Inf Retrieval (2014) 17:430–451
DOI 10.1007/s10791-013-9233-4

Wikipedia-based query phrase expansion in patent class search

Bashar Al-Shboul · Sung-Hyon Myaeng

Received: 23 March 2013 / Accepted: 16 September 2013 / Published online: 21 November 2013

Abstract Relevance feedback methods generally suffer from topic drift caused by word ambiguities and synonymous uses of words. Topic drift is an important issue in patent information retrieval as people tend to use different expressions describing similar concepts causing low precision and recall at the same time. Furthermore, failing to retrieve relevant patents to an application during the examination process may cause legal problems caused by granting an existing invention. A possible cause of topic drift is utilizing a relevance feedback-based search method. 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 phrases disambiguating the original query words. The idea was implemented for patent search where patents are classified into a hierarchy of categories, and the analyses of the experimental results showed not only the positive roles of phrases and words in retrieving additional relevant documents through query expansion but also their contributions to alleviating the query drift problem. More specifically, our query expansion method was compared against relevance-based language model, a state-of-the-art query expansion method, to show its superiority in terms of MAP on all levels of the classification hierarchy.

Keywords Patent search · Phrase-based query expansion · Wikipedia categories · Clarity · Retrievability

B. Al-Shboul
Department of Business Information Technology, The University of Jordan, Al-jubaiha, Amman 11942, Jordan
e-mail: bashar.shboul@gmail.com

S.-H. Myaeng (✉)
Department of Computer Science and Division of Web Science, Korea Advanced Institute of Science and Technology, 291 Daehak-ro (373-1 Guseong-dong), Yuseong-gu, Daejeon 305-701, South Korea
e-mail: myaeng@kaist.ac.kr

1 Introduction

Patent search addresses the task of retrieving relevant yet highly specific patent documents. Patent documents are generally difficult to process due to the variations of writing styles within different patent sections (i.e. abstract, claims, description) and among different patents (Atkinson 2008). Difficulty is exacerbated by two factors: (1) authors who intentionally write patents to obstruct their retrieval (Azzopardi et al. 2010) and (2) searchers (i.e. examiner, intellectual property office agent, etc.) who are in need of exhaustive knowledge of all the patents relevant to the patent at hand. Given that patent search is a sensitive application where overlooking a relevant patent will certainly cause legally expensive consequences (Jochim et al. 2011), there is a great demand to develop methods to overcome the difficulties caused by the variations. While several different approaches are possible to alleviate the problem, we focus on the methods by which search queries are enriched with contextually-relevant terms (e.g. synonymous words and phrases) obtained from local (Lavrenko and Croft 2001) or external resources (Bendersky et al. 2011; Al-Shboul and Myaeng 2011). In addition, we focus on the patent search task where the query patent is given with its semantic categories (i.e. IPC Codes) and other patents in the same categories are retrieved (Nanba et al. 2010), so that we can minimize the possibility of losing the semantically relevant patents that may look different on the surface.

Query expansion (QE) is one of the well-known techniques to enhance effectiveness in information retrieval (IR) and plays the role of disambiguating the context of a user query. Among several QE approaches, pseudo-relevance feedback (PRF) has been heavily studied because of its effectiveness without human intervention. It is 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 (Lang et al. 2010; Lv and Zhai 2009), especially in short queries, and inefficiency (Yin et al. 2009). While a patent document used as a query is long in patent search, a different type of topic drift would occur because of the way patents are written; newly filed patent applications often use terms that are different from those used in the previously granted patents.

On the other hand, there have been attempts to use lexical resources, most notably WordNet,[1] for QE. It is a complementary approach in that QE based on local collection statistics may fail due to the lack of relevant documents for a query, justifying the use of external collections for query enrichment (Kwok and Chan 1998). While WordNet has been exploited for various IR tasks including QE (Voorhees 1994; Vechtomova and Karamuftuoglu 2006, 2007; Magdy and Jones 2011), a common conclusion is that it seldom contributes to enhancement of retrieval effectiveness (Bai and Nie 2008), primarily because of its limited coverage of words and relations and the lack of contextual information for each word. Linguistically motivated lexical resources like WordNet fall short of the depth required for patent search QE. Even with richer external resources, the word sense disambiguation problem would remain when QE is based on words.

While adding terms to the original query with QE can have an effect of query context disambiguation and query enrichment, phrases containing a query term can play the same role because the surrounding word(s) in such a phrase provide additional contextual information. When phrases were used alongside with words for IR in general, however, the results were disappointing with only a slight improvement or even a decrease in

[1] http://wordnet.princeton.edu/.

effectiveness (Koster and Beney 2009; Koster et al. 2011). A reason is that phrases have different distributions over documents when compared to words (Lewis and Croft 1990) requiring different methods for estimating phrase probabilities. On the other hand, the use of noun phrases for expansion has been shown effective (Mahdabi et al. 2012). For patent search, however, handling phrases are particularly important because patent documents use phrases heavily due to their technicality. Unlike web search, patent documents used as queries in patent search contain quite a number of phrases that can be leveraged.

Motivated by the topic drift problem in PRF, the limitations of using WordNet for word-based QE (Voorhees 1994; Navigli and Velardi 2003) and the inability to improve retrieval effectiveness with automatically identified phrases, this paper proposes a novel QE approach utilizing Wikipedia[2] semantic annotations (i.e. categories) for query phrase expansion, especially geared toward patent search. Given that the major causes for topic drift are the ambiguity of query terms and the method adopted by PRF in weighting and selecting expansion terms from the top-ranked local documents, we attempt to alleviate these problems by expanding query phrases using external resources rather than using collection-dependent statistical phrases. More specifically, we propose to reduce query topic drift in QE by using both WordNet and Wikipedia as the source of expansion and at the same time by enriching a query with phrases for query context disambiguation. Using the two resources together allows for not only handling word synonyms but also compensating for the limitation of WordNet.

Aside from technically oriented motivations for devising a more effective QE method, we propose our method to make a contribution to patent search, which has arisen as one of the important information retrieval fields, especially for legal IR (Azzopardi and Vinay 2008; Maxwell and Schafer 2008). Current patent search systems use a keyword-based approach, and therefore their retrieval effectiveness relies on the quality of search keywords (Xu and Croft 2009). However, patents usually contain a large number of phrases because they often deal with technical vocabulary, and this aspect served as a strong motivation for us to use phrase-based QE. In addition to the unique characteristics of patent documents in terms of their vocabulary, usage, and structure, which make patent search different from others, individual patent documents are manually assigned to one or more classes from International Patent Classifications (IPCs). IPC has a hierarchy of four levels (Section, Class, Sub-Class (SC), and Group) with the "section" being the most general.[3] The group level can be divided into two main categories: Sub-Group (SG) and Main Group (MG). In this work, we chose to use SC, MG, and SG to represent semantic classes and evaluate the proposed method. For example, when a patent is assigned to the IPC code "G06F 17/30", its SC, MG, and SG are "G06F", "G06F 17", and "G06F 17/30", respectively. A group of patents belonging to a Sub-Class are said to be relevant to each other at a higher level of abstraction than those in the other two groups. The existence of the IPC hierarchy allows us to study generalization/specialization capabilities of phrases. Our retrieval task is thus to classify patent documents to those classes of the query as closely as possible, and the proposed query expansion approach was tested with the US Patent & Trademark Office (USPTO) patents provided for several NTCIR[4] tasks, and the CLEF-IP 2011 collection designated specifically for the patent classification tasks.

