Exploiting word embedding for heterogeneous topic model towards patent recommendation

Jie Chen, Jialin Chen, Shu Zhao, Yanping Zhang & Jie Tang

Scientometrics
An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy

ISSN 0138-9130
Volume 125
Number 3

Scientometrics (2020) 125:2091-2108
DOI 10.1007/s11192-020-03666-4

Springer

Scientometrics (2020) 125:2091–2108
https://doi.org/10.1007/s11192-020-03666-4

Exploiting word embedding for heterogeneous topic model towards patent recommendation

Jie Chen[1,2] · Jialin Chen[1,2] · Shu Zhao[1,2] · Yanping Zhang[1,2] · Jie Tang[3]

Received: 9 December 2019 / Published online: 17 August 2020

Abstract
Patent recommendation aims to recommend patent documents that have similar content to a given target patent. With the explosive growth in patent applications, how to recommend relevant patents from the massive number of patents has become an extremely challenging problem. The main obstacle in patent recommendation is how to distinguish the meanings of the same word in different contexts or associate multiple words that express the same meaning. In this paper, we propose a Heterogeneous Topic model exploiting Word embedding to enhance word semantics (HTW). First, we model the relationship among text, inventors, and applicants around the topic to build a heterogeneous topic model and learn the patent feature representation to capture contextual word semantics. Second, a word embedding is constructed to extract the deep semantics for associating multiple words that express the same meaning. Finally, with words as connections, the mapping from patent feature representations to patent embedding is established through a matrix operation, which integrates the information between the word embedding and patent feature representation. HTW considers the heterogeneity of patents and enhances the distinction or association among words simultaneously. The experimental results on real-world datasets show that HTW exceeds typical keyword-based methods, topic models, and embedding models on patent recommendations.

Keywords Patent recommendation · Heterogeneous topic model · Word embedding

✉ Shu Zhao
zhaoshuzs2002@hotmail.com

Jie Chen
chenjie200398@163.com

Jialin Chen
cjialin2019@outlook.com

Yanping Zhang
zhangyp2@gmail.com

Jie Tang
jietang@tsinghua.edu.cn

[1] Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Hefei 230601, China

[2] School of Computer Science and Technology, Anhui University, Hefei 230601, China

[3] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Introduction

Patent recommendation is a technique for recommending patent documents that contain similar content to a given target patent. Automatic patent recommendation, aiming to assist in retrieving patents, analyzing patent documents, applying for patents for new inventions, and so on, has emerged as a new popular research area in recent years. Although patent recommendation is important and widely used, with the explosive growth in patents, how to recommend relevant patents from the massive number of patents remains a difficult task.

There is a growing body of research that attempts to achieve automatic patent recommendation. Keyword-based methods, such as query keyword extraction or query expansion, have been introduced for matching keywords or phrases to find relevant patents through automatic extension of the keyword list (Lupu et al. 2017; Tannebaum and Rauber 2015; Wang et al. 2013). A topic model in which words have different meanings in each topic is used in patent recommendation (Krestel and Smyth 2013). The topic model avoids manually adjusting keywords by learning the word distribution containing specific semantic meanings of the same word in different contexts (Liu et al. 2015). Embedding methods represent each word or document as just one embedding, which extracts the implicit semantics in words rather than just considering the appearance of words (Helmers et al. 2019; Shalaby and Zadrozny 2018). However, inventors tend to intentionally use complex words and jargon to refer to the same concepts for establishing patent novelty in the patent application. This can be briefly illustrated with the words "pinna" and "auricle". These two words express the same meaning in some cases. "Auricle" is more commonly used, but "pinna" may be chosen by applicants. Meanwhile, less consistency in the meanings of words across patents in different technology domains leads to difficulty in using content-based retrieval (Chen et al. 2018). As an example, the word "property" indicates a thing or things that are owned by somebody in the legislation. However, this word also indicates an attribute or characteristic of something in other fields. Moreover, patent data have substantial bibliographic information that contains rich and useful insights of technological and economic value, but few studies have focused on these bibliographic information.

To utilize the heterogeneity of patents and simultaneously enhance the distinction or association among words, we propose a Heterogeneous Topic model exploiting Word embedding to enhance word semantics (HTW) towards patent recommendation, as shown in Fig. 1. First, we model the relationship among text, inventors and applicants around the topic for building a heterogeneous topic model to capture the patent feature representation. In the heterogeneous topic model, each word in each topic has different semantics, and the patent feature representation learned by the patent-topic-word structure captures the contextual word semantics. Second, a word embedding is constructed to extract the deep semantics for associating multiple words that express the same meaning. Finally, with words as connections, the word embedding matrix is utilized to map the patent feature representation to the vector space for learning the patent embedding. The patent embedding integrates the information between the word embedding and patent feature representation, which simultaneously enhances the association of similar words and the distinction of multiple meanings in a word.

The main contributions of our work are summarized as follows.

