Analyzing Topic Drift in Query Expansion for Information Retrieval from a Large-scale Patent DataBase

Bashar Al-Shboul  
Department of Business Information Technology  
King Abdullah II School for Information Technology  
The University of Jordan  
Amman, Jordan  
bashar.shboul@gmail.com

Sung-Hyon Myaeng  
Division of Web Science,  
Department of Computer Science  
Korea Advanced Institute of Science and Technology  
Daejeon, South Korea  
myaeng@kaist.ac.kr

Abstract—Topic drift has been recognized as a major reason for ineffective retrieval of documents using query expansion. Topic drift is very important in patent domain as document processing is dependable on the quality of the search. With the huge amount of available patents, we investigate on the concept of topic drift in patent domain by analyzing the topics of retrieval results obtainable with query expansion. We define and utilize topic attributes to distinguish drifting topics from other topic classes such as Vanishing, Appearing, and Rising Topics. As a result, this analysis can be used to compare two different query expansion methods for their relative effectiveness.

Keywords—Topic Drift, Topic Scope Focus, Rising Topics, Drifting Topics

I. INTRODUCTION

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 and moving the original query toward the real intent. While QE is considered an important technique in IR, it suffers the problem of topic drift, that is, the change in the query’s topic to an unintended direction after QE. It is argued that topic drift in QE methods is caused by erroneous extraction of additional query terms [14], biased query term weighting method [2], biased retrieval model [1], or their combinations.

Previous studies have tried to reduce the effect of topic drift by proposing different methods for predicting the query effectiveness [5], or by estimating the query drift [12]. This work is different in that it focuses on understanding the reasons behind the general notion of topic drift instead of predicting the query effectiveness, and also showing the drifting topics rather than estimating the query drift. Furthermore, it is different in that previous works predict the query drift/effectiveness depending on the retrieved documents from searching using the query [12]; instead, we study the query drift based on the topics in the retrieved documents.

A common method of showing the reduction of topic drift caused by a QE method is through enhanced precision and recall [1, 14]. While some previous studies mentioned the problem of topic drift as a reason behind ineffective retrieval after QE, and generally, but vaguely, indicate that it happens, no one to the best of our knowledge has shown what actually happens in QE in terms of topical changes and how they can be characterized.

The aim of our work is to do an in-depth study of the topic drift problem by developing a method for analyzing the topics of retrieval results before and after QE and understanding the relationship between QE and topical changes.

Motivated by the absence of an in-depth study that investigates the relationship between topic drift and QE, and the absence of methods for showing driftiness within queries, the major contribution of this work is to define then measure topic attributes (i.e. Topic Effectiveness, Topic Scope Focus), use them to distinguish drifting topics, among the expanded queries, from other topic classes (i.e. Vanishing, Appearing, and Rising Topics), and then study their influence on the retrieval effectiveness.

To address the problem of topic drift, we aim at identifying the drifting topics among the expanded queries topics. First, similarity between original query topics and expanded query topics are measured, then used to identify several topic classes such as topics similar to each other (i.e. coherent topics), topics that appear among the expanded query topics while absent among the original query topics (i.e. appearing topics), and topics that appear among the original query topics; however, absent among the expanded query topics (i.e. vanishing topics). Further, topic attributes are preliminarily measured then utilized to define more detailed topic classes such as focused rising topics (i.e. specific topics that may improve the query effectiveness), broad rising topics (i.e. general topics that may enhance effectiveness), and drifting topics (i.e. topics that cause the query topic drift). Generally, topic effectiveness represents the retrieval effectiveness of the topic query (i.e. a query generated from all the topic terms), while topic scope focus measures the topic terms cohesiveness and the topic distinguishability from other topics.

In patent search, topic drift is considered a serious problem due to the high dependability of patent processing according to the quality of the search. Motivated by the high importance of search in patent domain, [2] proposed an automatic QE method that improved the effectiveness by semantically expanding...
query phrases for Patent Class Search (i.e. searching for relevant patent classes and assign them to the query patent). While our main purpose of this work is not devising a new QE method, we compare queries extracted and then expanded from patents using a phrase-based QE method (i.e. WQPE), and Relevance-based Language Model (RM), a state-of-the-art word-based QE method [10]. By using WQPE we aim at studying the effect of semantically expanding queries using phrases from Wikipedia, and their effect on the topic drift. On the other hand, comparisons are made against RM because it is one of the heavily researched techniques in QE, and it has often been used as a comparison benchmark [9, 3]. A unique characteristic of patent documents is the manual assignment of individual document to one or more classes from IPC (International Patent Classifications). IPC has a hierarchy of four levels (Section, Class, Sub-Class (SC), and Group) with the “section” being the most general. 1 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 “G06F 17/30”, its SC, MG, and SG are “G06F”, “G06F 17”, and “G06F 17/30” respectively. A group of patents belonging to a class are said to be relevant to each other at different levels of abstraction. The existence of the IPC hierarchy allows us to study generalization/specialization of queries. The retrieval task is thus to classify patents to those classes of the query as closely as possible, and the proposed query expansion approaches (i.e. WQPE and RM) were formally tested on the US Patent & Trademark Office (USPTO) patents provided for several NTCIR tasks.