In our experiment, comparisons were mostly made against relevance-based language model (RM) (Lavrenko and Croft 2001) because it has often been used as a comparison

[2] http://www.wikipedia.org/.

[3] http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

[4] http://research.nii.ac.jp/ntcir/.

benchmark (Al-Shboul and Myaeng 2010; Yin et al. 2009; Lang et al. 2010). Among many PRF-based efforts (Lavrenko and Croft 2001; Yin et al. 2009; Lang et al. 2010; Bai and Nie 2008), RM has drawn much attention with its strong probabilistic ground and 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:

$$\#weight(w_1 Original\_Query\, w_2 Expansion\_Terms)$$

where w_1 and w_2 are normalized weights, set to 0.5 by default. Since expansion terms are not guaranteed to be relevant to the query topic in the first place, we try to find the optimal weight balance between the original query and the expansion terms for producing the best effectiveness and then compare our proposed model with the best possible results obtained from RM.

The contributions of this work are: (1) a new semantically-based query phrase expansion method utilizing Wikipedia as an external resource, (2) a method of complementing the query-drift effect caused by word-based QE with phase-based QE through an optimal combination of WordNet-based words and Wikipedia-based phrases in the context of patent search, (3) comparisons among a pure Wikipedia-based QE method, statistical word-based QE using PRF, and the proposed method in terms of retrieval experiments and analytical methods for their intrinsic characteristics, and (4) identification of the patent components that gave the best retrieval performance.

This paper is organized as follows. In Sect. 2, a review of the related work is discussed. Our proposed query expansion method is described in Sect. 3. Experiment goals, environment and results are discussed in Sect. 4. In Sect. 5, evaluations of intrinsic aspects are presented. Finally, we conclude in Sect. 6.

2 Related work

Among several automatic QE approaches, PRF has shown its value in improving retrieval effectiveness (Xu and Croft 1996). While several attempts to enhance PRF were reported (Lee et al. 2008; Cao et al. 2008; Lavrenko and Croft 2001), others tried different approaches such as expansion based on query characteristics (Xu et al. 2009), mining user logs (Cui et al. 2003), and using external resources (Kwok and Chan 1998). Among many PRF-based efforts (see for example (Lang et al. 2010; Bai and Nie 2008)), Relevance-based language model (RM hereafter) (Lavrenko and Croft 2001) has drawn much attention with its strong probabilistic ground. Assuming that a query and relevant documents are sampled from an unknown relevance distribution, it determines the probability of observing a term in a set of relevant documents based on its co-occurrence with the query terms. RM adds top ranked words to the relevance model of the query, and then a new query becomes a mixture of both original query words' model and relevance words' model. Recently, Bendersky et al. (2011) have utilized external resources for generating features used for weighting different classes of concepts in the query, as well as the latent concepts selected for expanding it. In our work, we follow the strategy of using external resources as a way of enriching and disambiguating queries with relevant words and phrases and then validate our approach by making a comparison against RM because it has often been used as a comparison benchmark (Lang et al. 2010; Bai and Nie 2008; Bendersky et al. 2011).

WordNet has been utilized for QE in different ways. For example, its semantic relations have been used to solve the problem of vocabulary mismatch (Voorhees 1994). In Navigli

and Velardi (2003), the query words were expanded separately by intersecting each word's synsets sharing a lexical relation in different resources. They showed that using WordNet synsets in addition to *"glosses"* as expansion words contributed significantly towards higher effectiveness. We compare our model with the best reported results in Navigli and Velardi (2003).

Wikipedia is an online collaborative contribution of the Web community to building an encyclopedia. It was utilized for IR-related tasks including word-sense disambiguation (WSD) (Navigli 2009), named-entity retrieval (Li et al. 2007), document clustering/classification (Banerjee et al. 2007), question answering (Ganesh and Varma 2009), prior art search (Lopez and Romary 2009), and QE (Li et al. 2007; Arguello et al. 2008). In Wikipedia, a page describing a concept expressed as a word or phrase belongs to one or more categories, each of which contains a category title (e.g. Information Retrieval), titles of the pages in this category (e.g. discount cumulative gain, generalized vector space model, etc.), and other related categories (e.g. Information Science). To the best of our knowledge, our work is the only one that utilizes the three semantic resources (individual categories, page titles under each category, and links to other related categories) for advanced QE whereas others used either links and anchored text for QE in a RM-based model for blog recommendation (Arguello et al. 2008) or simple categorical information (Li et al. 2007). Our work is based on deep understanding of the semantics available in Wikipedia, which is essential to the query phrase expansion, which is in turn critical for patent search.

Several types of phrases (i.e. noun phrases, head/modifier, bigrams, and others) have been used for different IR applications (Kapalavayi et al. 2009; Al-Shboul and Myaeng 2010; Arampatzis et al. 1998; D'hondt et al. 2013), but very rarely for query expansion (Jones and Staveley 1999). In Jones and Staveley (1999), for example, a phrase-based interactive system was proposed in which phrases were extracted using KEA (i.e. a machine learning-based phrase extraction tool) to link, browse, and query documents in a collection to retrieve a ranked list of similar documents. Our work is in line with them in its attempt to emphasize the importance of phrases for IR but unique in that phrases are recognized in queries for QE and in Wikipedia as candidates for expansion phrases. In a different effort (Al-Shboul and Myaeng 2010), semantic information extracted from DBpedia is utilized. A phrase is run against the DBpedia index using cosine similarity. SKOS (Simple Knowledge Organization System) of top ranked documents are used as thesauri to expand phrases. The work proposed in this paper shares the same idea of using an external resource for phrase-based QE but with greater depth and different methods.

3 Wikipedia-based query phrase expansion (WQPE)

Our QE method depends on expanding both query words and phrases although phrase expansion is the key. The goals behind this strategy are in twofold: (1) utilizing and providing additional contextual information with phrases for queries and (2) reducing vocabulary mismatches of both words and phrases. For query word expansion, we take a simple strategy of considering the first returned synset after searching WordNet instead of trying to further the widely studied area. On the other hand, we devised a novel method by which query phrases are expanded based on lexico-semantic matching against Wikipedia pages. As a preparation step, each Wikipedia page is transformed into a surrogate serving as a semantically oriented summary by gathering its categories (primary categories), page titles belonging to the primary categories, and category pages with in/out links to the

primary categories (i.e. secondary categories). A query phrase is then matched against each surrogate page's primary/secondary categories and titles and scored so that surrogate pages are ranked according to their lexico-semantic similarity values. Top ranked surrogate pages are selected and intersected to extract common phrases to be added to the query.

3.1 Query term extraction

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 by comparing the performance in our experiments as explained in Sect. 4. We first apply a stop word filter that uses a customized list collected from different sources (i.e. KEA,[5] InQuery, Lemur)[6]. We then apply Stanford POS Tagger[7] to the resulting text and subsequently Regular Expressions (RegEx) to extract keywords and keyphrases. This process is particularly useful for short fields like titles and abstract as the brevity does not warrant statistics-based methods (e.g. TF-IDF, Chi square, information gain) for word/phrase selection. Nouns, verbs and adjectives are extracted as keywords, and then phrases obtained by using RegEx. Note that the unigrams that are covered by a phrase are removed from the query as the phrase is considered as a disambiguated version of the word with additional contextual information. We use the following RegEx used to extract keyphrases:

$$(\text{VBG}|\text{VBN}|\text{JJ}|\text{JJR}|\text{JJS})(\text{NN}|\text{NNS}|\text{NNP}|\text{NNPS})+$$

where VBG and VBN are verbs, JJ, JJR and JJS adjectives, and NN, NNS, NNP and NNPS nouns of various forms such as singular, plural, and proper nouns. This RegEx was built based on our observations over the tagged query patents. An example is shown in Fig. 1.