(1) We propose a novel heterogeneous topic model that integrates the relationship among text, inventors and applicants around the topic and introduces the information of inven-

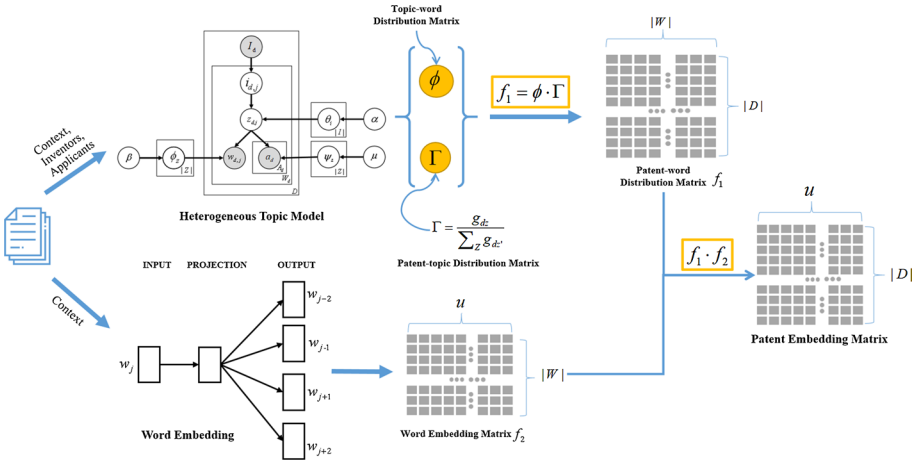

Fig. 1 The main idea of the proposed method HTW

tors and applicants into a patent feature representation to enhance the semantics of words in the context.

(2) A novel patent embedding model that utilizes a word embedding as the mapping matrix is proposed to simultaneously distinguish the meanings of the same word in different contexts and associate multiple words that express the same meaning.

(3) Experiments on three real patent datasets demonstrate that HTW outperforms word embedding models, also exceeding topic models and other representative document models on patent recommendation.

The remainder of this paper is organized as follows. "Related work" section reviews the related work on patent recommendation. "Preliminaries" section describes the notations used in this paper, presents the problem definition of patent recommendation and introduces the preprocessing of patent data. "Exploiting word embedding for heterogeneous topic model" section introduces the specific details of the proposed model. We present and analyze the experimental results in "Experiments" section. Finally, we conclude this paper and outline directions for future work in "Conclusion" section.

Related work

Patent recommendation techniques have been investigated to facilitate the search for similar or related patents, which can assist in retrieving patents, analyzing patent documents, applying for patents for new inventions and so on (Oh et al. 2013). A key aspect of patent recommendation is that the patent language is full of jargon and user-defined terminology, which leads to the overwhelming keyword-mismatch problem (Shalaby and Zadrozny 2019). Keyword-based methods, topic models and embedding methods have been widely used for patent recommendation and retrieval to solve the keyword-mismatch problem, but they always ignore the heterogeneity of patent data or fail to simultaneously address the problem of a word having multiple meanings and words that have the same meaning.

Keyword-based methods, such as query keyword extraction or query expansion, attempt to use word or phrase matching to find relevant patents. The existing patent search and analysis systems, such as Google Patent[1], Patentics[2], PriorArtSearch[3] and so on, are based on these methods. In the academic field, query keyword extraction techniques have been introduced for matching words or phrases to find relevant patents (Bashir and Rauber 2010; Ganguly et al. 2011; Xue and Croft 2009). For the problem of words that have the same meaning, a method of using a thesaurus to automatically include similar words for keywords has been proposed, but this method required manual management and expansion of the thesaurus (Lupu et al. 2017; Tannebaum and Rauber 2015; Wang et al. 2013). In response to this, pseudo-relevance feedback, which was given an initial search and used the *top k* search results in the initial search for extending the set of keywords, has been studied (Ganguly et al. 2011; Golestan Far et al. 2015; Mahdabi and Crestani 2012). This method still needed to screen the initial search results to adjust the search results and obtain more accurate words that have the same meaning. For this problem, Wang and Lin used WordNet and Wikipedia as knowledge bases to enrich the initial query with semantic-related concepts (Wang and Lin 2017). Besides, Arts et al. develop and validate a text-based patent similarity measure, which is also a keyword-based patent similarity measure (Arts et al. 2018, 2019). Patent recommendation based on keywords without considering semantics has great limitations. One of the most important points is that there may be few or no identical keywords in queried patents, but the main ideas between two patent documents can be quite similar.

