query. One of the heavily researched approaches is Pseudo-Relevance Feedback (PRF): an automatic query expansion method based on the assumption that top ranked documents retrieved for a query are the most relevant ones. While the terms in top-ranked documents are considered the best resource for selecting expansion terms, past research shows that PRF-like models suffer from several drawbacks such as query-topic drift [6], [7] and inefficiency [5]. There have been attempts to use lexical resources, most notably WordNet, for QE. As WordNet has been exploited in various IR tasks including QE [10], [13], [14], a common conclusion resulting from its limited coverage of words and relations, and the lack of contextual information for each word, is low effectiveness [11]. While words' ambiguity is a main reason behind using QE to disambiguate the query context, phrases can play the same role because the surrounding word(s) in a phrase provide additional contextual information. When phrases were used alongside with words for IR, however, the results have been disappointing with only a slight improvement or even a decrease in effectiveness [16]. A reason is that phrases have different distributions over documents when compared to words [15]. Motivated by the topic

M.V.M. Salem et al. (Eds.): AIRS 2011, LNCS 7097, pp. 115–126, 2011.

drift problem in PRF, inability to improve retrieval effectiveness with automatically identified phrases, and the limitations of using WordNet for word-based QE [10],[17], this paper proposes a novel QE approach utilizing Wikipedia semantic annotations (i.e. categories) for query phrase expansion. Our approach concentrates on reducing query topic drift using both WordNet and Wikipedia, to handle word synonyms and concurrently compensate the limitation of WordNet by enriching queries with phrases to disambiguate query context.

Wikipedia is an online collaborative contribution of the Web community to building an encyclopedia. Generally, it was utilized for IR-related tasks including Word-Sense Disambiguation (WSD) [9], Named-Entity Retrieval [24], Document Clustering/Classification [20], Question Answering [21], and QE [24], [28]. In Wikipedia, a page describing concept expressed as a word or phrase belongs to one or more categories, each of which contains a category title (i.e. Information Retrieval), titles of the pages in this category (i.e. Discount Cumulative Gain, Generalized Vector Space Model, etc.), and other related categories (i.e. Information Science). Our work differs from all other works in its utilization of individual categories, page titles under each category, and links to other related categories.

We test our method on patent search, which has arisen as one of the important information retrieval fields, especially for legal IR [1], [8]. As current patent search systems use a keyword-based approach, effectiveness of retrieval relies on the quality of search keywords [2]. Patents are particularly suitable for testing our method because they usually contain a large number of phrases because they often deal with technical vocabulary. Aside from several unique characteristics such as vocabulary, usage, and structure, patent search is unique in that each patent is manually assigned to one or more classes from the IPC (International Patent Classifications). A group of patents belonging to a class are said to be relevant to each other. IPC has a hierarchy of three levels, Sub Class (SC), Main Group (MG), and Sub Group (SG), with SC being most general. For example, when a patent is assigned to the IPC “G06F 17/30”, its SC, MG, and SG are “G06F”, “G06F 17”, and “G06F 17/30”, respectively. The IPC hierarchy allows us to study generalization/specialization capabilities of phrases as our retrieval task is to classify patents with or without query expansion. Our proposed query expansion approach has been tested on US Patent & Trademark Office (USPTO) patents provided by NTCIR.

In our experiment, comparisons are made against Relevance Model (RM) [4] because it has often been used as a comparison benchmark [3], [4]. Among many PRF-based efforts [4], [5], [6], [11], RM has drawn much attention with its strong probabilistic ground. Details of RM are omitted for brevity.

The contributions of this work are: (1) a thorough study of the effect of adding phrases to the baseline and RM in ranking documents. (2) an evaluation of the effect of RM and WordNet for word-based query expansion, their roles in generalizing/specifying the query topic, and their relevance to the query topic. (3) an exploration of the effect of utilizing Wikipedia for query phrase expansion and its role in generalizing/specifying the query topic, and (4) an analysis of the interaction between WordNet-based expanded words and Wikipedia-based expanded phrases, in addition to studying their effects on ranking documents and then finding an optimal weight ratio between them for achieving the best retrieval effectiveness.