3.2 Query term expansion

After a query is generated 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"). This process of expanding phrases, not words, is particularly important for patent search because a technology is often expressed in different word and phrases to minimize the possibility of being treated as similar to previously granted patents. It is important to maximize recall with various term expansions.

A phrase is expanded using the Wikipedia page surrogates. We first score the surrogate pages by employing the following phrase likelihood model:

$$P(ph_j|D) = \sum_{i \in G} \lambda_i p(ph_j|\theta_{i,j}) \quad (1)$$

where G = {Primary Categories of the page, Secondary Categories of the page, Titles of the pages under the Primary and Secondary Categories belonging to the page} represents the three different fields of a surrogate page D for which language models θ_i must be built

[5] http://www.nzdl.org/Kea/.

[6] www.lemurproject.org.

[7] http://nlp.stanford.edu/software/tagger.shtml.

Conception, filter, fable, removal, apparatus, vessel, job, providing, depression, associated, stuff, invention, holdfast, vas, setup, fiction, neckband, waterproofing, innovation, fabrication, interpolation, remotion, excogitation, material, fastening, pass, fixing, fastener, bags, design, "liquid passing", "sealing wax", "simple insertion", "force per unit area", "filtering liquids", "filterable seal", "passing play", "passing game", "pressure level", "low cost", "sealing problem", "flexible collar"

Fig. 1 An example query extracted from a patent title and abstract using RegEx

and $\lambda_i'\text{s}\left(\sum_{i\in G}\lambda_i = 1\right)$ are the mixture weights that represent the importance of the fields in calculating a score for each surrogate. They were set to 0.5, 0.3, and 0.2 empirically.[8] Primary Categories show the strongest associations among the documents under the same category whereas Secondary Categories defining the category context are used in case of category vocabulary mismatch. Titles become useful when the categories do not match although they are related to each other. This happens when a single concept is used and explained in two different domains. Phrase likelihood is estimated as:

$$p(ph_j|\theta_{i,j}) = \frac{c(ph_j, X_i)}{c(X_i)} \tag{2}$$

where $c(ph_j, X_i)$ is the count of phrase ph_j in the X_i field of the surrogate page and $c(X_i)$ is the count of all phrases in the field. After the relevant page surrogates are ranked for the phrase based on (1), the sets of phrases contained in the categories and titles of the top ranked page surrogates are intersected to find common phrases. In the current implementation, phrases found in 60 % or more pages among top k ($k = 5$) are added to the query.[9]

Finally, we apply an additional term filtering process to remove n-grams in the expanded query, which are covered by other longer n-grams. The rationale behind this duplicate removal process is that we consider the deleted terms (i.e. n-grams) have been specialized and thus disambiguated with the additional words found in the longer n-grams.

3.3 Patent retrieval and ranking

The goal of our patent search task is to retrieve as many documents as possible whose IPC codes contains the code of the query patent suppressing those whose IPC code is different. If a retrieved patent belongs to the IPC class to which the query patent belongs, it should be considered relevant to the query. As a result, relevancy and the evaluation results will vary depending on the level of abstraction used in IPC.

An expanded query generated by the method mentioned above (or other variations for the experiments) is run against the patent index using Okapi BM25 (see details about the experimental system in Sect. 4). After a ranked list of patents (R_Q) is retrieved by the query, they are re-ranked using the IPC information associated with them in order to suit the task defined above. A re-ranking algorithm should consider the fact that a search result is likely to contain multiple documents belonging to the same IPC class, as well as the

[8] Several combinations were tested on the whole query set in the NTCIR experiments to choose the particular combination.

[9] We ran a series of tests by varying k and the intersection percentage of the surrogates, where a particular phrase must be found, to pick 5 and 60 %, respectively. A combination of a larger k and a smaller percentage (i.e. less tight intersection) resulted in excessive phrases whereas a large percentage gave no or too few phrases unless k is sufficiently small (i.e. 5).

possibility that each document may have multiple IPC classes. Thus, our re-ranking process consists of three successive stages: *IPC Mapping*, *IPC Expansion*, and *IPC Scoring & Re-ranking*.

- *IPC Mapping* is the process of replacing a document ID with the IPC codes in <ID, score> pairs corresponding to the retrieved documents. For example, a pair of document ID and retrieval score <5188731, 8.60372> will be replaced with <{E03C 1/264, E03C 1/26}, 8.60372>, where "E03C 1/264" and "E03C 1/26" are the IPC codes assigned originally to the document 5188731.
- *IPC Expansion* considers the pairs containing more than one IPC and split them into separate entries. For the previous example, the <IPC, Score> is split into two pairs: <E03C 1/264, 8.60372> and <E03C 1/26, 8.60372>. The expanded list of pairs is denoted as R_{IPC}.
- *IPC Scoring & Re-ranking* re-scores redundant IPCs in R_{IPC} into the list L_Q that will contain distinct, but ranked in descending order according to their new scores, IPCs for the query Q. IPC codes are re-scored as follows:

$$Score(ipc) = \frac{\sum_{ipc \in L_Q} rank(ipc, R_{IPC})}{count(ipc, R_{IPC})} \tag{3}$$

where *rank* is a function returning the score of *ipc* in the R_{IPC}, and *count* returns the count of *ipc* in R_{IPC}. In this step, IPC codes with higher frequency and higher scores of the associated documents are favored.

4 Experiments

4.1 Experimental goals

Our experiments are centered around the research question, whether expanding query phrases using Wikipedia categories and page titles would help enhancing the effectiveness of patent search by limiting topic drift in word-based query expansion. To see the value of using phrases for query expansion, we examine word-based and phrase-based expansions separately as well as together. To put the value of the proposed method in context, it is compared against other word-based query expansion methods such as RM and WN-Gloss (Navigli and Velardi 2003), and a Wikipedia-based query expansion method (Arguello et al. 2008). Finally, as words and phrases have generally different distributions within the patents, an empirical study was conducted to find the best weight balance between words and phrases in the expanded queries.

4.2 Experimental setting

For our experiments, we first used NTCIR-6 USPTO patents of 1993–2002 (1,315,471 documents) as the main test collection (NTCIR Collection hereafter). For query expansion, we used Wikipedia Dump[10] for extracting Wikipedia articles. Query patents were selected randomly from those contained in the IPC groups having at least two patents so that at least one relevant patent remains in the group to be retrieved. They were removed from the document set that became the retrieval target. A total of 1,780 patents were selected as the

[10] http://dumps.wikimedia.org/enwiki/20100817 (last visited April 5th, 2011).

queries. Relevance judgments (i.e. the ground truth) of the queries were made based on IPCv9 crawled from the USPTO website for all patents.[11] A patent is deemed to be relevant for a query at one or more of the three levels (i.e. Sub-category (SC), Main Group (MG), or Sub-group (SG)) if its IPC matches with that of the query at least partially.