Topic models are able to automatically extract the keywords and main idea of a patent for relevant patent recommendation. Some researchers studied how to convert text into a list of words or a vector list of numbers based on bag-of-words (BOW) and recommend patents by mining the subject matter implied by the full text (Choi et al. 2018; Mahdabi et al. 2011; Verma and Varma 2011). Krestel et al. studied how to use the latent Dirichlet allocation and Dirichlet multinomial regression to profile patent documents and analyze their similarity (Krestel and Smyth 2013). Tang et al. proposed a topic model named the inventor-company-topic (ICT) model (Tang et al. 2012), which simultaneously models the topical aspects of inventors, companies and patent documents. The research field of the inventor contributes to the extraction of the topic of the patent, and the company's field also reflects the topic of the patent to some extent (Li et al. 2014). This model considers the heterogeneity of patent and extracts the semantics of the context, which is conducive to the aggregation of patent topics and improves the differentiation of patents in different fields. But the topic model is greatly influenced by word frequency. The high-frequency words will affect the result of the topic-word distribution, while the semantics of low-frequency words may be ignored. The topic model also ignores word co-occurrence information; thus, the obtained semantic information is not sufficiently accurate.

The embedding methods not only consider the word frequency in the BOW but also consider the word co-occurrence, which can extract the semantic information implied by the text and return the embedding matrix representing the semantic meaning of the patent document (Zhang et al. 2018). Jagendra et al. examined the possibility of using Word2vec (Mikolov et al. 2013) to automatically identify relevant words in the search results that are used to extend the query (Singh and Sharan 2016). Helmers et al. studied

[1] https://www.google.com/?tbm=pts.
[2] https://www.patentics.com/.
[3] https://www.uspto.gov/patent.

Table 1 Notations

Notations	Description	Notations	Description
D	The set of all patents	$w_{d,j}$	The j_{th} word in patent d
CP	The set of candidate patents	$i_{d,j}$	The inventor sampled for the j_{th} word in patent d
TP	The set of target patents	a_d	The applicants of patent d
A	The set of applicants	Θ	The probability distribution of inventor-topic
I	The set of inventors	Φ	The probability distribution of topic-word
W	The set of words in all patents	Ψ	The probability distribution of topic-applicant
Z	The set of topics	Γ	The probability distribution of patent-topic
t	Number of topics	α, β, μ	The dirichlet prior distribution of Θ, Φ, Ψ
v	Dimensions of matrix	f_1	The patent feature representation matrix of size $\|D\| \times \|W\|$
u	Dimensions of embedding	f_2	The word embedding matrix of size $\|W\| \times u$
$z_{d,j}$	The topic assigned to the j_{th} word in patent d	$pvec$	The patent embedding matrix of size $\|D\| \times u$

how to extract feature vectors using Word2vec, Doc2Vec (Le and Mikolov 2014) or other models to obtain the semantic information of texts in patents and finally used feature vectors to search or recommend patents (Helmers et al. 2019). Shalaby et al. embedded words to represent queried patents and then analyzed the semantic similarities between retrieved patents and queried patents based on vector-based representations (Shalaby and Zadrozny 2018). Li et al. attempted to combine a convolutional neural network with the word embedding method to uncover patents (Li et al. 2018). The embedding model learns each word or document with one embedding result, which leads to multiple meanings of the same word that cannot be distinguished in different contexts.

To our knowledge, there are few attempts that simultaneously associate different words with the same meaning and distinguish the meaning of the same word in different contexts. Most studies also ignore the existing information of patented inventors and applicants, which can be used to extract a more accurate topic distribution for the patent and improve the recommendation effect. The proposed approach establishes the association among text, inventors and applicants by taking the topic as the core to build a heterogeneous topic model and returns the patent feature representation, where the meanings of words are extracted based on context. A matrix operation is used to realize the mapping from the patent feature representation to the patent embedding by exploiting the word embedding and taking the word as the bridge.

Preliminaries

For the convenience of description and discussion, we introduce the definition of the patent recommendation problem and the associated notations in this section. In addition, we preprocess the text data before training the HTW model to reduce the influence of high-frequency words in the patent data. The preprocessing methods are also introduced in this section.

Notations

Table 1 describes the meanings of the symbols used in this paper.

Problem definition

Given a target patent set $TP = \{tp_1, tp_2, \ldots, tp_{|TP|}\}$ and a candidate patent set $CP = \{cp_1, cp_2, \ldots, cp_{|CP|}\}$, each cp_j *or* $tp_j = (W_j, I_j, A_j)$ has word list $W_j = [w_1, w_2, \ldots, w_{|W_j|}]$, inventor set $I_j = \{i_1, i_2, \ldots, i_{|I_j|}\}$, and applicant set $A_j = \{a_1, a_2, \ldots, a_{|A_j|}\}$. Our problem is as follows: for a target patent tp_j with $tp_j = (W_j, I_j, A_j)$, calculate the similarity $Pr(cp_{jm}|tp_j)$ between tp_j and each patent cp_{jm} in CP by a patent recommendation algorithm, and finally return $D_{relevant} = [cp_{j1}, cp_{j2}, \ldots, cp_{jk}]$, where the patents are related to tp_j, and if $m < n$, $Pr(cp_{jm}|tp_j) < Pr(cp_{jn}|tp_j)$.