This paper is organized as follows. In section 2, related work is presented. Our proposed framework is described in section 3. Experiment goals, environment and results are discussed in section 4. Finally, we conclude in section 5.

2 Related Work

Among the several automatic QE approaches, PRF has shown its effect in improving retrieval effectiveness [19]. While several attempts to enhance PRF were reported [12], [18], [4], others tried different approaches (i.e. expansion based on query characteristics [22], mining user logs [25], using external resources [26]). QE based on local collection statistics may fail due to the lack of relevant documents for a query, necessitating external collections were used for query enrichment [26]. In this work, we follow the strategy of using external resources as a way of enriching and disambiguating queries with relevant words and phrases.

Relevance feedback models have been proposed as query expansion methods. However, it is well known that it suffers from *topic drift* [6], [7], especially in short queries. Major causes are the ambiguity of query terms and the method adopted by PRF in weighting and selecting query candidate expansion terms (i.e. IDF). We attempt to alleviate these problems by expanding query phrases using external resources rather than collection-dependent phrases. RM was adopted in Lemur IR Toolkit, which we used to implement and test our model. Generally, in Lemur toolkit an RM query takes the form of *#weight(w_1 Baseline_Query w_2 Expansion_Terms),* where w_1 and w_2 are normalized weights, set to 0.5 by default. This way of scoring seems unfair considering that expansion terms are not guaranteed to be relevant to the query topic in the first place. In this paper, we try to find the optimal weight balance between the baseline query and the expansion terms for producing the best effectiveness and then compare our proposed model with the best possible results obtained from RM.

WordNet has been utilized for QE in different ways. For example, WordNet semantic relations have been used to solve the problem of vocabulary mismatch [10]. In [17], the query words were expanded separately by intersecting each word's synsets sharing a lexical relation in different resources. An interesting result shows that using WordNet synsets in addition to *"glosses"* as expansion words contributed significantly towards effectiveness. We compare our model with best reported results in [17].

Wikipedia has been utilized for QE in different ways. A link-based expansion approach is [28], where PRF is modified to rank links based on their target documents scores, and then use link texts as expansion terms. RM is then used for another round of expansion. This approach has shown a significant improvement over RM on blog documents. Our work is different from the state of the art works in that we utilize different parts from Wikipedia (i.e. categories, category links to relevant categories, and page titles under category pages), trying to expand different type of query terms (i.e. Phrases).

Several types of phrases (i.e. noun phrases, head/modifier, bigrams, and others) have been used for different IR applications [23], [3], [27], but very rarely for query expansion. In [3], for example, semantic information extracted from DBpedia is utilized. A phrase is run against DBpedia index using cosine similarity. SKOS (Simple Knowledge Organization System) of top ranked documents are used as thesauri to expand phrases. Additionally, all synonyms extracted from WordNet are added as expansion terms to the query used for research article classification over the IPC hierarchy. Our work shares the same idea of using an external resource for phrase-based QE but proposes a new method with a different resource, i.e. Wikipedia, and examines various options with a greater depth.

3 Proposed Method

Our method concentrates on providing contextual information to the query and avoiding vocabulary mismatches. Query words are expanded considering the most likely synset returned from searching WordNet, where synsets' ranking is based on word frequencies in the British National Corpus. In addition, phrases are also expanded as they contain contextual information for individual query words. Phrases are used to search the Wikipedia index to extract a set of Wikipedia categories (primary categories), their related categories (secondary categories), and the Wikipedia pages belonging to the primary categories, which are later processed to generate a set of candidate expansion phrases. Merging the original query with the candidate expansion words and phrases generates our expanded query.

3.1 Page Similarity

In Wikipedia, each titled page belongs to one or more categories. In computing page similarities, we consider pages belonging to the same category are similar to each other, instead of using the text in them. In addition, we use the related (secondary) categories linked to a category page of the primary category. Two categories with different titles can be seen similar to each other when they share related categories. For example, two pages whose categories are "Green Automobiles" and "Green Vehicles" are considered relevant to each other when the corresponding category pages share some secondary category names. However, their similarity is not as strong as the pages belonging to the same category. Further, page titles can be used for similarity calculation when the pages do not match in terms of their primary and secondary category names. We consider two pages are similar to some extent when there is an overlap between the titles of the pages belonging to the primary and secondary categories of the original pages being compared. For example, while "Precision & Recall" page and "Accuracy & Precision" page do not share any primary and secondary categories, they share several page titles belonging to their categories. Similarity between two Wikipedia pages is computed as follows:

$$Sim\left(P_1, P_2\right) = 0.5 \times PPC_{P_1,P_2} + 0.3 \times PSC_{P_1,P_2} + 0.2 \times PT_{P_1,P_2} \tag{1}$$

