Front. Comput. Sci., 2019, 13(5): 1048–1061
https://doi.org/10.1007/s11704-018-7056-6

RESEARCH ARTICLE

Patent expanded retrieval via word embedding under composite-domain perspectives

Fei WANG[1,2], **Tieyun QIAN** (✉)[1,2], **Bin LIU**[1,2], **Zhiyong PENG** (✉)[1,2]

1 School of Computer Science, Wuhan University, Wuhan 430072, China
2 State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China

Abstract Patent prior art search uses dispersed information to retrieve all the relevant documents with strong ambiguity from the massive patent database. This challenging task consists in patent reduction and patent expansion. Existing studies on patent reduction ignore the relevance between technical characteristics and technical domains, and result in ambiguous queries. Works on patent expansion expand terms from external resource by selecting words with similar distribution or similar semantics. However, this splits the relevance between the distribution and semantics of the terms. Besides, common repository hardly meets the requirement of patent expansion for uncommon semantics and unusual terms. In order to solve these problems, we first present a novel composite-domain perspective model which converts the technical characteristic of a query patent to a specific composite classified domain and generates aspect queries. We then implement patent expansion with double consistency by combining distribution and semantics simultaneously. We also propose to train semantic vector spaces via word embedding under the specific classified domains, so as to provide domain-aware expanded resource. Finally, multiple retrieval results of the same topic are merged based on perspective weight and rank in the results. Our experimental results on CLEP-IP 2010 demonstrate that our method is very effective. It reaches about 5.43% improvement in recall and nearly 12.38% improvement in PRES over the state-of-the-art. Our work also achieves the best performance balance in terms of recall, MAP and PRES.

Keywords patent retrieval, composite-domain perspective, double-consistency expansion, word embedding

1 Introduction

Patents have become a normal way for companies or organizations to protect investments and pursue interests. After the identification of unique technology, innovation could be granted a valid patent. The purpose of patent prior art search is to identify the uniqueness of patent technology. Namely, its goal is to prove the idea of innovation which has not been granted a patent or published in other scientific papers.

Different from Web document, a patent is a kind of semi-structure document containing a large number of technical terms. The identification of unique technology for the innovation needs massive mental work. Hence, patent prior art search attracts extensive attentions [1–13].

Challenges in patent prior art search are as follows: 1) Large quantity. According to statistics released by WIPO (World Intellectual Patent Organization) in 2016, the number of application of invention patents, one of three kinds of intellectual property rights, has reached 1,101,864, showing a 18.7% growth than that in 2015. 2) High recall. Patent prior art search aims to guarantee no infringement between the idea of innovation and any previous granted patents. Otherwise, it might result in a lawsuit of million dollars. 3) Strong ambiguity. The applicants have obscure style of writing and use the synonyms and hypernym-hyponyms, hence unusual terms are

Received February 16, 2017; accepted September 8, 2017

E-mail: peng@whu.edu.cn; qty@whu.edu.cn

used to express common semantics while uncommon semantics come from usual terms [14]. 4) Information dispersal. Patent is a semi-structure document containing *title*, *abstract*, *description*, *claim* and does not represent a focused information need. On the contrary, retrieval system uses the focused information need of keywords rather than documents to obtain relevant patent documents.

Patent prior art search generates patent queries with technical terms according to patent document. It contains patent reduction and patent expansion. Patent reduction selects core terms from query patents to generate queries while patent expansion strengthens retrieval semantics and reduces ambiguity by supplementing relevant terms.

Early works take a query patent as a single topic model for patent reduction. Those methods extract terms from the query patent through TF-IDF or the transformation of TF-IDF [2, 3]. Patent reduction of a single topic model could not cover all the technical characteristics of query patents. More recent works study the problem of diverse aspects or subtopics of patents [4–6], and propose to use clustering algorithms or term frequency to generate multiple queries. However, such queries can not present explicit aspect interpretations.

After the removal of stop words, about 12% of relevant patents have no shared terms with corresponding query patents [15]. Patent expansion is introduced to enrich retrieval semantics by adding relevant words. The related work for patent expansion has two expanded strategies of semantics [7, 8] and distribution [9–13, 16]. Semantic expansion uses repositories as external resources, which emphasizes semantic consistency and ignores distributed consistency. Distributed expansion expands a patent query by adding terms from pseudo-relevant patents or cited patents, which pursues distributed consistency and ignores semantic consistency. Hence, existing works split the relevance between distribution and semantics of terms. Furthermore, common repositories, such as WordNet or Wikipedia, collect usual words and common semantics, and they hardly meet the requirement of patent expansion for domain-aware terms [12].

In this paper, we propose a novel model to generate aspect queries, which is inspired by the finding that diverse aspects of patents have strong correlation with their classifications. In Fig. 1, a query patent (QP) has technical relevance with other three patents (SP1, SP2, SP3) which all come from evaluation set of CLEP-IP 2010. The query patent expects to obtain other three patents by patent prior art search. QP is classified into *section* C and *section* G. SP1 is classified into *section* B and *section* C. SP2 is classified into *section* B and *section* G. SP3 is classified into *section* C and *section* G. Obviously, QP and SP1 have the technical similarity on *section* C, QP and SP2 on *section* G. In contrast, QP and SP3 have the technical similarity both on *section* C and G simultaneously. Instead of clustering algorithm or term frequency, we use patent classifications to clear the boundaries for different query aspects. In order to accurately express different combinations of patent classifications, we propose the concept of *composite-domain perspective*. We generate a aspect query corresponding to the query aspect based on a specific *composite-domain perspective*.

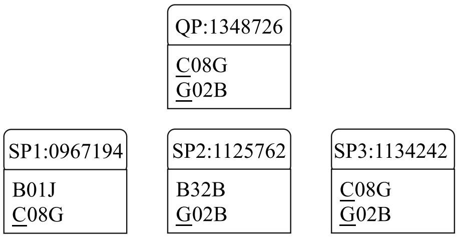

Fig. 1 Similarity among the query patent and similar patents

In order to obtain domain-aware candidate terms, we use the special domain resource to train semantic vector spaces by word embeding which has been widely used in semantic modeling. Besides, we are the first to implement double consistency expansion which takes semantics and distribution simultaneously into consideration to reduce semantic ambiguity. Finally, we design a fusion method based on the perspective weight and the rank of the retrieval patents to merge multiple retrieval results.

2 Related work

The retrieval topics of patent prior art search are full applications with fields of *title*, *abstract*, *description*, *claim* and so on. In previous works, researchers discussed different sections of a query patent to implement patent reduction. Some of previous works reported that fields of *title*, *abstract* and first *claim* contain the most concentrated core of technical characteristics. And patent queries of other fields obviously obtain a better performance because fields of *claim* and *description* have more technical details [2, 3, 17].

Patent reduction is to shorten a query patent and find a focused information need by removing the ambiguous and noisy terms. It has been studied for a few years. TF-IDF and its transformation have been widely applied into patent reduction [2, 3] which consider high term frequency in the query patent document and low term frequency in other patent documents to be a strong indicator of a good query term. Pseudo relevance feedback has also been used to reduce patent queries by removing the most dissimilar segments to the pseudo-relevant documents from the query patent [18].