Preprocessing

Prior to training HTW, we divide the text content into words and conduct a subsampling of high-frequency words that were sampled to reduce the impact of frequent words. Indeed, the topic model is a type of BOW model whose result is greatly influenced by the frequent words that provide little useful information. For example, "device" is a frequent word in patent data, and without aligning word frequency, it will appear with high probability in the topic-word probability distribution.

To balance the frequency between rare and frequent words, we refer to the subsampling of frequent words in the skip-gram model (Mikolov et al. 2013), and each word w_j in the training set is discarded with probability computed by Eq. (1):

$$P(w_j) = \left(\sqrt{\frac{f(w_j)}{th} + 1}\right) \times \frac{th}{f(w_j)} \tag{1}$$

where $P(w_j)$ is the probability of retaining word w_j, $f(w_j)$ is the frequency of word w_j, and th is a threshold, whose default value is 10^{-2}. Equation (1) indicates that $P(w_j)$ becomes smaller as $f(w_j)$ increases, and if $f(w_j) > 0.0026$, then word w_j has some probability of being deleted. By this subsampling equation, the word segmentation results will have a more uniform word frequency distribution.

Exploiting word embedding for heterogeneous topic model

In this section, we present a heterogeneous topic model that exploits a word embedding to enhance word semantics towards the patent recommendation, called HTW. This approach is composed of three phases. In the first phase, we propose a novel heterogeneous topic model that integrates the relationship among text, inventors and applicants around the topic to extract the semantics of words based on context. In the second phase, a word embedding is constructed to extract the deep semantics for associating multiple words that express the same meaning. Finally, with words as connections, the mapping from patent feature representation to patent embedding is established by a matrix operation such that the information between the word embedding and patent feature representation is integrated. With the

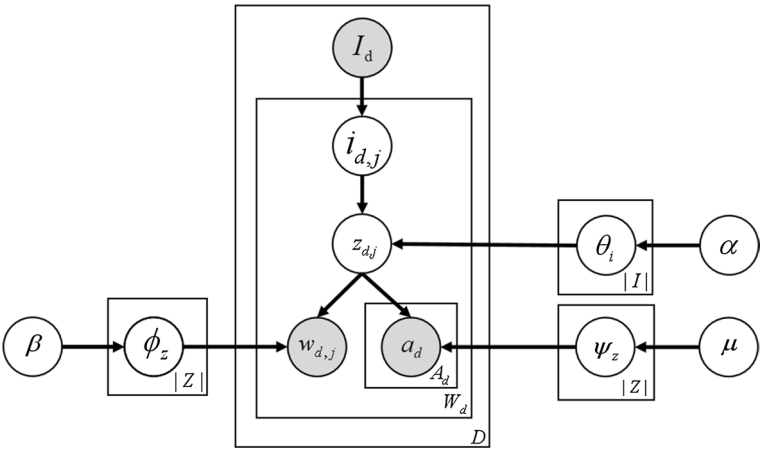

Fig. 2 Heterogeneous topic model containing inventor and applicant information

integration of information, the association of different words with the same meaning and the distinction of the same word with different meanings in different contexts are enhanced simultaneously. Next, we present a detailed illustration of the proposed approach.

Heterogeneous topic model containing inventor and applicant information

Researchers and companies often choose fields they are expert at or familiar with for in-depth research, so research results or developed products can often be used to infer their fields of expertise or interest. Conversely, the fields to which a research result belongs can also be inferred from the research fields of the owners. Based on this, the information of inventors and applicants can be used to improve the accuracy of patent topic distribution.

To establish the association among patent documents, inventors and applicants, we construct three multinomial distributions around the topic to model the heterogeneous topic model. Learning this heterogeneous topic model is to estimate the distribution $\boldsymbol{\Theta}$ of $|A|$ inventor-topics, the distribution $\boldsymbol{\Phi}$ of $|Z|$ topic-words, and the distribution $\boldsymbol{\Psi}$ of $|Z|$ topic-applicants. These three probability distributions are correlated to patent documents, inventors and applicants with the topic, and the learning process is illustrated in Fig. 2.

To more clearly illustrate the process in Fig. 2, it is described as follows: before writing a patent d, each inventor $i \in I_d$ would suggest which topic to incorporate in the patent according to his/her specialty (amount to the topic distribution $P(z|\boldsymbol{\Theta}_i)$); then, word $w_{d,j}$ is sampled from the proposed topic according to $P(w|\boldsymbol{\Phi}_z)$. After this process, each applicant $a \in A_d$ that owns patent d would be relevant to the suggested topics and could constitute a topic distribution of the applicant $P(z|\boldsymbol{\Psi}_a)$. In our heterogeneous topic model, the abstraction of the patent application writing process is represented by Eq. (2). It approximately describes how the text, inventor and applicant are related to each other with the topic as the core.

$$P(I_j = i, Z_j = z | W_j = w, C_j = c, Z_{\neg j}, C_{\neg j}, W_{\neg j}, I_d) \propto \Theta_{iz} \cdot \Phi_{zw} \cdot \Psi_{za} \tag{2}$$