Where PPC is percentage of primary category phrases match to all distinct primary categories, PSC is percentage of secondary category phrases match to all distinct secondary categories, and PT is percentage of page title phrases match to all distinct page titles. Phrases considered matching iff an exact string match exists. Weights assigned arbitrarily upon our assumption that PPC is the most important among all, followed by PSC then PT. Weight optimization is left as future work. For this computation, each Wikipedia page is indexed with primary and secondary categories and page titles of the categories. Indexing in this way will allow us to estimate Maximum Likelihood Estimate (MLE) of each query phrase under language models of each field (i.e. primary category field, secondary categories field, category page titles field), as will be explained later.

3.2 Query Term Extraction and Expansion

In patent search, terms are extracted from a query patent to form a search query. Key terms are extracted from different parts of query patents to see their roles in experiments as will be explained in section 4. We first apply a stop word filter that uses a customized list collected from different sources (i.e. KEA, InQuery, Lemur). We then apply Stanford POS Tagger to the resulting text and Regular Expressions (RegEx) to extract keywords and keyphrases. This process is particularly useful for short fields like titles and abstract as the brevity might not be enough for statistics-based methods (e.g. TFIDF, CHI, IG) of feature selection. Nouns, verbs and adjectives are extracted as keywords, and then phrases using RegEx. Further, all unigrams that are covered by a phrase are removed, as they are already disambiguated by additional contextual information in the phrase. The RegEx used to extract keyphrases is given by: (VBG|VBN|JJ|JJR|JJS)(NN|NNS|NNP|NNPS)+ where VBG and VBN are verbs, JJ, JJR and JJS represent adjectives, and NN, NNS, NNP and NNPS represent nouns and proper nouns in singular and plural forms. This RegEx was built based on our observations over the tagged query patents. After generating a query by extracting key terms from a query patent document, it is split into a bag of words (BOW) and a bag of phrases (BOP). Each entry in BOP and in BOW is expanded using Wikipedia and WordNet, respectively. A query phrase is expanded in order to alleviate any vocabulary mismatch by, for example, finding an alternative name for a technology (e.g. “speaker identification” and “speaker verification”). Often times, the same technology can be expressed in different phrases. A phrase is expanded using the Wikipedia index. To retrieve Wikipedia pages, we employ the following model:

$$P\left(ph_j|D\right) = \sum_{i\in G}\lambda_{\mathrm{i}}\ \mathrm{p}(\mathrm{ph}_j \mid \theta_{i,j}) \tag{2}$$

where G = {Primary Categories, Secondary Categories, Titles of the pages under Primary and Secondary Categories} represents three background language models, and $\sum_{i\in G}\lambda_i = 1$ are the mixture weights which were empirically set to 0.5, 0.3, and 0.2, respectively. Additionally, language models are estimated as:

$$p(ph_j \mid \theta_{i,j}) = \frac{c(ph_j,\ X_i)}{c(X_i)} \tag{3}$$

where *c(ph$_j$, X_i)* is the count of phrase *ph$_j$* in the X_i field of the index (with cosine similarity higher than 0.7), and *c(X_i)* is the count of distinct categories and page titles in the LMs. Further, top ranked 5 documents categories and titles are intersected to find common phrases. Phrases exist in 3 or more documents are added to the query. Finally, another cycle of word filtering is done to remove words covered by phrases trying to avoid duplicates, considering that those words are already disambiguated through a phrase or more.