Recently, some researchers [4–6] studied the diverse aspects or sub-topics on patent reduction. Kim et al. [4] used a decision tree to generate diverse queries while another work [5] made clustering algorithms group the terms to identify diverse query aspects. Far et al. [6] reported that a simple, minimal interactive relevance feedback approach where terms are selected from only the first retrieved relevant document with higher term frequency than retrieved irrelevant documents obtains the best result.

Patent terms have strong ambiguity. In order to reduce the ambiguity, previous works use external resources to expand queries. They usually adopt two strategies of semantics and distribution. The semantic expansion uses repositories such as Wikipedia [7] and WordNet [8] as external resources, which expands the queries with the synonymic and hypernym-hyponym words related to the query terms. The distributed expansion uses pseudo-relevant documents [9] as external resources and alleviates the query ambiguity by adding the query with distributed-relevant words.

Other studies also applied different external resources into distributed expansion [10–13, 16]. Mahdabi et al. [10] proposed to automatically disambiguate query terms by employing noun phrases which are extracted by using the global analysis and introduced a method to predict whether expansion based on the noun phrases would improve the performance. Previous works exploited proximity information to select expanded terms from a query-specific patent lexicon built based on different resources such as definitions of the International Patent Classification (IPC) [11, 16] or relevant patents with similar IPC code to the query patent [12]. Another recent work [13] built a time-aware weighted network based on patent citation and used a random worker to find influential documents as candidate expanded resource.

Our approach is based on patent domain and word embedding. The universal patent classified domain is International Patent Classification (IPC) which is currently being used by more than 100 patent-issuing bodies, such as patent office of Europe, USA, China and Japan. Word embedding is an unsupervised training method which can be applied to any type of text. Hence, our approach is not related to the special retrieval system and can be effectively applied to other patent datasets, such as Chinese patents or American patents.

3 Construction of patent queries

In this section, we introduce our work in details. Figure 2 presents the framework. We first perform pseudo relevance feedback (PRF) to get potential relevant patents for building composite-domain perspective converters and generate aspect queries for perspective retrieval. Secondly, we train the semantic vector space via word embedding under single-domain perspectives and perform expanded retrieval by taking semantics and distribution into consideration simultaneously. Finally, we propose a fusion method to merge multiple retrieval results based on perspective weight and patent rank in the perspective result. Table 1 is the symbol description.

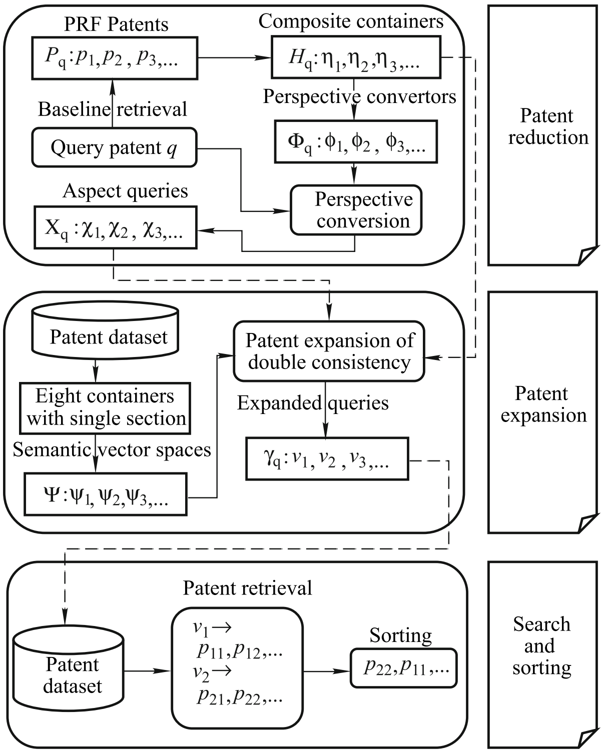
PRF Patents
P_q: p_1, p_2, p_3,...
Composite containers
H_q: η_1, η_2, η_3,...
Baseline retrieval
Perspective convertors
Query patent q
Φ_q: ϕ_1, ϕ_2, ϕ_3,...
Aspect queries
Perspective conversion
X_q: χ_1, χ_2, χ_3,...
Patent reduction
Patent dataset
Patent expansion of double consistency
Eight containers with single section
Expanded queries
Semantic vector spaces
γ_q: v_1, v_2, v_3,...
Ψ: ψ_1, ψ_2, ψ_3,...
Patent expansion
Patent retrieval
Patent dataset
$v_1 \rightarrow$ p_{11}, p_{12},...
$v_2 \rightarrow$ p_{21}, p_{22},...
Sorting
p_{22}, p_{11},...
Search and sorting

Fig. 2 Patent retrieval workflow under composite-domain perspectives

3.1 Definitions

In this section, we define some associated terms to create common context for our retrieval models.

Definition 1 (Patent domain) *Patent domain*, namely patent classified domain, refers to the patent technical classification of one patent. In this paper, we take IPC codes of patents as patent domains (more details in Section 3.2.2). In Fig. 1, the patent domains of QP are C08G and G02B.

Definition 2 (Perspective) *Perspective*, namely composite-domain perspective, is generated based on the *section* of IPC code. In Fig. 1, the *sections* of patent domains of QP are C(C08G) and G(G02B), then composite-domain perspectives of C, G and CG will be created. In order to underline a composite-domain perspective which contains only one *section*, we introduce the concept of *single-domain perspective*, such as C and G.

Table 1 Symbol description table

Symbol	Description
Q, q	Query patent set Q, $q \in Q$
P, p	Relevant patent set P, retrieved by a query patent, $p \in P$
M, m	Patent *section* domain set M, eg:$\{A, B\}$
G, g	Patent fine-grained domain set G, eg:$\{A01B\quad 11/02, \cdots\}$
A, α	Index set of composite-domain perspectives based on a topic A, eg:$\{A, B, AB\}$, $\alpha \in \mathrm{A}$
B, β	Index set of single-domain perspectives based on a topic B, eg:$\{\ A, B\ \}, \beta \in \mathrm{B}$
K, κ	Index set of single-domain sections/perspectives based on dataset. eg:$\{A, B, \ldots, G, H\}$, $\kappa \in \mathrm{K}$
H, η	Container set of composite-domain perspectives to store relevant patents, eg:{A={}, B={}, AB={}}, $\eta \in \mathrm{H}$
Φ, ϕ	Perspective converter set Φ, eg:$\{A{=}\{\},B{=}\{\}\ AB{=}\}$, $\phi \in \Phi$
Ψ, ψ	Semantic vector space set Ψ, eg:$\{A{=}\{\}, B{=}\{\}\}$, $\psi \in \Psi$
X, χ	Aspect query set X, $\chi \in \mathrm{X}$
Υ, υ	Expanded query set Υ, $\upsilon \in \Upsilon$
w	A term of query patents or relevant patents
c	A expanded term from a semantic vector space

Definition 3 (Query aspect) *Query aspects* are decided by the diverse aspect models, such as decision tree or clustering algorithm. One cluster center represents one query aspect in cluster algorithm while our model generates query aspects based on composite-domain perspectives. One query aspect corresponds to one composite-domain perspective and can generate one aspect query. In Fig. 1, QP generate three query aspects of C, G and CG, then three corresponding aspect queries will be created.

Definition 4 (Shared section) *Shared section* refers to the *section* of IPC code of the shared patent domain. In Fig. 1, the patent domains of QP are C08G and G02B while the patent domains of SP1 are B01J and C08G, then the shared *section* is C.