This paper is organized as follows. In section 2 we list related work. Proposed methods are described in section 3. In section 4 evaluation and experimental results are discussed. Finally, we conclude with section 5.

II. RELATED WORK

Much research has discussed the topic drift problem in QE by referring to it as a reason for ineffective retrieval [2, 1, 14]. Studies have tried to reduce the effect of topic drift by proposing different query performance prediction methods for predicting the query effectiveness pre-/post-retrieval [5], or by estimating the query drift [12]. Generally, prediction methods depend on scoring queries in different aspects according to the query terms and their probabilities or statistics in the collection or the retrieved document set (i.e. Clarity [16], Robustness [15], etc.) [5]. In [8], topic drift is measured through divergence between timely grouped documents. Our work is different in that topic drift is not defined differently from this work (i.e. in this work topic drift is defined as topical changes over time).

Other works concentrated on estimating topic drift in expanded queries as in [12]. Mainly, it is assumed that a high deviation in retrieval scores between the retrieved sets from an initial query and its expanded version is correlated to the reduction in topic drift. This work is different in that we focus on understanding the reasons behind the topic drift, and investigating the behavior of QE methods in causing queries to drift topics rather than estimating drift amount according to retrieved set of documents.

In [11], topic drift is introduced in expanding queries using expert profiles. Given a query, a set of ranked experts are selected to enrich the query with terms from their profiles (i.e. using terms in their research interests) and then run the query again to retrieve another ranked list of experts. Topic drift is measured in expert profiles through “cohesiveness” trying to predict the number of various expertise contained in the expert profile, and then decide which ones to use as candidate expansion terms. This attempt can be considered as a pseudo-relevance feedback like method of QE utilizing some parts of retrieved expert profiles after predicting their relevant expertise. Our work is different in that our intentions are not devising a new QE method, and that our goal is to study the effect of the existing QE methods on the topics generated of the retrieval set, in addition to studying the effect of the selected expansion terms on enhancing effectiveness.

III. PROPOSED METHODS

A topic in our work is defined as a set of words describing a piece of coherent information. In this section we describe the method with which topics are extracted from the original query (i.e. a patent document) and the expanded query. After that is a description of a method for similarity calculation between topics generated before and after QE so that we can identify vanishing and appearing topics. In addition, we define two attributes of a topic, Scope Focus and Effectiveness, by which an appearing topic can be further classified into drifting and rising topics:

A. Generating Topics

In order to generate topics from a query, we employ Hierarchical Dirichlet Process (HDP) [13], a very well-known method of identifying topics from a document collection. It is an extension of Latent Dirichlet Allocation (LDA) [4] that assumes a set of documents consist of several themes (i.e. topics), which are obscured by word-choice noise, and finds topic distributions over a document and term distributions over topics in a document. While LDA has been used for various applications, it suffers from a major problem of having to decide the number of topics a priori. In HDP, however, the number of topics is generated automatically. Additionally, HDP adopts the assumption that documents share one discrete distribution that can be viewed, in our case, as a very general topic with incoherent terms while the sub-distributions of it are sampled in such a way that each topic terms are distributed according to a Dirichlet distribution.

Since a patent used as a query may not be long enough to apply HDP, we first extend it by finding its IPCs and collecting the patent documents belonging to them to form a document pool that represent the query. This process makes sense because the IPCs represent the semantics of the query patent and the documents having the same IPCs can be considered belonging to the same semantic categories. We call the result Original Query Topics (OQTs). In order to analyze what
topical changes are made by an expanded query, on the other hand, we generate Expanded Query Topics (EQTs) by applying HDP to the set of documents retrieved by the expanded query. We applied HDP to the retrieved documents because it was not possible to identify IPCs from the expanded query per se, nor amenable to apply HDP directly to the expanded query.

Currently, the only complete relevance judgment for evaluating effectiveness in patent domain is by using IPCs, therefore a method assigning IPCs to topics have to be proposed in order to find a way to evaluate topics. In our work, IPC-to-Topic assignment takes the form of ranked list of IPCs returned from searching using a topic query (i.e. a query generated from all the topic terms). The best ranked IPCs among the retrieved patents of the topic query are assigned for the topic.