We employed Indri,[12] an indexing and retrieval tool of the Lemur Project and used Okapi BM25 as our default retrieval model. While Mean Average Precision (MAP) and recall were the main measures for effectiveness, IPC code matching was used for relevance judgments due to the absence of usual human-generated relevance judgments among query patents and the patent documents in the collection. MAP and recall for n retrieved documents (1,000 documents in this experiment) for |Q| (1,780) queries are computed as follows:

$$MAP = \frac{\sum_{q=1}^{|Q|} AvgP(q)}{|Q|} \tag{4}$$

$$AvgP(q) = \frac{\sum_{i=1}^{n} Precision(i).rel(i)}{\#Relevant\_Retrieved\_Documents} \tag{5}$$

$$Precision(n) = \frac{\#Relevant\_Retrieved\_Documents}{n} \tag{6}$$

$$Recall = \frac{\#Distinct\_Relevant\_Retrieved\_IPCs}{\#All\_Relevant\_IPCs} \tag{7}$$

where *AvgP(q)*, average precision for query q, is computed with *rel(i)*, an indicator function that returns 1 if the document at rank i is relevant and 0 otherwise. It computes precision every time a document is retrieved and takes an average. *Relevant_Retrieved_Documents* represents the number of documents with relevant IPC codes in R_Q, *Distinct_Relevant_Retrieved_IPCs* the number of distinct IPCs correctly retrieved, and *All_Relevant_IPCs* the number of IPCs assigned to the query patent. This method of evaluation follows the same concept used to evaluate NTCIR-8 patent mining tasks (Iwayama et al. 2007). It is worth mentioning that recall is calculated based on the retrieved IPC codes rather than the retrieved documents as the goal is to recommend IPC codes for the examiner to look at during the examination process. The patent search process therefore looks like patent classification using IR. However, it is slightly different from traditional machine learning-based classification process where the output is a class assigned for the patent.

On the other hand, the CLEF-IP 2011 Collection consisting of 3,118,089 patents in English, French, and German, among which 2,082,113 patents in English were provided.[13] In addition, 3,000 topics (patents), equally divided over the three languages, were provided for each task as participants were asked to classify them into IPCs. Two types of patent classification tasks have been offered: patent classification (task-1), where the participants were asked to classify topics into general IPC classes (i.e. Sub-Class level), and refined patent classification (task-2), were the participants were asked to classify patents into specific IPC classes (i.e. Sub-Group level). The Sub-Class of each topic was provided as additional information to be used in refining classification. For evaluation, we used the provided relevance judgments for each task and the same evaluation method adopted by the CLEF-IP tasks (i.e. trec_eval v9.0 provided by NIST).

[11] http://www.uspto.gov.

[12] http://www.lemurproject.org.

[13] Piroi, F.: CLEF- IP 2011: Track Guidelines. IRF, Vienna (2011).

Table 1 shows the average query length (i.e. number of terms) in both the Original and WQPE queries using different fields of the original patents. Queries are classified into three categories according to their lengths: short (i.e. Title (**T**), Abstract (**A**), Title + Abstract (**TA**)), intermediate (i.e. Claims (**C**), Title + Abstract + Claims (**TAC**)), and long queries (i.e. Title + Abstract + Specifications (**TAS**), All Patent Parts (**APP**)). The table also shows the exact number of words (**W**) and phrases (**P**) in each of the queries as well.

As can be seen in Table 1, WQPE added some terms to the short and intermediate queries but not to the long ones. This is caused by the large number of terms in the long queries corresponding to several redundant synsets that are a priori removed. The slight decrement in number of phrases on long queries is caused by removing short phrases (i.e. two words) that were covered by longer ones (i.e. four or five words). Note that RM generated a significantly larger number of words for each query type because it is based on a word based context free expansion.

4.3 Preliminary experiments

While RM-expanded queries generally follow a weighting structure that provides 50 % of the weight to the original query terms (i.e. Baseline part) and the other 50 % to the expanded ones (i.e. Relevance part), we attempted to optimize its performance by adjusting the ratio from 10 % for the baseline part and 90 % for the relevance part until a 90 % for the baseline part and 10 % for the relevance part was reached. To tune weights for RM queries, a set of preliminary experiments for all variants of weights were conducted, and the ratio of 60:40 was selected for Baseline/Relevance parts of the queries as they generated the best effectiveness. Likewise, WQPE required tuning since a balance between words and phrases in the expansion method seems important. A set of preliminary experiments with a subset of queries[14] (shown in Fig. 2a, b) show that fixing the weight of phrases twice as big as that of words gave the best results amongst all other combinations. At first, the weight of phrases was fixed to one, while words were assigned different weights from 1 to 8. Then, the weights of words were fixed to one, and then phrases were assigned variant weights between 1 and 8. While increasing the weight of words while fixing the weight of phrases caused a sharp fall in both MAP and recall, changing the weight of phrases while fixing the weight of words didn't always affect them negatively (as shown in Fig. 2). Therefore, giving higher weights to phrases than the 2:1 ratio found in our experiment would over-specify the query.

To determine which parts of a patent will generate the best query terms, preliminary experiments with various combinations of patent components were undertaken using the NTCIR collection. We found that a combination of Titles and Abstracts (**TA**) was the best as shown in Table 2. It was also observed, as in Table 3, that RM and WQPE also obtained their best results with the same combination of the two fields. Thus all the results reported subsequently in this paper are for queries extracted from those two fields.

4.4 Main experiments with the NTCIR collection

In our experiments two different baseline queries were used: unigram original queries and the unigram original queries plus phrases (generated using the regular expression in Sect. 3.1) after deleting the unigrams involved in the phrases (denoted as "word/phrase"). Expanded queries of different kinds are compared against these baselines so

[14] An additional set of queries were used to tune RM and WQPE. The set consists of 100 queries generated from randomly selected patents from the NTCIR collection.

Table 1 Average query lengths with all terms, words only, and phrase only

	T		A		TA		C		TAC		TAS		APP	
	W	P	W	P	W	P	W	P	W	P	W	P	W	P
Original														
All Terms	4.28		29.36		30.82		71.65		84.29		447.24		465.54	
Words or Phrases only	3.0	1.28	22.31	7.05	21.88	8.96	50.87	20.78	56.47	27.82	268.34	178.90	274.67	190.87
RM														
All Terms	54.28		79.36		80.82		121.65		134.29		497.24		515.54	
Words or Phrases only	53.0	1.28	72.31	7.05	71.88	8.96	100.87	20.78	106.47	27.82	318.34	178.90	324.67	190.87
WQPE														
All Terms	7.56		49.14		51.34		103.50		122.47		488.87		511.35	
Words or Phrases only	5.75	1.81	36.86	12.29	36.96	14.38	73.49	30.02	85.73	36.74	317.54	171.33	327.60	184.75

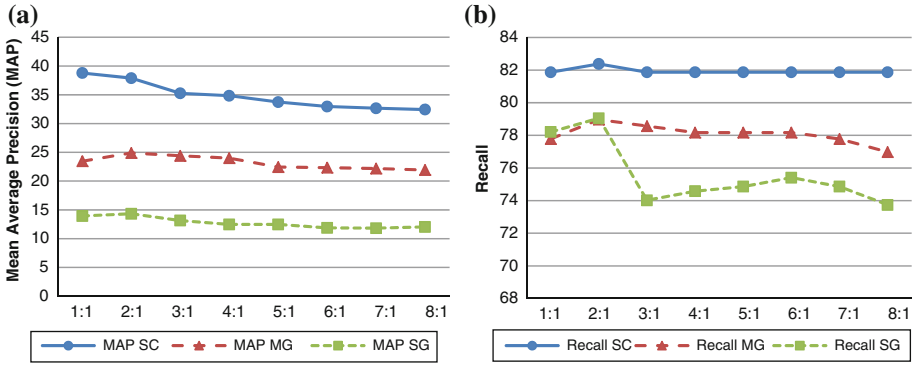

Fig. 2 (a) Changes in Mean Average Precision (MAP) with phrase-to-word weight ratios (b) Changes in Recall with phrase-to-word weight ratios