3.2 Composite-domain perspective model

In this section, we first discuss the shortcomings of existing works on the quantitative technical similarity and propose our perspective quantitative strategy. We then introduce our methods to construct composite-domain perspective converters and implement perspective conversions for query patents.

3.2.1 Measuring the technical similarity of patents

Patent prior art search generates patent queries based on quantitative strategy of patent technical similarity. Patent search normally has two quantitative strategies, i.e., the single-topic model [2,3] and multiple-topic model [4–6]. The single topic strategy takes a query patent as an indivisible technical unit and generates a single query for patent retrieval. Multiple topic strategy thinks that every patent has more than one technical units and generates multiple aspect queries. Compared with the single topic strategy, multiple topic strategy is closer to the way people think about technology. However, multiple topic strategy uses global statistics such as co-occurrence terms and high term frequency to cluster terms as queries. These queries can not find explicit semantic interpretations and not represent independent technical units.

In section one, we have reported our findings that patent technical similarity is related to patent classification. We find that some patents have more than one technical classified domains, and some relevant patents with different technical classified domains only share a part. In Fig. 1, QP and SP1 are comparative patents with technical similarity. QP is classified into C08G and G02B while SP1 is classified into B01J and C08G. QP and SP1 have the shared classified domain of C08G. We try to measure technical similarity based on a part of shared classified domains of comparative patents and propose a concept of *composite-domain perspective* which encourages to generate aspect queries from different technical perspectives.

Current retrieval framework takes a query or an indexed document as a indivisible unit of semantic interpretation, which measures the similarity between a query and a document based on the union of technical classified domains rather than the intersection.

In Fig. 3, query patent *QP* has technical similarity with three relevant patents *SP* on different classified domains (Fig. 3(a)). Supposing that different classified domains of patents have equivalent technical characteristics. In the single topic model, a retrieval system uses smooth technology to align the query and document, which reduces the similarity value for which are measured on the union of technical classified domains rather than the intersection, such as QP and SP1, QP and SP2 (Fig. 3(b)). We use the composite-domain perspectives to segment the technical characteristics according to patent classifications and generate multiple aspect queries such as QP1, QP2 and QP3 (Fig. 3(c)). Composite-domain perspective model implements the aspect retrieval by measuring the technical similarity on the shared classified domains. In Fig. 3(d), the technical similarity of SP1 and QP1 are measured on the shared section A, SP2 and QP2 on section C, and SP3 and QP3 on both section A and C. They will improve the rank of relevant patent SP1, SP2 and SP3 in the retrieval results compared to the rank in the single topic model.

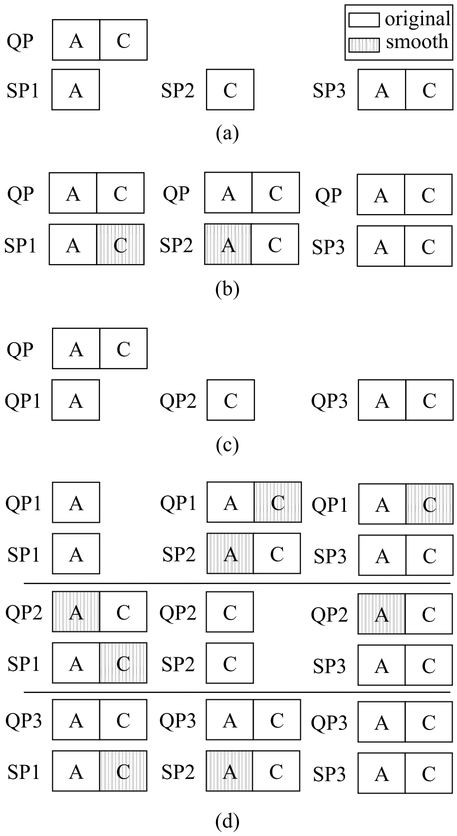

Fig. 3 Different quantitative strategies of technical similarity. (a) A topic with three relevant patents; (b) similarity of the single topic model; (c) generation of aspect queries; (d) similarity of the composite-domain perspective model

3.2.2 Constructing aspect queries

We first determine the basic unit of classified level for composite-domain perspectives. International Patent Classification (IPC) is the universal classification of patent applications. The IPC hierarchical system arranges the IPC codes in a tree-like structure with five components of *section*, *class*, *subclass*, *group* and *subgroup*. For example, $A01C$ 12/14 is a patent classification of IPC code which has multiple granularity, such as *section*(A), *class*($A01$), *subclass*($A01C$), *group*($A01C$ 12) and *subgroup*($A01C$ 12/14).

We takes *section* as the basic unit of classified level for the perspective because *section*, which is the division of technical industry, usually has a greater distinction. A single-domain perspective is generated based on only one section while a composite-domain perspective corresponds to one or more sections. Patent similarity will be measured under the common perspective. When a query patent has more than one sections, similarity will be measured under more than one different perspectives. For example, if a query patent has two sections of A and B, the technical characteristic will be measured under three perspectives of A, B and AB.

A patent document presents all the works of technical innovation which probably has more than one technical domains. We construct perspective converters to reveal the technical characteristics under the different composite-domain perspectives.

Though *section* is the basic unit to quantify technical characteristics, perspective converters are constructed based on more fine-grained classification, such as *subgroup*, because a *section* still contains massive fine-grained classifications. The method to construct perspective converters for composite-domain perspectives is presented in Algorithm 1.

Algorithm 1 Construction of perspective converters for composite-domain perspectives

```
Input: query patent ($q$)
Output: perspective converters ($\Phi_q$)
1: obtain potential relevant patents $P_{top-k}$ through baseline retrieval based on $q$;
2: obtain section domain set $M_q$ and fine-grain domain set $G_q$ of $q$;
3: generate index set of composite-domain perspective $\mathrm{A_q}$ based on section domain set $M_q$;
4: for each index $\alpha \in \mathrm{A_q}$ do
5:     generate a container to store relevant patents, $\eta_\alpha \rightarrow \mathrm{H}$;
6: end for
7: for each relevant patent $p \in P_{top-k}$ do
8:     obtain section domain set $M_p$ and fine-grain domain set $G_p$ of the $p$;
9:     for each index $\alpha \in \mathrm{A_q}$ do
10:        if $M_p \cap M_q \in \alpha$ and $G_q \cap G_p \neq \emptyset$ then
11:            $p \rightarrow \eta_\alpha$;
12:        end if
13:    end for
14: end for
15: for each container $k_\alpha \in K$ do
16:     operate formula (1),(2) to construct perspective converters based on $k_\alpha$ and implement normalization $\overline{\phi_\alpha} \rightarrow \overline{\Phi}$;
17: end for
```

The procedure of Algorithm 1 is illustrated as follows. We perform baseline retrieval and obtain $TOP-K$ relevant patents as document resources to construct perspective converters for composite-domain perspectives (line 1). Based on *section* level classification, we generate the index set of composite-domain perspectives and containers to store relevant patents (lines 2–6). Relevant patents are distributed into a container based on their *section* level classification and fine-grained level classification (lines 7–14). A relevant patent is distributed into a container, which means that the relevant patent and the container meet two conditions simultaneously as follows: 1) the shared *section* level classification of the query patent and the relevant patent is a part of the container. 2) the query patent and the relevant patent share a part of fine-drained level classifications. Finally, our approach to construct perspective convertors is the following. First, we estimate the relevance of a patent in the container with the composite-domain perspective in Eq. (1). Then we use those relevant patents from a container to generate the language model as a perspective convertor in Eq. (2) (lines 15–17).