B. Similarity

While several topic similarity/dissimilarity methods have been proposed in the past such as KL-Divergence, Cosine Similarity, Jaccard’s Coefficient, we opt for JS-Divergence that has shown its superiority in measuring topic similarity [4, 7] over others. As a word-based topic similarity method, dissimilarity between two topics in JS-Divergence is given by:

\[
JS(t_1, t_2) = \frac{1}{2} (KL(t_1||M) + KL(t_2||M))
\]

where

\[
M = \frac{1}{2} (t_1 + t_2)
\]

where \(t_1\) and \(t_2\) are topics, KL represents the Kullback-Leibler Divergence between two probability distributions P and Q given as:

\[
KL(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i}
\]

Due to the frequent use of non-identical but synonymous terms in patent documents, using words alone is likely to measure dissimilarity between topics incorrectly, that is, some topics may contain similar but synonymous terms, and yet be considered dissimilar. Thus, we have proposed another method for topic dissimilarity that includes IPCs. The goal is to detect the dissimilarity between the topics on a semantic level (i.e. terms). A possible way of measuring similarity between topics is utilizing IPCs assigned per topic; hence, measuring the dissimilarity between two ranked lists of IPCs. IPC similarity for utilizing IPCs assigned per topic; hence, measuring the dissimilarity between the topics on a semantic level (i.e. terms).

One important property for this similarity is that it is asymmetric, meaning that:

\[
Sim_{\text{mix}}(t_1, t_2) \neq Sim_{\text{mix}}(t_2, t_1)
\]

This is due to the different IPC ranks in their assigned IPC lists. A possible modification to make this relation symmetric is by measuring the total divergence to the average topic mixture similarity, as below:

\[
\text{Dissim}_{\text{mix}}(t_1, t_2) = \frac{1}{2} (KL(Sim_{\text{mix}}(t_1, t_2)||M) + KL(Sim_{\text{mix}}(t_2, t_1)||M))
\]

where

\[
M = \frac{1}{2} (Sim_{\text{mix}}(t_1, t_2) + Sim_{\text{mix}}(t_2, t_1))
\]

As dissimilarity between topics could be measured according to terms in addition to IPCs, a hybrid method to combine both dissimilarities is proposed as follows:

\[
\text{Dissim}_{\text{hybrid}}(t_1, t_2) = \lambda JS(t_1, t_2) + (1 - \lambda)\text{Dissim}_{\text{mix}}(t_1, t_2)
\]

where \( \lambda \) is a weighting factor.
C. Topic Attributes

In order to characterize a topic for its goodness and hence determine whether it will be a drifting topic to be suppressed or a rising topic to be retained in the expanded query, two attributes for topics are introduced: topic effectiveness and topic scope focus.

1) Topic Effectiveness

It is defined to be the retrieval effectiveness of a topic when it is used as a query. It is devised to evaluate an individual topic for its relative contribution to effectiveness of retrieving documents based on the original information need. This measure helps determining whether a topic would contribute to a better query after expansion.

Note that this attribute is used to evaluate topics generated out of an expanded query only in retrospect, for the purpose of evaluating a QE method. In reality, however, this information would not be available for individual topics as there would be no relevance judgments for a query or a topic.

2) Topic Scope Focus

A role of QE is to disambiguate the original query. We argue that query disambiguation with QE is closely related to narrowing the potential scope of the query or moving it toward a position in a vector space that represent the true information need. This can be done by adding terms more closely related to the information need or adjusting the term weights.

Given our premise that the more focused a topic is, the better, we attempt to measure topic scope focus for an identified topic based on its similarity to the original query topics and at the same time its independent cohesiveness that can be measured by the characteristics of the terms representing the topic.

More precisely, topic scope focus is measured by cohesiveness of the terms representing the topic and their dissimilarity to those in other topics. Assuming that the term probability to a topic is the membership of that term to a hypothetical cluster, and that the dissimilarity between two topics can be measured with the hybrid dissimilarity method defined earlier, a scope focus function for a topic can be defined as:

\[
Scope(t) = 1 - \frac{\sum_{w \in t} P(w|t)}{|t|} - \frac{\sum_{\text{set}, \text{set}} \text{Dissim}_{	ext{Hybrid}}(\{t\} \text{set})}{|T| - 1}
\]

where \(\text{Dissim}_{\text{Hybrid}}\) is the hybrid topic dissimilarity defined earlier while \(T\) is the set of all topics. The first part of the equation represents the average cohesiveness in the topic \(\sum_{w \in t} P(w|t)/|t|\) measuring the average of term probabilities in the topic. We assume that a cohesive topic has terms with a higher average probability. The second part shows the average distinguishability of a topic from others. That is, a focused topic is distinguishable if it is highly dissimilar to other topics generated from the original query.