Table 2 A comparison among different patent fields (and field combinations) used as queries for the original unigram queries

	Original (Unigram)					
	Sub-Class		Main-Group		Sub-Group	
	MAP	Recall	MAP	Recall	MAP	Recall
T	29.57	82.01	19.28	79.55	11.59	61.30
A	35.89	83.30	21.30	81.91	14.36	64.16
TA	**39.19**	**84.40**	**23.99**	**80.03**	**15.64**	**72.71**
C	26.67	83.04	17.38	80.21	11.19	68.99
TAC	26.81	83.26	17.93	81.11	11.80	60.67
TAS	12.34	82.14	7.57	45.57	4.19	44.39
APP	15.82	82.04	10.55	45.73	6.67	36.58

Numbers in bold are the best results obtained for each IPC level for the original unigram query

that the performance differences we observe are not just due to the addition of new phrases to the unigram baselines. The first query expansion is only with Wikipedia as in the second groups in Table 4: a state-of-the-art Wikipedia-based query expansion method (denoted by WikiLinks) and WQPE-Wikipedia (expanding original queries using Wikipedia but not the WordNet-based expansion). The second expansion, the third group in Table 4, is only with WordNet: a state-of-the-art method and WQPE as denoted by WN-Gloss and WQPE-WordNet without using Wikipedia. Finally, we compare our proposed method (WQPE) using both WordNet and Wikipedia against RM, which is considered a state-of-the-art benchmark in PRF-based query expansion, and two other phrase-based QE methods.

Between the two baselines, a slight decrease in mean average precision (MAP) as well as recall was observed with the word/phrase queries at all classification levels compared to the unigram queries. Our analysis showed that the retrieved sets for the two cases contain almost the same relevant documents but the rankings were different. Recall was not expected to be improved because the phrase expansion did not add new terms. However, in order to understand the decrease in MAP, we did further analysis and found some of the

Table 3 A comparison among different patent fields (and field combinations) used as queries for RM and WQPE

	Sub-Class				Main-Group				Sub-Group			
	RM		WQPE		RM		WQPE		RM		WQPE	
	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall	MAP	Recall
T	47.02	82.7	50.58	84.4	30.66	78.79	33.27	82.58	18.43	81.56	19.73	87.64
A	50.7	84.75	52.14	84.88	33.87	82.53	36.33	82.79	22.84	86.12	24.54	86.42
TA	**50.86**	**84.91**	**54.34**	**84.97**	**34.1**	**82.84**	**38.37**	**82.92**	**22.85**	**86.4**	**25.82**	**86.95**
C	42.4	84.34	27.66	83.13	27.63	79.84	16.77	73.71	17.8	77.9	9.66	58.72
TAC	42.62	84.69	26.84	83.51	28.51	81.26	16.76	75.05	18.76	80.56	10.09	61.84
TAS	19.62	82.91	9.89	80.72	12.03	72.45	4.13	62.22	6.66	54.68	1.91	35.37

Numbers in bold are the best results obtained for each IPC level for each query

Table 4 A comparison between the original, RM-expanded, and WQPE-expanded queries

Query	SC (3,155)		MG (3,801)		SG (5,391)	
	MAP	Recall	MAP	Recall	MAP	Recall
Unigram	39.19	84.40	23.99	80.03	15.64	72.71
Word/Phrase	36.75	84.12	22.15	78.19	13.32	70.69
WQPE-Wikipedia	33.97	84.21	20.9	78.37	12.23	68.72
WikiLinks (Arguello et al. 2008)	50.57	77.65	37.94	73.30	22.06	78.52
WQPE-WordNet	40.19	84.43	24.66	80.53	15.84	75.71
WN-Gloss (Navigli and Velardi 2003)	22.39	82.12	12.95	75.24	9.25	61.07
RM (Lavrenko and Croft 2001)	52.99	83.99	36.07	81.76	22.86	87.01
WQPE	**54.34 (2.55 %)**	**84.97 (1.17 %)**	**38.37$^{\alpha}$ (6.38 %)**	**82.92 (1.42 %)**	**25.82$^{\alpha}$ (12.95 %)**	**86.95 (−0.07 %)**

α indicates significance at ($p = 0.05$) in a t test

phrases such as “the like”, “first member”, or “number one” did more harm than good in ranking. As a result, all the comparisons in Table 4 are against the unigram queries.

Table 4 shows a comparison between WQPE-Wikipedia that expands query phrases only and the expansion method using hyperlinked words and phrases within Wikipedia pages (i.e. WikiLinks). The result is that the latter outperformed the former in terms of MAP. Since the WQPE-Wikipedia result is very close to the baseline word/phrase query case, a possible reason for the lower performance is because few expansion phrases were added to the original phrases not allowing them to make a significant effect on the query effectiveness by themselves.

A similar comparison was made between two methods of using WordNet for expansions: WQPE-WordNet and WN-Gloss (Navigli and Velardi 2003). While the latter is a method to expand query terms using their glosses in WordNet, the former has shown higher effectiveness at all of classification levels. However, the increase obtained by WQPE-WordNet was not statistically significant compared to the unigram original queries. It was found that while 76 % of words added by WordNet (representing on average more than 23 % of all query words) had very high IDF weights, many of them were irrelevant to the patent topics. Another interesting phenomenon is 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 expanded queries due to the high IDF values and affected the search negatively, causing the query topics to drift.

As a reference model to be compared against our proposed method, we chose RM. RM is shown to be generally superior to the original queries and Wikipedia- and WordNet-expanded queries in terms of both MAP and recall. However, as terms in the expanded parts are selected according to a context-free statistical model, many of them seem irrelevant to the query topic. An example is shown in Fig. 3 where words with underlines are either irrelevant or general terms added by RM to the query. Note that the WikiLinks based query expansion performed comparably to RM. This indicates that selected hyperlinked texts contain good expansion terms comparable to the ones selected using RM.

filter, invention, material, collar, vessel, bag, liquid, means, side, end, air, view, apparatus, present, device, section, outer, body, such, provide, surface, comprising, according, inner, pressure, oil, attached, element, plastic, top, flexible, method, fastening, embodiment, preferred, portion, media, container, system, object, seal, comprises

Fig. 3 An example RM expansion terms where underlined words are irrelevant to the query topic

alphabetical, assigned, attribute, auxiliary, base, brief, conclusion, cycles, hdm, high, increasing, institute, invention, japanese, joining, lee, machine, main, manner, memory, name, natural, necessary, new, nippon, number, numerical, objects, order, ordinary, parallel, plurality, pre-determined, preferred, present, producing, record, relation, relational, resulting, same, several, show, significant, simple, single, speed, storing, summary, table, transfer, translation, value, version, central processing unit, computer address, data processor, information processing system, trading operations, word form, unit of measurement

Fig. 4 An example WQPE expanded query where underlined terms are expansion phrases

Another example is shown in Fig. 4 where the underlined phrases are expanded ones added by WQPE (the Wikipedia expansion part). While RM-based expansion may cause a topic drift, the new query helps disambiguating general terms like "computer", "data", and "information" by including the underlined phrases and move the query into a more specific technical domain for a better focus. This query specification role of the proposed method would be particularly useful when the original query contains noisy words (e.g. "same", "show", "lee", "pre-determined") and general terms shared among different fields (e.g. "relational", "table", "storing"). The query shown in Fig. 4 achieved approximately 14 % improvement in MAP over the corresponding RM-expanded query.