To learn the three multinomial distributions (inventor-topic distribution $\boldsymbol{\Theta}$, topic-word distribution $\boldsymbol{\Phi}$, and topic-applicant distribution $\boldsymbol{\Psi}$) in Eq. (2), Gibbs sampling (Griffiths and Steyvers 2004), which is easy to implement and very effective, is exploited to sample the bibliographic information of patents. Gibbs sampling samples one component of the uniform distribution at one time while leaving the other components unchanged. In Gibbs sampling, the three multinomial distributions are calculated with prior parameters α, β, μ. The sampling process is shown as follows:

$$\boldsymbol{\Theta}_{iz} = \frac{g_{iz} + \alpha}{\sum_Z (g_{iz'} + \alpha)} \tag{3}$$

$$\boldsymbol{\Phi}_{zw} = \frac{g_{zw} + \beta}{\sum_W (g_{zw'} + \beta)} \tag{4}$$

$$\boldsymbol{\Psi}_{za} = \frac{g_{za} + \mu}{\sum_A (g_{za'} + \mu)} \tag{5}$$

where g represents the number of times that the topic, word or applicant is sampled and α, β, μ are the prior parameters. We typically set $\alpha = 50/t, \beta = 0.01, \mu = 0.01$.

Each time after all patents have been sampled, the three learned distributions are updated to integrate the information in each distribution. For patent data, each patent and its inventors or applicants are associated by a uniform distribution. For each sampling process, the distribution of topic applicants is for all applicants of a patent. Therefore, when updating with sample results, it is necessary to take the situation in which one patent corresponds to multiple applicants into account. Considering that there may be multiple applicants for a patent, we average the $\boldsymbol{\Psi}$ of the A_d for each patent d. Then, the probability of generating a patent is calculated as Eq. (6):

$$\begin{aligned} &P(I_j = i, Z_j = z | W_j = w, A_j = a, Z_{\neg j}, A_{\neg j}, W_{\neg j}, I_d) \propto \\ &\frac{g_{iz} + \alpha}{\sum_Z (g_{iz'} + \alpha)} \cdot \frac{g_{zw} + \beta}{\sum_W (g_{zw'} + \beta)} \cdot \sum_{A_d} \frac{g_{za} + \mu}{\sum_A (g_{za'} + \mu) \cdot |A_d|} \end{aligned} \tag{6}$$

After Gibbs sampling, each word, inventor and applicant of a patent will be sampled to a topic that associates items containing relevant information. The basic idea of the heterogeneous topic model is to group words into various topics according to their semantic meanings. The distribution of topic words ($\boldsymbol{\Phi}$) will represent the collective semantics of words under this topic. That is, the semantics of each word in different topics will have some distinction by clustering the topic of words. However, the ultimate goal of building a heterogeneous topic model is not only to assign the meanings of words to different topics but also to extract the semantics of context-based words.

Hence, we calculate the topic distribution of patents (patent-topic $\boldsymbol{\Gamma}$) by Eq. (7) based on the Gibbs sampling results. The topics have been assigned to patents such that the semantics of the patent full-text is represented by the distribution of topic.

$$\boldsymbol{\Gamma}_{dz} = \frac{g_{dz}}{\sum_Z g_{dz'}} \tag{7}$$

The patent-topic-word structure containing riching in semantic information is build based on $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$. And the information between $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ are integrated by Eq. (8):

$$f_1 = \boldsymbol{\Gamma} \cdot \boldsymbol{\Phi} \tag{8}$$

Here, $\boldsymbol{\Gamma}$ is a matrix of size $|D| \times t$, $\boldsymbol{\Phi}$ is a matrix of size $t \times |W|$, and f_1 is a matrix of size $|D| \times |W|$ parameters. Equation (8) takes the topic as the bridge to establish the connection between patents and words and achieves the semantic extraction of context-based words. The semantics of each word in different patents will have some distinction by clustering the topics of words. Each patent $d \in D$ will obtain their distribution representation of words $\overrightarrow{f_1(d)}$.

Based on the Gibbs sampling results, each word will have different meanings in different contexts, which includes the semantic information of the context and the information of the research interests of inventors and applicants. The heterogeneous topic model that we proposed establishes the association among text, inventors and applicants with the topic as the core, and it also addresses how to obtain the semantic information of words in different contexts.

Word embedding enhances heterogeneous topic model

Word embedding is an effective approach for capturing the semantic information of words and can be used to measure word similarities. For each target patent $tp_j \in TP$ with word sequence W_j, the objective of the word embedding model is to map all words into a low-dimensional space $\mathbb{R}^v, v \ll |W|$. The learned result is able to effectively represent the semantic information in word w. $f_2 : W \to \mathbb{R}^v$ is the mapping function from words to word embeddings. Here, v is a parameter that specifies the number of dimensions, and f_2 is a matrix of size $|W| \times u$ parameters. With the training of the word embedding model, the mapping function $f_2 : W \to \mathbb{R}^v$ is learned for mapping each word $w \in W$ into its vector space and returns the word embedding $\overrightarrow{f_2(w)}$. The similarity of two words w_m and w_n can simply be measured with the inner product of their word embeddings.