In our method, every query word is expanded, where the first synset found in WordNet is considered as an expansion candidate term. That is because different feature selection approachs (i.e. Chi, IG, DF,..etc) have shown different weaknesses in weighting terms (i.e. preference of terms with specific characteristics). For example, a statistical model such as IDF weights the term based on its discriminating power assuming that terms with low document frequency are more discriminating than others. To alleviate such an issue, query words are expanded separately aiming to reduce the effect of terms with higher weights over others with low weights, granting the chance for all query words to be expand regardless of their distribution over the corpus. This method of word expansion was followed hoping that WordNet expansion words will help in the case of vocabulary mismatch between the query and the collection. Furthermore, in patent domain IPC skewness is a very serious problem for patent search tasks (i.e. State-of-Art). In our indexed USPTO collection, less than 50 IPCs embrace more than 13% of the documents, showing a long-tailed distribution where term sparseness for the majority of IPCs is the major problem.

3.3 Patent Retrieval

Using the expanded queries generated by our model, they were run against patent index using Okapi BM25 with default variables set in Lemur where, k_1 and b are BM25 variables set to *1.2* and *0.75* respectively, while *IDF* is described in [29].

After a ranked list of patents R_Q is returned from a query, we re-rank them using the IPC information associated with them. The main idea is that if a retrieved patent belongs to the IPC to which the query patent belongs, it should be considered relevant to the query. A re-ranking algorithm should consider the fact that a search result is likely to contain multiple documents belonging to the same IPC and that each document may have multiple IPCs itself. Our re-ranking process is divided into three successive stages: *IPC Mapping*, *IPC Expansion*, and *IPC Scoring*. The process of generating a new list of documents' IPC(s) with their scores is called *IPC Mapping*. Next, patents with multiple IPCs are expanded into several entries holding the same score, and rank, as the whole patent. This stage is called *IPC Expansion*. The expanded list is denoted as R_{IPC}. Final stage is re-scoring IPCs as an IPC will probably occur several times in R_{IPC}. For that purpose list L_Q= {*distinct IPCs in R_{IPC} for query Q*}. IPCs are re-scored as follows:

$$Score\left(ipc\right) = \frac{\sum_{ipc \in L_Q} rank\left(ipc, R_{IPC}\right)}{count(ipc, R_{IPC})} \tag{4}$$

In this case, IPCs with higher frequency on higher scores in R_{IPC} are favored as they are assumed to be highly relevant compared to other IPCs. IPCs are further re-ranked based on their new scores.

4 Experiments

Our experiments aim at answering the following series of research questions:

- What is the effect of using phrases in the baseline query in addition to words?
- What are the best weights assigned to baseline and expansion parts in RM query?
- What is the effect of query word expansion using WordNet?
- What is the effect of using Wikipedia for query phrase expansion?
- What is the effect of combining both Wikipedia and WordNet expansions?
- What is the best weight balance between words and phrases in our expansion model?

For our experiments, we used Wikipedia Dump of August 2010, in addition to Indri, which was used for indexing Wikipedia articles and NTCIR-6 USPTO patents of 1993-2002 (~1.3M documents). Queries were patents selected randomly from those contained in the IPCs having at least two patents. The query patents were then removed from the Index. A total of 1,780 patents were selected as the queries for experiments. Relevance judgments of the queries were made based on IPCv9 crawled from the USPTO website. Query patents were mapped to their corresponding IPCv9 to generate the relevance judgment. In all our experiments, we used Okapi BM25 as the retrieval model for the combined query of words, phrases, and their expansions in our model. To determine which parts of a patent will generate the best query terms *preliminary experiments* with various combinations of patent parts were undertaken to find that a combination of Titles and Abstracts was the best. Thus all the results reported in this paper are for queries extracted from those two parts. During experiments, Precision and Recall were used to evaluate our work; however, depending on IPCs instead of documents. They are given as follows:

$$Precision(Q) = \frac{Relevant\_Retrieved\_IPCs(Q)}{All\_Retrieved\_IPCs(Q)} \tag{5}$$

$$Recall = \frac{Distinct\_Relevant\_Retrieved\_IPCs}{All\_Relevant\_IPCs} \tag{6}$$

where Q is Query, Relevant_Retrieved_IPCs represents the number of documents with relevant IPCs in R_Q, All_Retrieved_IPCs is set to number of top ranked documents (i.e. 1000 in our case), Distinct_Relevant_Ret-rieved_IPCs is the number of distinct IPCs correctly retrieved, while All_Relevant_IPCs represents number of IPCs assigned to the query patent. Same precision was used in evaluating NTCIR Patent Mining tasks; however, Recall was modified to evaluate based on IPCs. The number of relevant IPCs for each classification level is given beside the classification level notation (i.e. "SC (3155)" indicates that at Sub-Class level, there are 3,155 relevant IPCs for the whole set of 1,780 query patents used for experiments)