The highlight of the experimental results is the comparison between the proposed method and the rest mentioned already. It is shown that WQPE outperformed all the other QE methods including RM. The result is attributed to 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 less ambiguous queries than RM expanded ones, thereby limiting the topic drift phenomenon of word-based QE. 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. As shown in Table 4, WQPE performed better than RM at all levels of classification.

As a case study for illustration of the effects of the proposed query expansion methods, we selected two baseline queries with their PRF and WQPE expansions and observed the changes in IPC rankings for each retrieved set. Patent 5'188'731, for example, has two relevant IPCs. The average rank of the IPCs for the baseline query was 189th while it moved to the 28th after PRF expansion and then to the 8th after WQPE. Another example is patent 5′254′127 where three relevant IPCs exist. The IPCs average rank was the 87th for the baseline query while it became the 9th and the 5th for RM and WQPE expanded queries, respectively. These examples show the effect of query expansions in improving the average ranks of relevant IPCs. As can be seen in the experiments, the combination of WordNet-expanded words and Wikipedia-based expanded phrases exploited the best of both to achieve the best MAP and recall.

4.5 Experiments with the CLEF-IP 2011 collection

We ran experiments with the CLEF-IP 2011 collection for further validation and generalizability of our findings from the experiments reported in the previous sub-section. In

Table 5 Comparison against the two best performing systems in CLEF-IP 2011 SC level classification

Query	P	R	F_1
NIJMEGEN (Verberne and D'hondt 2011)	0.5436	0.8563	0.6186
WISENUT (Seo et al. 2011)	0.2885	0.8389	0.4032
RM	0.4888	0.8910	0.6312
WQPE	**0.6682**$^{\alpha}$	**0.9123**$^{\alpha}$	**0.7714**$^{\alpha}$

The superscript α indicates significance at ($p = 0.05$) in a t-test

Table 6 Comparison against the two best performing systems in CLEF-IP 2011 SG level classification (more refined from SC)

Query	P	R	F_1
NIJMEGEN (Verberne and D'hondt 2011)	0.0731	0.0622	0.0609
WISENUT (Seo et al. 2011)	0.2930	0.4954	0.3328
RM	0.3333	0.5642	0.4190
WQPE	**0.4110**$^{\alpha}$	**0.6589**$^{\alpha}$	**0.4925**$^{\alpha}$

The superscript α indicates significance at ($p = 0.05$) in a t-test

particular, we have applied our method, as well as RM,[15] for the two tasks at CLEF-IP 2011, Patent Classification (task-1), and Refined Patent Classification (task-2) (Piroi et al. 2011). Our experiments used the English part of the CLEF-IP 2011 collection covering 1,000 topics (patents) in each task, requiring an index of 2,082,113 documents for Indri. Similarly to the previous experiments with the NTCIR collection, query terms have been extracted from the titles and the abstracts of each topic patent. Results are reported in Table 5 (task-1) and Table 6 (task-2). It is important to mention that for task-2, WQPE has been applied without the additional SC information provided for each topic targeting to check the effectiveness without such information.

As in Table 5, the NIJMEGEN (Verberne and D'hondt 2011) method, which was shown at the contest to be most successful in classifying topics for SC, failed in classifying them further down in the IPC hierarchy (i.e. SG) as shown in Table 6. While the WISENUT (Seo et al. 2011) method was shown to be the best for the deeper IPC classification level (i.e. SG), its performance drops significantly, especially in precision, for the task at a higher level classes (i.e. SC) as shown in Table 5. In both of the tasks, WQPE gave a significant improvement in both precision and recall over the two best performing methods in the respective tasks at CLEF-IP 2011. For the sake of completeness and generalizability of the superiority of the proposed method over RM, we show the performance of RM for the same set of queries. Although RM is shown to be better than NIJMEGEN and WISENUT, it is inferior to WQPE. The improvements in F_1 are 22.21 and 17.54 % over RM for the two tasks, respectively. The experimental results based on the two different setting indicates the importance of using both words and phrases to increase coherency and reduce the influence of the topic drift phenomenon in word-based QE methods.

[15] The same settings for both WQPE and RM described and used in 4.4 have been used.

5 Evaluations of intrinsic aspects

In addition to precision and recall, other methods for evaluating various aspects of queries or retrieval systems exist. Retrievability (Azzopardi and Vinay 2008) is a measure to calculate the portion of indexed documents that can be retrieved through querying. It has a great importance in the patent domain as finding all relevant documents to a search query (i.e. query patent) is an important feature in a patent search engine (Lupu et al. 2011). Now that WQPE has proven its effectiveness against other QE methods (shown in Tables 4, 5, and 6) in terms of MAP, precision, and recall values computed for different IPC levels, it is also important to prove the intrinsic value of the proposed method by checking the extent to which all relevant documents can be accessed. For brevity, all analysis results reported in this section will be based on the results obtained from the main experiments based on the NTCIR collection only. While retrievability was originally devised to evaluate the influence of an IR system on accessibility to a collection, it can be used to show the effect of expanded queries against the original ones by fixing the retrieval model or engine. More specifically, we aim at showing our proposed QE method contributes towards reducing a retrieval bias incurred by the characteristics of a retrieval system (e.g. different document length normalization methods and methods for considering document–document relationships as in PageRank). Retrievability can capture the extent to which a QE method makes it possible to discover the documents that cannot be retrieved by the original queries, which cannot be measured by either MAP or recall we used with IPC-based relevance.

Clarity (Cronen-Townsend et al. 2002) is another way of predicting query performance which is strongly correlated to query effectiveness and inversely correlated with ambiguity. It was originally designed to identify ineffective queries without relevance information, by computing the relative entropy between a query language model and the corresponding collection language model. It assumes that a query whose highly ranked documents are about a single topic (coherent documents) has a model that is characterized by large probabilities for a small number of terms, while a query with highly ranked documents from mixed topics (incoherent documents) has a smoother model similar to the whole collection model. In this work, clarity is used to compare expanded queries generated by different methods; aiming specifically at showing WQPE expanded queries are less ambiguous than others with respect to the document collection. In a sense, retrievability and clarity are complementary to each other in that they roughly correspond to the notions of recall and precision, respectively, while not relying on relevance judgments.

5.1 Retrievability

We investigated the effect of the WQPE method on reducing the retrieval bias, which has a direct bearing on retrievability. OKAPI BM25 was used to retrieve patents for three different QE methods because it has been reported to have the lowest retrieval bias among several retrieval methods (Azzopardi et al. 2010; Bashir and Rauber 2010).

First, retrievability of each document is calculated according to Azzopardi and Vinay (2008) as follows:

$$r(d) = \sum_{q \in Q} f(k_{dq}, c) \tag{8}$$

where f is a cost function, k_{dq} is the rank of document d that contains query term q, and c is the number of examined results. $f(k_{dq},c)$ shall return 1 if $k_{dq} < c$, and 0 otherwise.