The \(\text{Scope}\) score will fall in the range of \([0, 1]\), where the zero indicates a very distinguishable topic with rare incoherent terms or a topic with very coherent; however commonly shared, terms. The best \(\text{Scope}\) in this case is closer to one\(^3\), meaning that the topic is highly coherent, and highly distinguishable from other topics, as well.

D. Topic Classes

Since not all topics are equally important or exist before and after query expansion, we need to classify them accordingly. Some topics may be preserved from the original queries while others newly appear or disappear after the expansion process. Following are the details of how different topic classes are defined and used.

1) Coherent Topics

Two topics are said to be coherent to each other when their similarity is high enough. By measuring coherency between topics, we can identify the topics that remain in the expanded query and at the same time determine the degree to which the changes have occurred to the original topic or the degree to which the content remains in the preserved topic. The main reason behind studying coherent topics is to identify the changes among topics, which can help identifying query topic drift.

Depending on the value of a coherency threshold, the numbers of various types of topics will vary. As such, determining a threshold value would influence the query expansion process and hence overall effectiveness of an expanded query. That is, it is an optimization problem. In this study, however, the goal is to clearly explain the topic drift issue in QE and therefore various threshold values are examined empirically to see their effects.

2) Vanishing Topics (VT)

Original query topics (OQTs) are said to be vanishing if they show incoherency with all the topics identified from an expanded query or EQTs. We hypothesize that VTs can cause the original query to be ambiguous, and thus removing them is supposed to enhance the resulting query’s retrieval effectiveness. Since determining VTs are affected by the coherency threshold, the hypothesis can be proven or rejected depending on the threshold values we examine. Conversely, we can determine the threshold value by measuring how much effectiveness we gain by identifying VTs. We empirically examine the effect of eliminating VTs based on various coherency threshold values. In this work, experiments on VTs have been omitted for brevity.

3) Appearing Topics (AT)

An EQT is said to be appearing if it is incoherent to all the OQTs. While ATs can be new topic introduced by the new terms in the expanded query, they can be latent topics that existed in the original query but buried with other stronger terms, which emerge with the aid of the new terms in the expanded query.

ATs generally have different importance as generating topics does not care whether topics are coherent or not \([6]\), alternatively, it focus on generating topics according to assumed term and topic distributions (i.e. Dirichlet), or

\(^3\)Even though the zero score can be achieved by in-cohesive and common topics we rely on HDP to sample terms of general topics out, and that dissimilarity between topics is generally high (i.e. in average over 0.45)
optimizing likelihood-based measures. Therefore, an AT may be a very specific topic that disambiguates one important aspect of the information need. However, it is also possible that it is a drifting topic that directs the original information need in a totally incorrect direction. As a result, we can classify ATs into two: rising and drifting topic categories. Rising Topics are supposed to increase effectiveness of the query whereas drifting topics do harm on effectiveness of the query. Determining a threshold dynamically for an expanded query is difficult and beyond the scope of this paper; in this work, we examine different threshold values empirically and attempt to show this notion of drifting topics would help determining goodness of an expanded query and hence improving retrieval effectiveness.

IV. EVALUATION

We now present the results of our experiments on different topic classes. Mainly, we aim at showing that Rising Topics are helpful for generating a more effective query than the original query or the OQTs. In addition, we aim at showing that the concept of topic focus is helpful in selecting topics among the rising topics to form an even more effective query than the rising topics only query.

We have randomly sampled 103 queries (patents) from the NTCIR-6 USPTO\(^4\) collection. As the number of documents used to generate topics has to be reasonable, we made sure that each IPC associated with a patent included in the collection contains at least 100 documents. For generating topics from both original and expanded queries, we used HDP that is configured to stop at 100 iterations on all queries with 10 burn-in iterations, while all other HDP parameters set to the values listed in [13]. For fair comparisons between an original query and its expanded version, the same number of documents was selected for generating topics for both.

In order to show that the analysis results depend on different QE methods, we employed two: Wikipedia-based Query Phrase Expansion (WQPE) [2] and Relevance-based Language Model (RM) [10] methods. While RM was chosen because it is a strong probabilistic model and has been heavily used as the benchmark of QE methods, WQPE is a phrase-based QE method that has proven to be more effective than RM in patent class search. It is interesting to note that both methods have generated much more topics on average (153 topics for WQPE, and 168 topics for RM) than the original queries (11.3 topics). A possible reason is that expanded query terms may retrieve documents from various contexts, which would generate additional topics. Since each of the original queries was quite coherent by itself and the documents added to it all share the same IPCs, on the other hand, it is natural that HDP recognizes fewer topics.