4.1 Baseline Queries

In our experiments two different baseline queries were used: unigram baseline queries and the unigram baseline queries plus phrases after deleting the unigrams involved in the phrases (denoted as “word/phrase” in Table 1). Table 1 shows a significant decrement in MAP (Mean Average Precision) with the word/phrase queries. This can be attributed to the generality of added phrases, while unigrams succeeded to rank relevant documents higher than word/phrase did. However, the recall drop was less severe, meaning that phrases disambiguated the context almost similarly to unigrams.

Table 1. A comparison between the baseline, RM, and our method (OM)

Query	SC (3155)		MG (3801)		SG (5391)	
	MAP	Recall	MAP	Recall	MAP	Recall
Unigram	39.19	84.40	23.99	80.03	15.64	72.71
Word/Phrases	36.75	84.12	22.15	78.19	13.32	70.69
OM Wikipedia	33.97	84.21	20.9	78.37	12.23	68.72
Wiki-Links	50.57	77.65	37.94	73.30	22.06	78.52
OM WordNet	40.19	84.43	24.66	80.53	15.84	75.71
WN-Gloss	22.39	82.12	12.95	75.24	9.25	61.07
RM	52.99	83.99	36.07	81.76	22.86	87.01
Our Model	54.34	84.97	38.37	82.92	25.82	86.95

4.2 Relevance Model

In our work, as in Lemur, RM expanded queries generally follow a weighting structure that provides 50% of the weight to the original query terms, and the other 50% to the expansion terms. To tune weights for RM queries, a *preliminary set of experiments* for variants of weights have been undertaken, and 0.6/0.4 was selected for Baseline/Relevance parts of the queries as they generated the best result (Experiment results were not listed in the paper due to page limit). As can be seen in Table 1, our model gave significantly better MAP values compared to RM but only slightly better results in recall. With further analysis it was found that RM performance worsens as participation from RM expansion part in scoring documents increases (i.e. 78% or more of query weight determined by RM expansion part). However, performance enhanced after weight participation decreased to less than 77%.

4.3 WordNet

In our method (an experiment of expanding baseline query using WordNet, and skipping Wikipedia expansion part), 76% of words added by WordNet (representing more than 23% of all query words) had very high IDF; however, many of them were irrelevant to the patent topic. Further analysis revealed that WordNet expansion terms exist twice as many in the irrelevant retrieved documents as in the relevant ones. Those terms had the highest weights in the query and affected the search negatively, causing the query topic to drift. Table 1 also shows that queries consisting of words and their WordNet expansions work slightly better than those phrase queries at the general level (SC) but much better at a specific retrieval level (SG Level). An interesting finding is that in each WordNet expanded query there are in average 8 words overlapped with phrases in the title-abstract queries, representing 26% of average query length, and having a good chance covering general query topic. Furthermore, our method of WordNet expansion was compared to the WordNet expansion using glosses described in [17], denoted as WN-Gloss in Table 1. The result shows that QE using WordNet words from the first synset retrieves a better result than that of using WordNet glosses.

4.4 Wikipedia

In this section, two different sets of experiments were performed. First, to understand the effect of our proposed idea of Wikipedia expansion on query performance, this expansion has been performed on the baseline queries (annotated as OM Wikipedia). Even though this experiment's MAP result was the second worst amongst all, it is important to note its recall result, which is only slightly lower than those of most other cases at the general level of IPC, and worsened at more specific classification levels. Examining Wikipedia expansion phrases, it was found that these results can be attributed to one (or more) of the following reasons:

- Wikipedia titling policy which requires that category titles are topic descriptors specific enough to be distinguished from each other but general enough to cover more specific concepts in each page.
- The number of relevant/irrelevant expansion phrases. As in retrieval process all phrases has exactly the same weight. This weighting policy was basically followed aiming not to judge a phrase based on its existence in Wikipedia index because it might be a new or uncommon phrase. On the other hand, the equal weighting policy might seem have helped irrelevant phrases to drift the query topic; however, this seems to be a low possibility considering that relevant phrases appear, in average, more frequently in relevant documents (i.e. In average, 2.7 expansion phrases exist in relevant documents compared to 2.1 phrases in irrelevant ones)
- Wikipedia categories can be unreliable sometimes because it is possible to arrive at complete different sets of categories from similar pages (i.e. Wikipedia pages "precision and recall" and "accuracy and precision").