After the distribution of relevant patents according to the technical classified domains, each container obtains some relevant patents. We evaluate the relevance between relevant patents and technical characteristics of composite-domain perspectives using in Eq. (1).

$$P(p|\theta_\alpha)=\begin{cases} H(\theta_p,\theta_{\alpha+\Delta}) - H(\theta_p,\theta_\alpha), & \alpha \subsetneqq \alpha{+}\Delta,\\ H(\theta_p,\theta_\alpha), & \alpha = \text{ALL}. \end{cases} \tag{1}$$

We assume p denotes a relevant patent from the container. α (eg: $\{\text{A}\}$) denotes the index of current container or perspective which corresponds to one or more IPC sections while $\alpha + \Delta$ (eg: $\{\text{A, B}\}$) denotes the augmented IPC sections. α and $\alpha + \Delta$ all come from the section domain set of the query patent. $H(\theta_p, \theta_{\alpha+\Delta})$ denotes relative entropy between patent document p and relevant patents stored in container of $\alpha + \Delta$, while $H(\theta_p, \theta_\alpha)$ denotes relative entropy between patent document p and relevant patents stored in container of α. We evaluate relevance between a relevant patent and the technical characteristic of a patent perspective by the difference of relative entropy. The difference assigns higher scores to patents which contain specific terms and are more similar to α and less similar to $\alpha + \Delta$. When α and the query patent have the same IPC sections, relevance is evaluated by relative entropy itself.

We construct perspective convertors to present domain-relevant technical characteristics as follows.

$$P(w|\theta_\alpha) = Z_w \sum\nolimits_{w\in p, p\in k_\alpha} P(w|p)P(p|\theta_\alpha). \tag{2}$$

$P\,(w|\theta_\alpha)$ denotes the weight of technical terms in the language model θ_α while $P\,(w|p)$ denotes the weight of technical terms in patent document p. And $Z_w = \sum_{w\in\theta_\alpha} \frac{1}{P(w|\theta_\alpha)}$ is defined as term-specific normalization factor.

Perspective conversion adjusts the technical characteristic of query patents based on the technical characteristic of composite-domain perspectives, which strengthens the technical characteristics of query patents that are similar to the composite-domain perspectives and inhibit the difference. We present three approaches of perspective conversions to generate aspect queries.

- Linear conversion (LINE)

$$P_\alpha(w|\chi) = \lambda P(w|q) + (1-\lambda)P(w|\phi_\alpha). \tag{3}$$

- Multiplicative conversion (MULTIPLY)

$$P_\alpha(w|\chi) = P(w|q) * P(w|\phi_\alpha). \tag{4}$$

- Relative-entropy conversion (RE)

$$P_\alpha(w|\chi) = \gamma P(w|q) + (1-\gamma)RE(P(w|q), P(w|\phi_\alpha)). \tag{5}$$

$$RE(P(w|q), P(w|\phi_\alpha)) = \\ Z_{ww}\begin{cases} P(w|q)\log\frac{P(w|q)}{P(w|\phi_\alpha)}, P(w|q) \geqslant P(w|\phi_\alpha),\\ P(w|\phi_\alpha)\log\frac{P(w|\phi_\alpha)}{P(w|q)}, P(w|q) < P(w|\phi_\alpha). \end{cases}$$

Linear conversion is a linear integration of the query language model and the perspective language models in Eq. (3) while multiplicative conversion takes the products of the term frequency in different language models as aspect queries in Eq. (4). Relative-entropy conversion is a linear combination of the query language model and the relative entropy of the two language models in Eq. (5). In order to ensure the non-negative property, relative-entropy conversion is divided into two cases with differential processings. Z_{ww} is the normalization factor. The effectiveness of conversions will be discussed in the experiment section.

3.3 Expansion model with double consistency

In Section 3.2, we realize aspect retrieval under composite-domain perspectives through perspective conversions and retrieval information all comes from query patents. However, patent documents often have obscure style of writing and semantic ambiguity, which affects the retrieval performance.

In this section, we introduce a novel patent expansion method to reduce semantic ambiguity. Expanded retrieval with double consistency is implemented based on aspect queries and semantic vector spaces (introduced in Section 3.3.1). In order to meet the domain-aware expanded requirement, we use word embedding to train the semantic vector spaces as expanded resources (introduced in Section 3.3.2).

3.3.1 Construction of double-consistency expanded queries

Patent expansion has two opposite effects. It enriches the retrieval semantics by adding relevant terms. However, not all contribute to the improvement of performance. Patent expansion has two kinds of strategies, i.e., semantic expansion and distributed expansion. The former realizes semantic consistency and ignores distributed consistency while the latter pursues distributed consistency and ignores semantic consistency. Those expanded approaches with only one consistency will hurt the performance.

We introduce a novel patent expanded model with double consistency which combines distributed consistency with semantic consistency. In order to implement patent expansion with double consistency, we determine the information resources for distributed consistency and semantic consistency. We take mutual information to guarantee distributed consistency which uses term distribution to explain technical characteristics. We take semantic vector spaces to guarantee semantic consistency which use euclidean distance of term vectors to quantify semantic relevance.

Our proposed expansion model consists of three steps. In the first step, we use mutual information to quantify the distributed consistency between candidate terms and aspect queries under the single-domain perspective in Eq. (6). In the second step, we calculate the double consistency of candidate terms by the combination of the distributed relevance and the semantic similarity under the single-domain perspective in Eq. (10). In the third step, we fuse the aspect language model (Eqs. 3–5) and the accumulation of double consistency under the single-domain perspectives to generate the double consistency under the composite-domain perspectives in Eq. (11).

Semantic vector spaces provide domain-aware terms with similar semantics as candidate terms. We also constrain that expanded terms have higher distributed relevance with terms of aspect queries. We use mutual information to quantify the distributed relevance between semantic similar terms and aspect queries under the single-domain perspective.

$$R_\beta(c;\chi) = Z_{c\in\kappa(w)} \sum\nolimits_{w'\in\chi} P_\beta(c,w') \log \frac{P_\beta(c,w')}{P_\beta(c)*P_\beta(w')}, \quad (6)$$

where c denotes a candidate term with similar semantics to query term w in semantic vector space ψ_κ. single-domain section κ and single-domain perspective β correspond to the same IPC section. $R_\beta(c;\chi)$ denotes the relevance between candidate term c and aspect query χ under single-domain perspective β. And $Z_c = \frac{1}{\sum_{c\in\kappa(w)} R_\beta(c;\chi)}$ denotes the candidate-specific normalization factor.

$$P_\beta(c,w') = \frac{n_\beta(c,w')}{N_\beta}. \quad (7)$$

$$P_\beta(c) = \frac{n_\beta(c)}{N_\beta}, \quad (8)$$

$$P_\beta(w') = \frac{n_\beta(w')}{N_\beta}. \quad (9)$$

We calculate parameters based on the relevant patent set of single-domain perspective β, namely $\eta_{\alpha=\beta}$. $n_\beta(c,w')$ denotes the number of patent documents that term c and w' appear in. $n_\beta(c)$ denotes the number of patent documents that term c appears in while $n_\beta(w')$ denotes the number of patent documents containing the term w'. And N_β denotes the number of relevant patent documents.