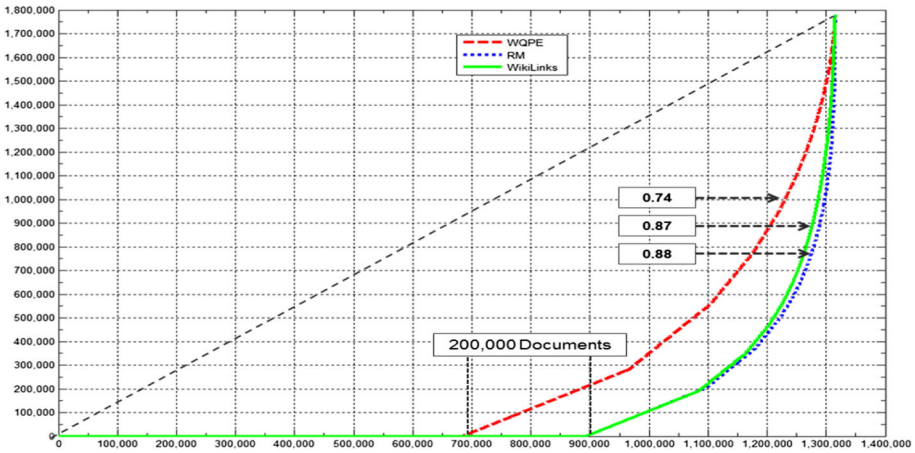

Fig. 5 Lorenz Curves for three query expansion methods, and their Gini-Coefficient scores

Calculation results can be further analyzed using *Lorenz Curve* where documents are sorted according to their retrievability scores in ascending order, plotting a cumulative score distribution. In Fig. 5, the x-axis and y-axis represent the number of documents and the cumulative retrievability scores, respectively. If the retrievability of documents is distributed equally (all documents are equally retrievable) the Lorenz Curve will be linear. On the other hand, a curve skewed to the right indicates a greater amount of bias the retrieval method has as in Fig. 5.

In our experiment, top ranked 100 retrieved documents were considered in calculating retrievability for a query, similarly to the way described in Azzopardi and Vinay (2008), Bashir and Rauber (2010). As can be seen in Fig. 5, WQPE increased the retrievability by almost 200,000 documents compared to RM and WikiLinks QE methods shown by the shrinking areas under the ideal line.

Figure 5 also shows the corresponding Gini Coefficient (G) for each of the QE methods shown in the Lorenz curves. *Gini coefficient* is used to summarize the amount of bias in a Lorenz curve and computed as follows (Azzopardi and Vinay 2008):

$$G = \frac{\sum_{i=1}^{n}(2.i - n - 1).r(d_i)}{(n-1)\sum_{j=1}^{n} r(d_j)} \tag{9}$$

where n is the number of documents in collection. When G is equal to 0, all documents are equally retrievable (the ideal line of retrievability), while G equaling 1 indicates that only one document is retrievable, and all other documents have retrievability scores of 0. It is reported that WQPE has the smallest coefficient (0.74) showing closer line to the ideal retrievability while RM and WikiLinks have shown more skewed curves from the ideal line (Gini Coefficient equals to 0.87 and 0.88 for WikiLinks and RM, respectively).

5.2 Clarity

Clarity is defined as a measure of relative entropy between the query language model and the collection language model. It is also a query retrieval prediction method that is strongly correlated to query effectiveness and query ambiguity (Cronen-Townsend et al. 2002). It assumes that a highly coherent and less ambiguous query has a model characterized by

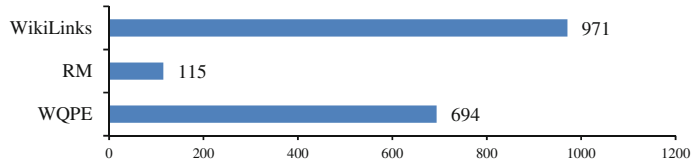

Fig. 6 Clarity comparisons among different query expansion methods

large probabilities for small number of terms; however, the language model for a low coherent query has a word probability distribution that is likely to be similar to that of the model of the whole collection. Since clarity was proposed for words rather than phrases, we extended it to handle phrases as follows:

$$Clarity(Q) = C_{Word}(Q) + C_{Phrase}(Q) \tag{10}$$

where C_{Word} is the clarity score generated from our expanded query's words as defined in Cronen-Townsend et al. (2002):

$$C_{Word} = \sum_{w \in V} P(w|Q) \log_2 \frac{P(w|Q)}{P_{Coll}(w)} \tag{11}$$

where $P_{Coll}(w)$ is the is the relative frequency of the word in the entire collection, $P(w|Q)$ is the query language model, and V is the entire vocabulary of the collection. C_{Phrase} in (10) is the score generated from the expanded phrases and computed in a similar way.

The average clarity score for WQPE expanded query words was 44 compared to 72 and 69 for RM and WikiLinks respectively, while the average clarity score for WQPE expanded query phrases was 5,192 compared to 2,223, and 4,899 to RM and WikiLinks, respectively. An important issue that arises from these results is the big difference in clarity scores between words and phrases caused by the difference between the frequencies. While words have generally higher frequencies, phrases have much less due to their positional dependency over other neighboring words, where smaller collection probability generates higher clarity score. This seems reasonable considering that phrases are less ambiguous than words, and thus they are supposed to have higher clarity values.

An analysis of query-based clarity scores between all QE methods has been made. In computing the clarity values for the expanded queries resulting from the different expansion methods, we normalized C_{words} and C_{phrase} so that total clarity does not favor a QE method that adds more phrases, which have much higher clarity values in general. Figure 6 show that WQPE method generates queries with much higher clarity than RM although the expanded queries from WikiLinks have the highest clarity.

6 Conclusions

We have proposed a new method of using Wikipedia categories and WordNet together for query word and phrase expansion (i.e. WQPE) in the task of IPC-class based patent search. 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 indicates that added phrases have the effect of decreasing topic drift caused by RM by showing better effectiveness. Furthermore, analysis shows that WQPE

expanded queries are less ambiguous than RM queries measured by clarity, which contributes to a reduction of topic drift, and more adequate for patent retrieval tasks as they have higher retrievability. Based on the obtained results, WQPE shows promising results as an adequate QE method for patent search and retrieval, especially for professional search tasks (i.e. Prior-art Search, Invalidity Search, etc.).

While this research centers on the idea of showing the feasibility of utilizing semantically-related phrases as expansion terms, it is not clear how they interact with the words expanded by WordNet in improving effectiveness. Understanding the reasons behind such improvements over a state-of-the-art method requires in-depth analysis of the experimental data, which will lead to a new hypothesis about query expansion or help developing a new methodology for investigating on the notion topic drift in automatic query expansion. We are currently working on these issues as an extension to the work reported in this paper. An additional issue that we aim at studying in the future is the effectiveness of using the POS tagger on patents, specifically on our set of topics, by showing some analysis on its performance.

Acknowledgments This research was supported by the Ministry of Science, ICT & Future Planning (MSIP), Korea, under the ITRC support program (NIPA-2013-H0301-13-3005) supervised by the National IT Industry Promotion Agency (NIPA) and by World Class University (WCU) program under the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007). We would like to thank the reviewers for their valuable comments and suggestions that helped improve the quality of this paper.

References

Al-Shboul, B., & Myaeng, S. (2010). IRNLP@KAIST in the subtask of Research Papers Classification in NTCIR-8. In *Proceedings of the 8th NTCIR workshop meeting on evaluation of information access technologies: Information retrieval, question answering and cross-lingual information access* (pp. 331–335). Tokyo, Japan.