Furthermore, our model was compared to another state-of-the-art method of utilizing Wikipedia for query expansion [14]. As can be seen, the state-of-the-art model (annotated as Wiki-Links in Table 1) performed better in terms of MAP than our proposed method of Wikipedia expansion (without WordNet part) at all the classification levels and also better in recall at the SG classification level. However, it performed worse in recall at the more general classification levels (i.e. SC, MG). This indicates that expansion phrases added by our method are more general than terms added from hyperlinked texts, because they performed better on recall over higher classification levels; however, they are not enough for retrieving relevant documents higher in rank as can be concluded from MAP results. Note that the Wiki-Links based query expansion performed comparably to RM. This indicates that selected hyperlinked texts contain good expansion terms comparable to the ones selected using RM.

4.5 Proposed Method

Our proposed method depends on the interaction between Wikipedia and WordNet expansions. All words and phrases were considered for expansion under the hypothesis that phrases provide good contextual information and hence valuable assistance to context disambiguation. Words and phrases were merged together in order to generate a more coherent, and a less ambiguous query than a RM expanded query. Our analysis reveals that phrases play an important role of promoting relevant documents in the ranked list since IDF values for phrases are usually higher than those of the words. It is worth mentioning that baseline queries consisted on average of 31 terms, and after applying our expansion method, the number of terms was increased to become 51. As shown in Table 1, our method performed better than RM at all levels of classification. As can be seen, the combination of WordNet-expanded words, and Wikipedia-based expanded phrases exploited the best of both to achieve the best MAP and recall. As in Table 2, our method shows significantly better precision@N, using a t-test at p = 0.05, especially at the specific classification levels where precision is significantly higher.

Table 2. Precision@N, Our Method (OM) vs. Relevance Model (RM)

	SC		MG		SG	
	RM	**OM**	**RM**	**OM**	**RM**	**OM**
P@5	18.81	19.84	15.21	16.88	14.29	16.21
P@10	11.23	11.74	9.74	10.64	9.81	11.16

4.6 Phrase Weight Balance

As our proposed method performed the best amongst all other cases, it seems important to find the best weight balance between words and phrases in the expansion method. A *preliminary experiment with a subset of queries* showed that fixing phrase weights twice as big as that of words gave the best results amongst all other combinations. However, giving higher weights to phrases over a certain limit (i.e. twice as much in our experiment) will generalize the query more, and decrease effectiveness as well (Experiment detailed results were not listed in the paper due to page limit). The same experiment was performed on RM by modifying phrase weights in the baseline part of the expanded query; however, the results shows that the higher weight assigned to phrases in the baseline part worsens the results in terms of both MAP and Recall. The reason is that RM expansion part still dominates the weighting of the query, with minor participation from the baseline part. Furthermore, increasing the weights of phrases in the baseline part decreases their weighting effect severely.

5 Conclusions and Future Works

We have proposed a new method of using Wikipedia categories and WordNet at the same time for query word and phrase expansion in the task of search for patents. In a series of experiments using IPC categories and USPTO patents, our proposed method has shown the usefulness of expanding query phrases with Wikipedia categories of two kinds and titles, when they are used together with expanded words. Our analysis of the experimental results reveals that added phrases control topic drift caused by PRF. Our future work includes incorporating the Wiki-Links idea to our proposed expansion method, an IPC re-ranking model is being researched, as well as devising a more accurate method for keywords expansion by using a more accurate synset recommender method, so that we can pinpoint where the expanded phrases play an essential role.

Acknowledgement. This research was partially supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-C1090-1111-0008).