Semantic vector spaces provide domain-aware terms and semantic similarity for query terms. We calculate double consistency of candidate terms through mutual information and semantic similarity under the single-domain perspective.

$$RS_\beta(c|\chi,w) = R_\beta(c;\chi) * SIM_\kappa(c|w), \quad (10)$$

where $RS_\beta(c|\chi,w)$ denotes the double consistency between candidate term c and aspect query χ under single-domain perspective β. $R_\beta(c;\chi)$ is the distributed relevance in Eq. (6) and $SIM_\kappa(c|w)$ denotes the semantic similarity between query term w and candidate term c under single-domain section κ, which is the euclidean distance evaluated in semantic vector space ψ_κ.

$$RS_\alpha(c|\chi) = P_\alpha(w|\chi) * \sum\nolimits_{\beta\in\alpha} RS_\beta(c|\chi,w). \quad (11)$$

$RS_\alpha(c|\chi)$ denotes the double consistency between candidate term c and aspect query χ under composite-domain perspective α, which is evaluated by the technical characteristic of query term w in aspect query χ and the accumulation.

3.3.2 Expanding resource based on word embedding

We implement patent expansion with double consistency to improve the retrieval performance. Current works take repositories to guarantee semantic consistency. However, common repositories collect usual words and common semantics, which hardly meets the requirement of patent expansion. Hence, we propose a method that uses patent documents to train semantic vector spaces as domain-aware expanded resources.

Word embedding trains the semantic vector spaces by the statistics of neighbor word distribution. The model maps words to vectors and the euclidean distance of vectors could be used to represent the semantic distance of words [19].

This work assumes that patents with similar classification tend to use the same terms. We create eight containers corresponding to eight IPC sections. When all the IPC codes of a patent share one section, we put the patent into the corresponding container. We use patent documents in each container to train semantic vector spaces and obtain eight semantic expanded resources.

Table 2 shows the semantic similar terms and the similarity of *mouse* in the semantic vector space of IPC section A and H. Section A represents $HUMANNECESSITIES$ and term *mouse* has some semantic similar terms, such as *rat*, *rabbit*, *mice*, *murine*. Section H represents $ELECTRICITY$ and term *mouse* also has some semantic similar terms, such as *keyboard*, *trackball*, *click*, *touchscreen*.

Table 2 Similar terms and similarity of term *mouse* in IPC section A and H

Mouse			
Section A		Section H	
Term	Similarity	Term	Similarity
Rat	0.7936	Amouse	0.8162
Rabbit	0.6809	Keyboard	0.7304
Mice	0.6535	Trackball	0.6781
Murine	0.6337	Click	0.5976
Micewhich	0.5695	Touchscreen	0.5528
Pig	0.5426	Touchpanel	0.5410

In order to underline the originality of one innovation, applicants tend to create new vocabulary that may not fully comply with the grammatical or usage habit. The mismatched words of *touchscreen* and *touchpanel* may frequently appear in patent documents while words of *touch screen* and *touch panel* may be popular among ordinary people in the early stage of one invention. In preprocessing stage, patent documents with some special characters and formulas will be converted into general texts for index. Such pretreatment usually generates some misspelled words, such as *amouse*(namely *a mouse*), micewhich(namely *mice which*). There are a large number of mismatched and misspelled words in full text indexing of patent documents. Common repositories would not contain these words but we can tackle these problems by selecting expansion terms from semantic vector space, which will obviously contribute to the performance of patent retrieval.

3.4 Fusion ranking model

We achieve more than one ranking lists of relevant patents through the retrieval of multiple queries generated by a query patent under composite-domain perspectives. We combine them and obtain a single ranking list based on the technical relevance. We fuse the relevant patents from different ranking lists based on the weight of composite-domain perspectives and the rank of relevant patents in the list as expressed below:

$$P(p,\chi) = \sum\nolimits_{\alpha} \frac{N_\alpha(\chi) - o_\alpha(p,\chi)}{N_\alpha(\chi)} * W(\alpha), \tag{12}$$

where $N_\alpha(\chi)$ denotes the number of relevant patents in a ranking list obtained by aspect query χ. $o_\alpha(p,\chi)$ denotes the rank of relevant patent p in the ranking list. $P(p,\chi)$ is the relevance accumulation between aspect queries and relevant patent p under different composite-domain perspectives. $W(\alpha)$ denotes the weight of composite-domain perspective α.

$$W(\alpha) = \frac{n_\alpha}{N}, \tag{13}$$

where N denotes the number of relevant patents obtained from baseline retrieval. n_α denotes the number of relevant patents stored in composite-domain container η_α.

We have introduced our approach of three models, namely composite-domain perspective model, expansion model with double consistency and fusion ranking model. In next section, we will evaluate the effectiveness of our three models.

4 Experiments

In this section, we describe our experimental setup and four comparsion experiments under various settings. In each comparsion, we report the baselines and analyse experimental results.

4.1 Experimental setup

4.1.1 Dataset and evaluation setup

Patent dataset In this study, we used a large patent dataset of CLEP-IP 2010 released by CLEP Intellectual Property track. CLEP-IP 2010 provides a topic set which are patent applications and have *title*, *abstract*, *description* and *claims*. Patent applications are annotated with the metadata tags, such as IPC classes. CLEP-IP 2010 consists of 2.6 million distinct patent documents published between 1985 and 2001 and almost covers all the patent technical classified domains. CLEP-IP 2010 has been taken as evaluation set in most recent works [6,8,10–13,16,18] and some other tasks. In our experiments, we used the English subsection of CLEP-IP 2010. The English test set of CLEP-IP 2010 corresponds to 1348 topics (patent applications).

Preprocessing During indexing and retrieval, both documents and queries are stemmed by Stanford CoreNLP. Stopword removal is performed by using the stop-word list [4]. We also remove all the formulas and numeric references. The retrieval experiments described in this paper are implemented by using Lucene.

Evaluation measures The performance of this work is quantified by recall and mean average precision (MAP). We also report the evaluation results of patent retrieval evaluation score (PRES [20]) which combines MAP and recall in one single score and is designed for recall-oriented applications.

Recall.

$$Recall\,(q,\ Z) = \frac{T\,(q,\ Z)}{z}, \tag{14}$$

where Z is the number of patents to be checked by the user (cut-off value) and z is the number of relevant patents.

MAP.

$$MAP\,(q,\ \mathrm{Z}) = \frac{\displaystyle\sum_{i=1}^{t} \frac{i}{r_i\,(q, Z)}}{t}, \tag{15}$$

where r_i is the rank of relevant patent and t denotes the number of relevant patents obtained by the retrieval.

PRES.

$$PRES\,(q, Z) = 1 - \frac{\dfrac{\sum r_i}{z} - \dfrac{z+1}{2}}{Z}, \tag{16}$$

$$\sum r_i = \sum_{i=1}^{zR} r_i + zR\,(Z + z) - \frac{zR\,(zR - 1)}{2}, \tag{17}$$

where R is the number of relevant patents in the checking window and $\sum r$ is the summation of ranks of relevant patents.