As discussed in the Introduction section, we want to exploit the word embedding model to implement the association of different words with the same meaning. To integrate word embedding information into a heterogeneous topic model, the word embedding is regarded as a mapping matrix, and the mapping from patent to vector space is realized by Eq. (9):

$$pvec = f_1 \cdot f_2 \tag{9}$$

where $pvec$ is the patent embedding with size $|D| \times u$. After obtaining patent embedding $pvec$, words with the same meaning are associated, and meanings of the same word in different contexts are distinguished. Then, patents that study the same topic and have similar semantics are closer in vector space.

Based on the distance between patent embeddings, the similarity $Pr(cp_m|tp_n)$ between the target patent $tp_n = (W_n, I_n, A_n)$ and the candidate patent $cp_m = (W_m, I_m, A_m)$ is calculated by cosine similarity:

$$Pr(cp_m|tp_n) = \frac{\overrightarrow{pvec(cp_m)} \cdot \overrightarrow{pvec(tp_n)}}{||\overrightarrow{pvec(cp_m)}|| \cdot ||\overrightarrow{pvec(tp_n)}||} \tag{10}$$

Finally, we recommend the *top k* most similar patents for each target patent. Here, k is a parameter specifying the number of related patents recommended. The process of the

Table 2 Datasets

	Train	Test	Average cited	Inventors	Applicants
A61-Full	25,881	2500	17.5	44,502	6892
A61-Sampling	432	10	15.1	1048	292
USPTO	20,763	1987	10.6	49,442	8368

algorithm is described in Algorithm 1. In Algorithm 1, lines 2 to 9 describe the heterogeneous topic model, lines 10 and 11 describe how to enhance the heterogeneous topic model with word embedding, and lines 12 to 18 describe how the HTW model makes patent recommendations.

In the HTW model, the heterogeneous topic model takes the topic as the core and integrates the text, inventor and applicant information into the patent feature representation containing the context information of words. In addition, the word embedding has been exploited to map patents to the vector space and learn the patent embedding. With the exploitation of word embedding for heterogeneous topic models, the extraction of context-based word semantics and the association of words with the same meaning have been enhanced.

Algorithm 1: HTW model

Input: target patent tp_j; candidate patents CP; $\alpha = 50/t, \beta = 0.01, \mu = 0.01$; k
Output: $D_{relevant}$

1 **begin**
  // algorithm of heterogeneous topic model
2 **Initialize:** Θ, Φ, Ψ;
3 **begin**
4 **foreach** $d \in D$ **do**
5 Update $\Theta_{I_d}, \Phi_Z, \Psi_Z$ via Eq.(6);
6 Calculate Γ_d via Eq.(7);
7 **end**
8 $f_1 = \Gamma \cdot \Phi$;
9 **end**
  // exploiting word embedding for enhanced heterogeneous topic model
10 Calculate f_2 by word embedding model;
11 $pvec = f_1 \cdot f_2$;
  // algorithm of patent recommendation
12 **begin**
13 **foreach** $cp_m \in CP$ **do**
14 Calculate $Pr(cp_m|tp_j)$ via Eq.(10);
15 **end**
16 sort $Pr(tp_j)$ from big to small;
17 $D_{relevant} = Pr(tp_j)[:k]$;
18 **end**
19 **Return** $D_{relevant}$;
20 **end**

Experiments

Dataset

In our experiments, we employ three benchmark datasets: A61-Full, A61-Sampling, and USPTO. The statistics of datasets are listed in Table 2. The test set and the training set are completely separate in the training. We use the model trained by the training set to study the text embedding of the test set. We introduce these three datasets as follows:

A61-Full[4] collects 2,500 patents with the *Cooperative Patent Classification scheme* (CPC) number A61 published in 2015 as the target patents. All patents cited by the 2,500 patents are taken as part of the dataset, and 1,000 patents are randomly selected from the patents never cited by the 2,500 patents as another part of the dataset. If patent d_1 is cited by patent d_2, then the patent pair (d_1, d_2) is labeled "cited"; otherwise, it is labeled "random". Therefore, the first dataset contains 2,470,736 patent pairs with the "cited/random" label. For this dataset, we use the target patents as test set, and the other patents are used as training set.

A61-Sampling[4] is a subset of A61-Full. Ten target patents are extracted from the A61-Full, and the patents cited by these patents are included. At the same time, some patents are randomly selected. Patents in this dataset are manually labeled as "relevant" or "irrelevant". These labels indicate whether the cited patent is relevant to the target patent and whether important prior art is missing from the search results. For this dataset, we use these ten target patents as test set, and the other patents are used as training set.