References

1. Azzopardi, L., Vanderbauwhede, W., Joho, H.: Search system requirements of patent analysts. In: Proc. of SIGIR 2010 (2010)
2. Xue, X., Croft, W.B.: Transforming patents into prior-art queries. In: Proc. of SIGIR 2009 (2009)
3. Al-Shboul, B., Myaeng, S.H.: IRNLP@KAIST in the subtask of Research Papers Classification in NTCIR-8. In: Proc. of NTCIR-8 (2010)
4. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proc. of SIGIR 2001 (2001)
5. Yin, Z., Shokouhi, M., Craswell, N.: Query Expansion Using External Evidence. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 362–374. Springer, Heidelberg (2009)
6. Lang, H., Metzler, D., Wang, B., Li, J.T.: Improved latent concept expansion using hierarchical markov random fields. In: Proc. of CIKM 2010 (2010)
7. Lv, Y., Zhai, C.: Adaptive relevance feedback in information retrieval. In: Proc. of CIKM 2009 (2009)
8. Maxwell, K., Schafer, B.: Concept and Context in Legal Information Retrieval. In: Proc. of JURIX 2008 (2008)
9. Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2), Article 10 (2009)
10. Voorhees, E.: Query expansion using lexical-semantic relations. In: Proc. of SIGIR 1994 (1994)
11. Bai, J., Nie, J.Y.: Adapting information retrieval to query contexts. Inf. Process. Manage. 44(6), 1901–1922 (2008)
12. Lee, K., Croft, B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: Proc. of SIGIR 2008 (2008)

13. Vechtomova, O., Karamuftuoglu, M., Robertson, S.: On document relevance and lexical cohesion between query terms. Information Processing & Management 42(5), 1230–1247 (2006)
14. Vechtomova, O., Karamuftuoglu, M.: Query expansion with terms selected using lexical cohesion analysis of documents. Information Processing & Management 43(4), 849–865 (2007)
15. Lewis, D., Croft, B.: Term clustering of syntactic phrases. In: Proc. of SIGIR 1990 (1989, 1990)
16. Koster, C., Beney, J.: Phrase-based document categorization revisited. In: Proc. of PaIR 2009 (2009)
17. Navigli, R., Velardi, P.: An analysis of ontology-based query expansion strategies, workshop on adaptive text extraction and mining (ATEM 2003). In: 14th European Conference on Machine Learning (ECML 2003) (2003)
18. Cao, G., Nie, J.Y., Gao, J., Robertson, S.: Selecting good expansion terms for pseudo-relevance feedback. In: Proc. of SIGIR 2008 (2008)
19. Xu, J., Croft, B.: Query expansion using local and global document analysis. In: Proc. of SIGIR 1996 (1996)
20. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using Wikipedia. In: Proc. of SIGIR 2007 (2007)
21. Ganesh, S., Varma, V.: Exploiting structure and content of Wikipedia for Query Expansion in the context of Question Answering. In: Recent Advances in Natural Language Processing (RANLP 2009), Bulgaria (2009)
22. Xu, Y., Jones, G., Wang, B.: Query dependent pseudo-relevance feedback based on Wikipedia. In: Proc. of SIGIR 2009 (2009)
23. Kapalavayi, N., Murthy, S., Hu, G.: Document classification efficiency of phrase-based techniques. In: IEEE/ACS International Conference on Computer Systems and Applications (2009)
24. Li, Y., Luk, W., Ho, K., Chung, F.: Improving weak ad-hoc queries using Wikipedia as external corpus. In: Proc. of SIGIR 2007 (2007)
25. Cui, H., Wen, J., Nie, J., Ma, W.: Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering 15(4), 829–839 (2003)
26. Kwok, K., Chan, M.: Improving two-stage ad-hoc retrieval for short queries. In: Proc. of SIGIR 1998 (1998)
27. Arampatzis, A., Tsoris, T., Koster, C., Van Der Weide, T.: Phrase-based information retrieval. Information Processing & Management 34(6), 693–707 (1998)
28. Arguello, J., Elsas, J.L., Callan, J., Carbonell. J.G.: Document Representation and Query Expansion Models for Blog Recommendation. In: Proc. of ICWSM 2008 (2008)
29. Robertson, S., Jones, K.: Relevance weighting of search terms. Journal of the American Society for Information Science 27, 129–146 (1976)