4.1.2 Parameter setup

In this section, we describe the details of parameter setup in our experiment. We use the training topics of CLEP-IP 2010 for tuning the parameters of our model. This training set consists of 196 English topics with multiple sections.

Scoring function Scoring function is used to score the relevance between a topic and patent documents. Current works used LMDS (Dirichlet Smoothing), LMJMS (Jelinek-Mercer Smoothing), BM25 and VSM to score the relevance. We are interested in finding the suitable scoring function for our model. Figure 4 presents the retrieval results of four scoring functions on our model. VSM significantly outperforms the other three scoring functions on recall and MAP. VSM also obtains the best performance balance in terms of three evaluation measures. Hence, we take VSM as the scoring function for our model.

Number of retrieval terms Figure 4 presents the results of different numbers of retrieval terms in patent retrievals. The results suggest that the number of retrieval terms have an immense influence on the results and a specific number of retrieval terms hardly presents the performance comprehensively. So we evaluate the performance of our work in a number range of retrieval terms, such as [10-100, step 10].

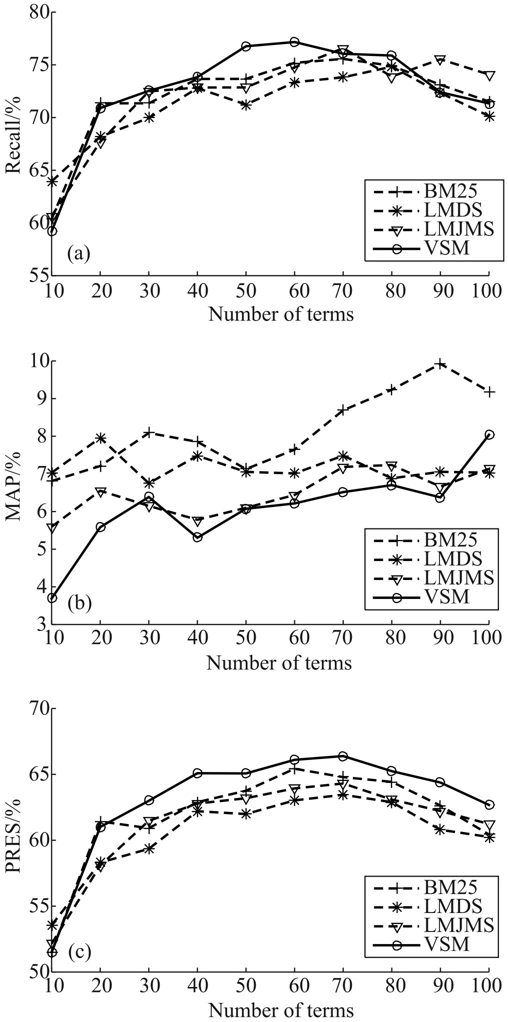

Fig. 4 Retrieval performance of four sorting methods. (a) Recall; (b) MAP; (c) PRES

4.2 Experimental evaluation

4.2.1 Effectiveness of composite-domain perspective model

In this section, our goal is to predict whether aspect queries under composite-domain perspectives are effective. *Oracular* is an excellent method to generate multiple queries based on the iterative retrieval [6]. In each iteration, it selects retrieval terms based on the following constraints: 1) relevant patents have common IPC codes with the query patent. 2) a word is selected as a term for next iterative query only when word frequency of pseudo-relevant patents is higher than non-relevance patents. And *Oracular* achieved the state-of-the-art performance. Therefore, we use *Oracular* as a baseline and 484 English topics with multiple IPC sections of CLEP-IP 2010 for evaluation.

We use the training set of CLEP-IP 2010 for tuning the parameters and the optional values of λ and γ are set to 0.5 and 0.6 for all the experiments. Due to the immense performance difference with different numbers of terms, we select the best performance of *Oracular* as a baseline. The results are shown in Fig. 5. Compared with *Oracular*, composite-domain perspective model constantly achieves better performance and keeps a relatively large advantage in terms of recall, MAP and PRES. The average recall of *Oracular* is 50.71% while linear conversion of our model obtains 71.89% when we use 40 terms for the retrieval presented in Fig. 5(a). This demonstrates that section is the suitable IPC level as the basic unit of classification level for composite-domain perspectives. *Oracular* is an iterative retrieval and generates multiple queries without explicit interpretations, which results in semantic drift on topics with multiple IPC sections. Our work generates multiple queries with a rational interpretation that technical similarity has relevance to technical classified domains.

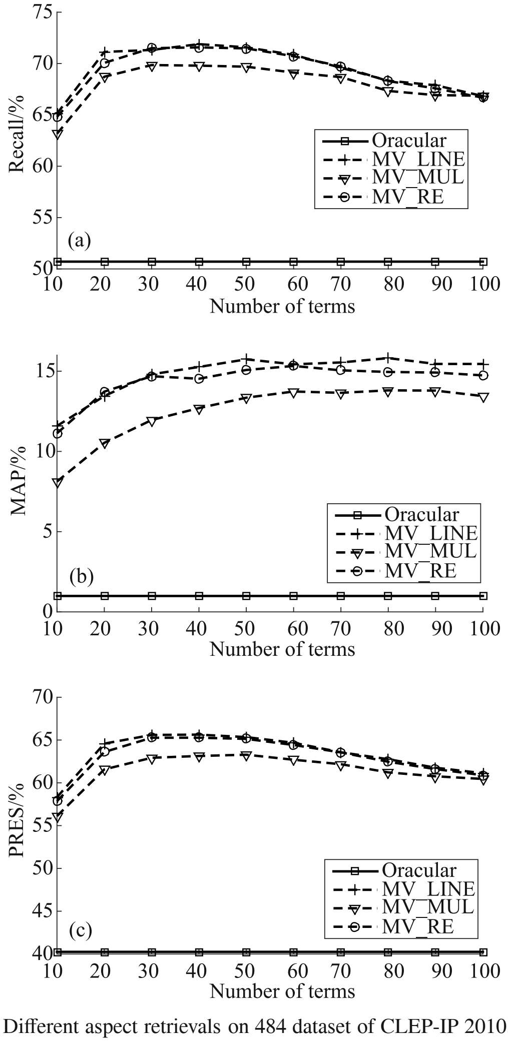

Fig. 5 Different aspect retrievals on 484 dataset of CLEP-IP 2010. (a) Recall; (b) MAP; (c) PRES

Figure 5 shows that linear conversion and relative-entropy conversion obtain the similar performance and maintain the slightly better than multiplicative conversion. The three Subfigures in Fig. 5 look like downward parabolas, which suggests that our work is very effective on identifying key terms of query patents for aspect queries according to different composite-domain perspectives.

Our model is also helpful to determine the best number of key terms for a query in practical retrieval application. In Fig. 5, all the performance goes up with the increase of terms at first and then remains stable. So we could find the inflection number of key terms to achieve the best balance in terms of recall, MAP and PRES simultaneously.

4.2.2 Effectiveness of expanded model with double consistency

In this section, we wish to examine the performance of our expanded model which combines distributed consistency and semantic consistency.

In order to verify the effectiveness of double consistency expansion, we present the results in the form of *Less Than Set* (LTS). A LTS represents a topic set where the performance (such as recall, MAP and PRES) of each topic in the aspect retrieval without patent expansion is less than a specific threshold. For example, a LTS of 10% denotes a set of query patents, where the recall of each topic without patent expansion is less than 10%.