USPTO dataset is a subset extracted from the entire *United States Patent and Trademark Office* (USPTO) set[5]. We first collect the patent data of 2018 from the complete set of USPTO and randomly selected 2% as the seed patent. Then the corresponding citations of the seed patent are extracted and the citations in the seed patent set are filtered to ensure the separation of the seed patent set and the citation patent set. Finally, we filter for seed patents with less than 5 or more than 25 citations, and the remaining seed patents are used as test sets, while the other patents are used as training sets. This dataset covers all categories of the CPC.

Evaluation

To evaluate the effectiveness of the method, the patent recommendation problem is modeled as a binary classification problem. The experiments of A61-full and A61-sampling datasets follow paper (Helmers et al. 2019), which are evaluated through calculating AUC and AP for the label pairs "cited/random" and "relevant/irrelevant". For the label "relevant/irrelevant", if $Pr(cp_m|tp_n)$ is higher than the threshold 0.5, then patent cp_m is considered a relevant patent; otherwise, it is considered an irrelevant patent. Generally, a patent is similar to its citation (Chen 2017), so the citation information is used to calculate AUC, AP, and Recall to evaluate the experiments on the USPTO dataset. so the experiments of the USPTO dataset are evaluated by using citation information to calculate AUC, AP, and Recall.

[4] https://github.com/helmersl/patent_similarity_search.

[5] http://www.patentsview.org/download

The recall value is obtained according to the recommendation results, and the *receiver operating characteristic* (ROC) curve is plotted according to the recall value. The *area under the ROC curve* (AUC) can be used to evaluate the ability of the recommendation model to distinguish between relevant and irrelevant patents.

Although AUC is an effective measure for distinguishing between relevant and irrelevant, random patents selected in the A61-Full dataset have little overlap with the target patents. Thus, they can easily be distinguished, which means that the AUC value may be high. To further understand the recommendation effect of the algorithm, *average precision* (AP) is used to evaluate our method. According to the similarity, the ordering result returned for patents can be obtained, and precision and recall can be calculated according to the ordering result for all thresholds. The AP score is computed as follows:

$$AP = \sum_{k} (Recall@k - Recall@k - 1) \cdot Precision@k \tag{11}$$

where $Recall@k$ is evaluated by Eq.(12), $Precision@k$ is evaluated by Eq.(13):

$$Recall@k = \frac{1}{|TP|} \sum_{j=1}^{|TP|} \frac{|R_p \bigcap T_p|}{|T_p|} \tag{12}$$

$$Precision@k = \frac{1}{|TP|} \sum_{j=1}^{|TP|} \frac{|R_p \bigcap T_p|}{|R_p|} \tag{13}$$

where R_p is the *top k* patent list recommended based on a target patent tp_j. T_p is the set of patents citing tp_j or relevant to tp_j.

AP can solve the limitation of a single point value of recall and precision and obtain a result that reflects global performance.

Baselines

To verify the effectiveness of the HTW model, we select the baseline from the keyword-based model (TF-IDF), topic model (LSA, ICT) and embedding model (Word2vec, Doc2vec) to conduct experiments:

(1) *TF-IDF* TF-IDF is a kind of BOW model, which can be used for the automatic extraction and extension of keywords. To verify the effect of Word2vec, we conducted two groups of experiments based on TF-IDF using the TF-IDF model alone and combining TF-IDF with Word2vec.

(2) *Latent Semantic Analysis (LSA)* (Deerwester et al. 1990) LSA is an analysis method based on latent semantics proposed by Deerwester. LSA uses the dimensionality reduction method of singular value decomposition (SVD) of a matrix to analyze the structure of a document. Thus, the relevance analysis of the text can be performed in the vector space of dimensionality reduction, and the semantic relations latent in the text can be mined by SVD and so on.

(3) *ICT* (Tang et al. 2012) ICT (inventor-company-topic) and other topic models are developed on the basis of LDA (Blei et al. 2003), while LDA and other topic models introduce hyperparameters to establish a multilayer Bayesian structure on the basis of LSA.