Al-Shboul, B., & Myaeng, S. (2011). Query phrase expansion using wikipedia in patent class search. In *Proceedings of the 7th Asia conference on information retrieval technology (AIRS)* (pp. 115–126).

Arampatzis, A., Tsoris, T., Koster, C., & Van Der Weide, T. (1998). Phrase-based information retrieval. *Information Processing and Management, 34*(6), 693–707.

Arguello, J., Elsas, J., Callan, J., & Carbonell, J. (2008). Document representation and query expansion models for blog recommendation. In *Proceedings of the conference of the American association for artificial intelligence (AAAI)*.

Atkinson, K. (2008). Toward a more rational patent search paradigm. *1st ACM Workshop on Patent IR* (pp. 37–40).

Azzopardi, L., Vanderbauwhede, W., & Joho, H. (2010). Search system requirements of patent analysis. In *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 775–776).

Azzopardi, L., & Vinay, V. (2008). Retrievability: An Evaluation measure for higer order information access tasks. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM)* (pp. 561–570).

Bai, J., & Nie, J. (2008). Adapting information retrieval to query contexts. *Information Processing and Management, 44*(6), 1901–1922.

Banerjee, S., Ramanathan, K., & Gupta, A. (2007). Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)* (pp. 787–788).

Bashir, S., & Rauber, A. (2010). Improving retrievanility of patents in prior-art search. In *Proceedings of the 32nd European conference on advances in information retrieval (ECIR)* (pp. 457–470).

Bendersky, M., Metzler, D., & Croft, W. (2011). Parameterized concept weighting in verbose queries. In *Proceedings of the 34th international ACM SIGIR conference on research and development in Information Retrieval (SIGIR)* (pp. 605–614).

Cao, G., Nie, J., Gao, J., & Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 243–250).

Cronen-Townsend, S., Zhou, Y., & Croft, W. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 299–306).

Cui, H., Wen, J., Nie, J., & Ma, W. (2003). Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 829–839.

D'hondt, E., Verberne, S., Koster, C. H. A., & Boves, L. (2013). Text representations for patent classification. *Computational Linguistics, 39*(3), 755–775.

Ganesh, S., & Varma, V. (2009). Exploiting structure and content of wikipedia for query expansion in the context of question answering. In *Proceedings of the international conference on recent advances in natural language processing (RANLP)* (pp. 103–106). Borovets, Bulgaria.

Iwayama, M., Fujii, A., & Kando, N. (2007). Overview of classification subtask at NTCIR-6 patent retrieval task. In *Proceedings of the sixth NTCIR workshop meeting on evaluation of information access technologies: Information retrieval, question answering and cross-lingual information access (NTCIR-6)*.

Jochim, C., Lioma, C., & Schutze, H. (2011). Expanding queries with term and phrase translations in patent retrieval. In *Proceedings of the second international conference on multidisciplinary information retrieval facility (IRFC)* (pp. 16–29).

Jones, S., & Staveley, M. (1999). Phrasier: A system for interactive document retrieval using keyphrases. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 160–167).

Kapalavayi, N., Murthy, S. N. J., & Hu, G. (2009). Document classification efficiency of phrase-based techniques. *IEEE/ACS International Conference on Computer Systems and Applications (AICCSA) 2009* (pp. 174–178). doi:10.1109/AICCSA.2009.5069321.

Koster, C., & Beney, J. (2009). Phrase-based document categorization revisited. In *Proceeding of the 2nd international workshop on patent information retrieval (PaIR)* (pp. 49–56).

Koster, C., Beney, J., Verberne, S., & Vogel, M. (2011). Phrase-based document categorization. In *Current challenges in patent information retrieval* (pp. 263–286).

Kwok, K., & Chan, M. (1998). Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 250–256).

Lang, H., Metzler, D., Wang, B., & Li, J. (2010). Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM)* (pp. 249–258).

Lavrenko, V., & Croft, W. (2001). Relevance-based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 120–127).

Lee, K., Croft, B., & Allan, J. (2008). A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 235–242).

Lewis, D., & Croft, W. (1990). Term clustering of syntactic phrases. In *Proceedings of the 13th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 385–404).

Li, Y., Luk, W., Ho, K., & Chung, F. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 797–798).

Lopez, P., & Romary, L. (2009). PATATRAS: Retrieval model combination and regression models for prior art search. In *Proceedings of the 10th cross-language evaluation forum conference on multilingual information access evaluation: Text retrieval experiments (CLEF)* (pp. 430–437).

Lupu, M., Mayer, K., Tait, J., & Trippe, A. J. (Eds.). (2011). *Current challenges in patent information retrieval. The Information Retrieval Series* (1st ed., Vol. 29).

Lv, Y., & Zhai, C. (2009). Adaptive relevance feedback in information retrieval. In *Proceeding of the 18th ACM conference on information and knowledge management (CIKM)* (pp. 255–264).

Magdy, W., & Jones, G. (2011). A study on query expansion methods for patent retrieval. In *Proceedings of the 4th workshop on patent information retrieval (PAIR)* (pp. 19–24).

Mahdabi, P., Andersson, L., Keikha, M., & Crestani, F. (2012). Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 505–514).

Maxwell, K., & Schafer, B. (2008). Concept and context in legal information retrieval. In *Proceedings of the 2008 conference on legal knowledge and information systems: JURIX 2008: The twenty-first annual conference* (pp. 63–72).

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2010). Overview of the patent mining task at the NTCIR-8 workshop. *NTCIR-8* (pp. 293–302).

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computer Survey*, *41*(2), Art. No. 10. doi:10.1145/1459352.1459355.

Navigli, R., & Velardi, P. (2003). An analysis of ontology-based query expansion strategies. In *Proceedings of workshop on adaptive text extraction and mining (ATEM), 14th European conference on machine learning (ECML)* (pp. 22–26). Cavtat-Dubrovnik, Croatia.

Piroi, F., Lupu, M., Hanbury, A., & Zenz, V. (2011). CLEF-IP 2011: Retrieval in the intellectual property domain. *CLEF 2011.*

Seo, H., Han, K., & Lee, J. (2011). CLEF-IP 2011 working notes: Utilizing prior art candidate search results for refined IPC classification. *CLEF Notebook Papers/Labs/Workshop.*

Vechtomova, O., & Karamuftuoglu, M. (2006). On document relevance andlexical cohesion between query terms. *Information Processing and Management, 42*(5), 1230–1247.

Vechtomova, O., & Karamuftuoglu, M. (2007). Query expansion with terms selected using lexical cohesion analysis of documents. *Information Processing and Management, 43*(4), 849–865.

Verberne, S., & D'hondt, E. (2011). Patent classification experiments with the linguistic classification system LCS in CLEF-IP 2011. In *Proceeding of: CLEF 2011 labs and workshop, notebook papers* (pp. 19–22). Amsterdam, The Netherlands.

Voorhees, E. (1994). Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 61–69).

Xu, J., & Croft, W. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 4–11).

Xu, X., & Croft, W. (2009). Transforming patents into prior-art queries. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 808–809).

Xu, Y., Jones, G., & Wang, B. (2009). Query dependent pseudo-relevance feedbackbased on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 59–66).

Yin, Z., Shokouhi, M., & Craswell, N. (2009). Query expansion using external evidence. In *Proceedings of the 31th European conference on IR research on advances in information retrieval (ECIR)* (pp. 362–374).