Our experiments indicate that expanded retrieval outperforms unexpanded retrieval in terms of recall, MAP and PRES. In Table 3, the average recall of a LTS of 10% without expansion is 1.72% while the average recall of the LTS reaches 4.29% when adding 8 expanded terms, achieving 149% relative improvement. The higher performance a LTS without expansion reaches, the less improvement the LTS after expansion achieves. These results are in line with our expectation. They suggest that double consistency is beneficial to expand terms which share technical characteristic with terms of query patents.

With more expanded terms, we find that improvement becomes less and finally performance declines. This could help us to find a explicit boundary point of expanded term number to obtain the best performance.

From Table 3, we also find that patent expansion with double consistency just has a slight absolute performance improvement for two reasons: a) patent expansion is a doubleedged sword, which enriches relevant semantics and non-relevant semantics simultaneously. b) the result of unexpanded retrieval also has great influence on the performance of expanded retrieval.

$$\zeta = |P_E(q) - P_{UE}(q)|. \tag{18}$$

Table 3 Performance comparison of unexpanded and expanded retrieval

Type[1]	W40E*[2]	10/%[3]	30/%	50/%	60/%	80/%	100/%
Recall	W*E00	1.72	9.40	27.20	31.22	47.00	71.32
	W*E04	3.33	10.79	27.96	31.79	47.32	71.36
	W*E08	4.29	11.30	28.13	32.05	47.53	71.16
	W*E17	4.29	11.44	28.15	32.06	47.51	71.00
	W*E30	4.29	11.44	28.06	32.00	47.47	70.98
MAP	W*E00	3.02	7.43	10.57	11.71	12.74	13.65
	W*E04	3.30	7.62	10.66	11.78	12.81	13.68
	W*E08	3.38	7.68	10.70	11.78	12.83	13.69
	W*E17	3.48	7.75	10.75	11.83	12.87	13.78
	W*E30	3.50	7.77	10.76	11.84	12.88	13.79
PRES	W*E00	1.87	10.19	24.01	30.18	45.41	64.58
	W*E04	2.43	10.64	24.20	30.35	45.46	64.46
	W*E08	3.38	11.30	24.57	30.73	45.45	64.32
	W*E17	3.11	11.16	24.40	30.59	45.30	64.22
	W*E30	3.11	11.16	24.32	30.54	45.24	64.20

[1] All the results are evaluated in linear conversion on CLEP-IP 2010
[2] $W40E*$ represents that each aspect query contains 40 terms from the query patent and a specific number expanded terms
[3] 10% is the average performance of a LTS which represents a topic set where the performance of each topic in the aspect retrieval is less than 10%

We find that the retrieval results of some topics just have slight improvements after expansion but notably pull down the average performance. We define ζ to quantify the performance change between expansion and unexpansion as shown in Eq. (14). We remove a query patent from a LTS when ζ of the query patent is less than a specific threshold σ.

We set $\sigma = 10^{-4}$ and Fig. 6 presents the result in terms of recall. W40E0 represents that every query contains 40 terms of a query patent while W40E17 denotes that every query is a combination of 40 terms of a query patent and 17 expanded terms. The average recall of a LTS of 10% without expansion is 0.99% while the average recall of the LTS reaches 45.09%. Compared Table 3 and Fig. 6, we find that patent expansion of double consistency fails to greatly improve the absolute average performance but has a significant improvement on a part of query patents, especially for a topic with an originally poor result.

4.2.3 Effectiveness of fusion ranking model

In Section 4.1.2, we introduce four scoring functions and select VSM for our work. VSM has been used to score the relevance of a query and relevant patents in the baseline retrieval, the aspect retrieval and the expanded retrieval. In this section, we wish to examine the effectiveness of fusion ranking model. Hence, we take all the four functions as our baselines to merge the aspect retrieval results based on the relevant score. We use 484 English topics with multiple IPC sections of CLEP-IP 2010 as our experimental evaluation set.

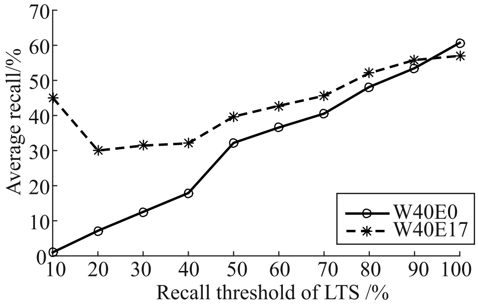

Fig. 6 Recall comparison of unexpanded and expanded retrieval under composite-domain perspectives with $\sigma = 10^{-4}$

Figure 7 shows the results of our fusion ranking method for merging multiple retrieval ranking lists. Our proposed method achieves the best results in terms of recall and PRES. This indicates that the weight of composite-domain perspectives and rank of aspect ranking lists are the effective features to quantify the relevance between a query and relevant patents. As for MAP, our work is slightly weaker than VSM but superior to other three fusion methods. MAP is usually used as the central evaluation measure treating recall and precision equally while recall and PRES are designed for recall-oriented applications [20]. Patent retrieval is a recall-oriented task and our method obtains the highest recall and PRES. Therefore, we think that the fusion ranking model is a successful design for multiple-result fusion in our work.

4.2.4 Comparisons with baselines

In this section, we compare our method with the following three methods representing the latest state-of-the-art works on all the test set of CLEP-IP 2010.

ARCIP ARCIP [10] used a language modeling framework to score single terms from query patents and train a regression model to predict the effectiveness of the phrases for patent expansion.

QDPI QDPI [11] proposes a proximity-based query expanded method to address the term mismatch problem in patent retrieval by calculating the proximity information between the lexicons of IPC definition pages and query terms of a query patent in the pseudo patent documents.

QDMCN QDMCN [13] performs query refinement by making use of the time-evolving patent citation network and term distribution of the relevant patents.

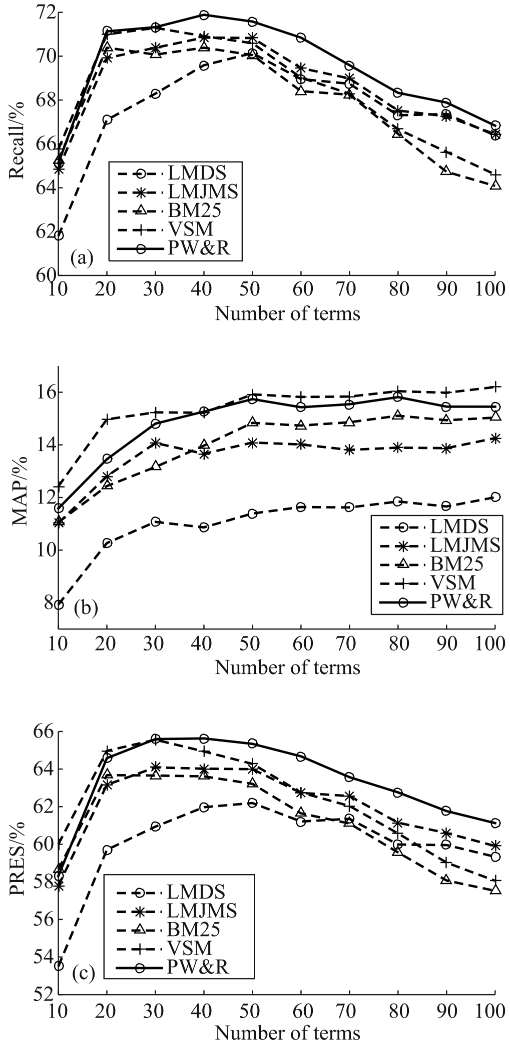

Fig. 7 Fusion ranking comparison on 484 dataset of CLEP-IP 2010. (a) Recall; (b) MAP; (c) PRES

Table 4 presents the comparison results. Our proposed approaches precede existing works in terms of recall and PRES. LINE achieves the average 5.43% improvement over QDMCN in terms of recall while LINE+ obtains the average 12.38% improvement over the state-of-the-art in terms of PRES. And expanded retrieval with double consistency obtains the higher performance than composite-domain perspective retrieval.