Our major findings in this paper are summarized in Table 1. Showing three different sets of results (Original Queries, EQTs after eliminating drifting topics, and EQTs after eliminating drifting topics and then eliminating unfocused topics) over three different patent classification levels (SC, MG, and SG), the table shows the effect of eliminating drifting topics showing an effectiveness improvement in higher classification levels. By eliminating unfocused topics in addition to drifting topics, effectiveness improved on all classification levels to reach at least 13.97% and 11.72% for WQPE and RM respectively, showing higher improvement over EQTs with drifting topics eliminated (13.09% for WQPE compared to 11.28% for RM).

<table>
<thead>
<tr>
<th>Queries (Baseline)</th>
<th>RM</th>
<th>WQPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>50.38</td>
<td>52.88</td>
</tr>
<tr>
<td>MG</td>
<td>34.26</td>
<td>35.87</td>
</tr>
<tr>
<td>SG</td>
<td>27.9</td>
<td>28.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EQTs after eliminating Drifting Topics</th>
<th>RM</th>
<th>WQPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>(6.43%)</td>
<td>(4.75%)</td>
</tr>
<tr>
<td>MG</td>
<td>(4.32%)</td>
<td>(5.10%)</td>
</tr>
<tr>
<td>SG</td>
<td>(0.39%)</td>
<td>(0.78%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EQTs after eliminating Drifting Topics then eliminating Unfocused Topics</th>
<th>RM</th>
<th>WQPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>(18.44%)</td>
<td>(20.39%)</td>
</tr>
<tr>
<td>MG</td>
<td>(21.75%)</td>
<td>(17.95%)</td>
</tr>
<tr>
<td>SG</td>
<td>(11.2%)</td>
<td>(13.97%)</td>
</tr>
<tr>
<td></td>
<td>(11.28%)</td>
<td>(13.09%)</td>
</tr>
</tbody>
</table>

As our focus in this paper is identifying drifting topics and showing the effect of removing them, we have conducted a series of experiments trying to identify drifting topics among other topic classes, and then investigate the relationship between effectiveness and the elimination of drifting topics from the expanded queries to prove that it will lead to enhancement in retrieval effectiveness. After several preliminary experiments on coherent and vanishing topics we have realized that expanded query drifting topics exist among appearing topics only, thus further experiments are performed on appearing topics.

For each of the appearing topics we measured effectiveness by generating a query from the topic terms. As individual topics achieved low effectiveness results we have decided to divide appearing topics according to their effectiveness and find the most effective amount of appearing topics that will contribute towards generating an effective query. An important act as been concluded that is, by eliminating 75% of the appearing topics, the remaining topics will generate better effectiveness than using all the topics generated from the expanded query. It has been also concluded that those 75% can be safely recognized as drifting topics, as the higher percentage of topics considered among them, the lower obtained effectiveness. On the other hand, the other 25% of the topics

---

are recognized as the rising topics. Complete results of this experiment were omitted for brevity.

In order to examine the topic focus importance, rising topics were further sorted according to their focus score, and various percentages have been tested. It has been concluded that selecting the top 25% focused topics among the rising topics will produce a query with higher effectiveness than using rising topics only, or even the original query. Additional experiments on the rising topics, by trying smaller percentages of the most focused topics (i.e. 18.75%, 12.5%, and 6.25%), has been made; however, selecting over-focused topics has proven to harm the reported effectiveness. Complete experiments and their full results were also omitted for brevity.

V. CONCLUSIONS

We have proposed a new framework where topics of expanded queries can be analyzed. In particular, we showed how drifting topics can be identified in experiments and demonstrated the negative role they play in retrieval effectiveness and positive role of rising topics. It has been proved that topic pairwise similarities and topic attributes can be significantly useful in identifying different topic classes. The main value of the work in this paper lies in the finding of different topic types and the possibility of analyzing expanded queries and selecting useful topics for retrieval effectiveness.

An obvious limitation in our work is the use of topic effectiveness assuming that we have the relevance judgment, basically because our main goal is to study the drifting topics. However, we believe that finding a replacement for predicting the topic effectiveness represents a challenge for our future work. Additionally, we aim at utilizing our analysis for devising a new QE method that utilizes topics instead of ranking documents or terms for expanding queries.

ACKNOWLEDGMENT

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-3005) supervised by the NIPA (National IT Industry Promotion Agency).

REFERENCES