In terms of MAP, our work significantly exceeds QDPI and QDMCN but is slightly weaker than ARCIP for two reasons: a) a query is generated by combining the query patent with perspective conversion, which results in a slight semantic divergence. b) patent expansion improves the recall, but expands some relevant terms that are different from the terms of query patent and most relevant patents, which lowers the ranks of most relevant patents.

Table 4 Retrieval performance comparison on CLEP-IP 2010

Method	Recall/%	MAP/%	PRES/%
ARCIP	65.00	15.60	56.70
QDPI	65.95	10.50	55.40
QDMCN	67.68	7.80	57.84
LINE	71.32	13.65	64.58
MULTIPLY	69.07	12.29	62.63
RE	70.97	12.92	64.51
LINE+[1]	71.36	14.33	65.00
MULTIPLY+	69.96	12.53	63.05
RE+	71.05	13.31	64.53

[1] + represents the aspect retrieval with double consistency expansion

Patent retrieval is a recall-oriented task. In Table 4, we find that existing works tend to sacrifice MAP to improve recall and PRES which are designed for recall-oriented applications. Compared with existing works, our work obtains the best balance in terms of all the three evaluation measures.

5 Conclusion and future work

In this paper, we perform the patent prior art search including both patent reduction and patent expansion. We first measure the technical similarity of patents under composite-domain perspectives and implement patent reduction to generate aspect queries. We then realize the patent expansion in the semantic vector space. Our expansion distinguishes itself by double consistencies which overcome the shortage of single distributed or semantic consistency. Finally, a novel fusion method, taking the perspective weight and the rank of relevant patents into consideration, is proposed to merge multiple retrieval results. Our experiments verify the effectiveness of the three models and our work achieves the best performance balance in terms of all three evaluation measures.

Our work can be expanded in several ways. First, we can further study the performance prediction of original retrieval, which could be used to determine whether to perform patent expansion with double consistency. Second, we can define a better merging method for multiple retrieval results to overcome the semantic divergence of composite-domain perspectives.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61232002, 61572376), the Science and Technology Support Program of Hubei Province (2015BAA127) and the Wuhan Innovation Team Project (2014070504020237).

References

1. Zhang L, Li L, Li T. Patent mining: a survey. ACM SIGKDD Explo-

rations Newsletter, 2015, 16(2): 1–19

2. Xue X, Croft W B. Automatic query generation for patent search. In: Proceedings of the 18th ACM International Conference on Information and Knowledge Management. 2009, 2037–2040
3. Xue X, Croft W B. Transforming patents into prior-art queries. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009, 808–809
4. Kim Y, Seo J, Croft W B. Automatic boolean query suggestion for professional search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011, 825–834
5. Kim Y, Croft W B. Diversifying query suggestions based on query documents. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2014, 891–894
6. Far M G, Sanner S, Bouadjenek M R, Ferraro G, Hawking D. On term selection techniques for patent prior art search. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015, 803–806
7. Al-Shboul B, Myaeng S H. Query phrase expansion using wikipedia in patent class search. In: Proceedings of the 7th Asia Information Retrieval Symposium. 2011, 115–126
8. Magdy W, Jones G J F. A study on query expansion methods for patent retrieval. In: Proceedings of the 4th Workshop on Patent Information Retrieval. 2011, 19–24
9. Kishida K. Pseudo relevance feedback method based on taylor expansion of retrieval function in NTCIR-3 patent retrieval task. In: Proceedings of the ACL-2003 Workshop on Patent Corpus Processing. 2003, 33–40
10. Mahdabi P, Andersson L, Keikha M, Crestani F. 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. 2012, 505–514
11. Mahdabi P, Gerani S, Huang J X, Crestani F. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2013, 113–122
12. Wang F, Lin L. Domain lexicon-based query expansion for patent retrieval. In: Proceedings of the 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. 2016, 1543–1547
13. Mahdabi P, Crestani F. Query-driven mining of citation networks for patent citation retrieval and recommendation. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. 2014, 1659–1668
14. Judea A, Schütze H, Brügmann S. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In: Proceedings of the 15th International Conference on Computational Linguistics. 2014, 290–300
15. Magdy W, Leveling J, Jones G J F. Exploring structured documents and query formulation techniques for patent retrieval. In: Proceedings of the Workshop on Cross-Language Evaluation Forum for European Languages. 2009, 410–417
16. Mahdabi P, Crestani F. Patent query formulation by synthesizing multiple sources of relevance evidence. ACM Transactions on Information Systems, 2014, 32(4): 1–30
17. Cetintas S, Si L. Effective query generation and postprocessing strategies for prior art patent search. Journal of the Association for Information Science and Technology, 2012, 63(3): 512–527
18. Ganguly D, Leveling J, Magdy W, Jones G J F. Patent query reduction using pseudo relevance feedback. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011, 1953–1956
19. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. 2013, arXiv preprint arXiv:1301.3781
20. Magdy W, Jones G J F. PRES: a score metric for evaluating recall-oriented information retrieval applications. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2010, 611–618

Fei Wang is a PhD candidate at the School of Computer Science, Wuhan University, China. His current research interests are in the database, complex data management, patent mining, information retrieval and natural language processing. He received the ME degree in Computer Science from Chengdu University of Information Technology, China in 2014.

Tieyun Qian is a professor at the State Key Laboratory of Software Engineering at Wuhan University, China. She received her BS degree in Computer Science from Wuhan University of Technology, China in 1991, and her PhD degree in Computer Science from Huazhong University of Science and Technology, China in 2006. Her current research interests include text mining, Web mining, and natural language processing. She has published over 30 papers in leading conferences including ACL, EMNLP, SIGIR, etc. She is a member of CCF and ACM. She has served as program committee member of many premium conferences: WWW, COLING, DASFAA, WAIM, and APWeb.

Bin Liu is a lecture at the School of Computer Science, Wuhan University, China. Bin Liu received the PhD, BS, and ME degree in Computer Science from Wuhan University, China. His current research interests are in the database, data mining, complex data management and natural language processing.

Zhiyong Peng received the BS and ME degree in Computer Science from Wuhan University and Changsha Institute of Technology of China, respectively. He received PhD degree from Kyoto University of Japan in 1995. He is a professor at Wuhan University. Prior to join Wuhan University in 2000, he worked as a researcher at the Advanced Software Technology and Mechatronics Research Institute of Kyoto from 1995 to 1997 and was a member of the technical staff at Hewlett-Packard Laboratories, Japan from 1997 to 2000. His current research interests are in the database, trusted data management, and complex data management.