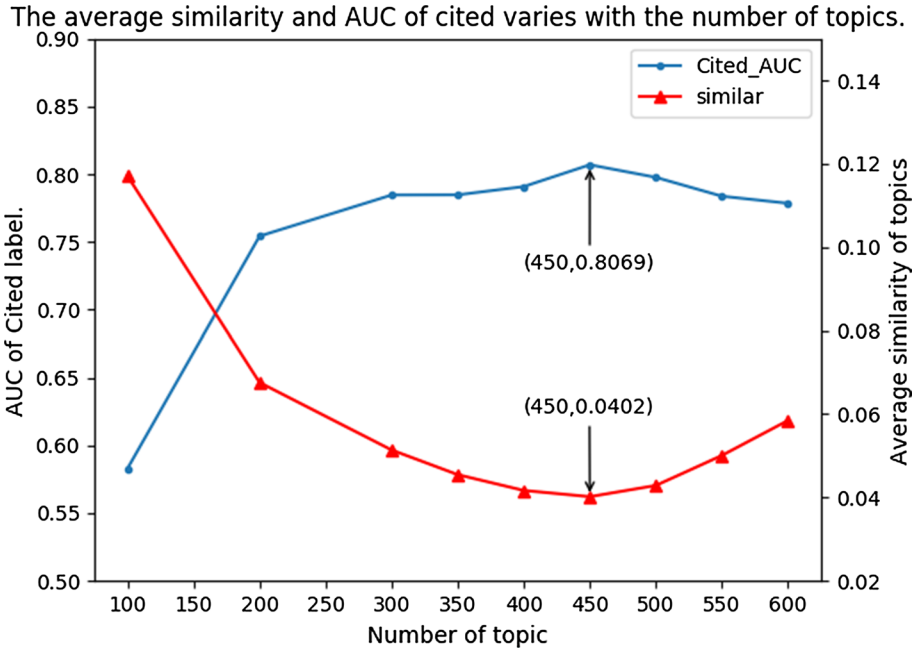

Fig. 3 The average similarity and AUC of cited varies with the number of topics

Most topic models use the Gibbs sampling method to solve and derive the distribution of the model to explore the semantic information of text.

(4) *Doc2vec* (Le and Mikolov 2014) Doc2vec is a method proposed by Le and Mikolov to convert paragraphs or sentences into a feature matrix based on Word2vec. Doc2vec treats paragraphs as separate words and uses the Word2vec approach to train paragraphs.

Experimental results and analysis

The selection of topic number is extremely important to the topic model, and different topic number settings can lead to a huge gap in results. To choose the best number of topics, we calculated the average similarity between topics (Cao et al. 2008):

$$sim(m,n) = \frac{\overrightarrow{\Phi(m)} \cdot \overrightarrow{\Phi(n)}}{||\overrightarrow{\Phi(m)}|| \cdot ||\overrightarrow{\Phi(n)}||} \tag{14}$$

where $sim(m, n)$ represents the similarity between topic m and topic n. $\overrightarrow{\Phi(m)}$ and $\overrightarrow{\Phi(n)}$ are the topic-word distributions of topic m and topic n obtained by the topic model, respectively.

$$avg\_sim = \frac{\sum_{m=1}^{t-1} \sum_{n=m+1}^{t} sim(m,n)}{t \times (t-1)/2} \tag{15}$$

The *avg_sim* is the average similarity between topics.

Table 3 Experiment Results on A61 Dataset

Models	AUC			AP		
	A61-sampling		A61-full	A61-sampling		A61-Full
	Cited	Relevant	Cited	Cited	Relevant	Cited
TF-IDF (Helmers et al. 2019)	0.8063	0.8118	0.9560	0.7095	0.5274	0.4705
LSA (Helmers et al. 2019)	0.7075	0.7798	0.9361	0.5921	0.4787	0.3257
ICT	0.7920	0.8223	0.9435	0.7239	0.5340	0.4105
TF-IDF +W2V (Helmers et al. 2019)	0.8544	0.8408	0.9410	0.7354	0.5443	0.4019
Doc2vec (Helmers et al. 2019)	0.8138	0.7658	0.9314	0.6829	0.4749	0.3121
HTW	**0.8705**	**0.8528**	**0.9631**	**0.7576**	**0.5520**	**0.5037**

Table 4 Experiment Results on USPTO Dataset

Models	USPTO-Abstract					USPTO-Claims				
	R@50	R@75	R@100	AP	AUC	R@50	R@75	R@100	AP	AUC
TF-IDF	0.366	0.417	0.458	0.262	0.926	0.370	0.432	0.476	0.268	0.957
LSA	0.358	0.412	0.452	0.252	0.942	0.384	0.441	0.481	0.296	0.961
ICT	0.336	0.421	0.483	0.269	0.938	0.379	0.458	0.475	0.278	0.940
TF-IDF +W2V	0.420	0.467	0.499	0.322	0.949	0.405	0.452	0.486	0.324	0.955
Doc2vec	0.345	0.403	0.437	0.271	0.932	0.391	0.425	0.443	0.260	0.927
HTW	**0.436**	**0.486**	**0.519**	**0.344**	**0.958**	**0.431**	**0.478**	**0.514**	**0.350**	**0.965**

The paper (Cao et al. 2008) proves that when the average similarity between topics is smaller, the results obtained by the topic model can better distinguish the hidden topic content of different documents, and the effect is better.

We calculate the average topic similarity and the AUC score of different topic numbers in the A61-Sampling dataset by the model proposed in this study, and the results are shown in Fig. 3. As shown in Fig. 3, the A61-Sampling dataset minimizes the average similarity of topics and maximizes the AUC score when the number of topics is 450. The reason is that the dataset is selected under the same classification data. There is a certain correlation between the patents. At the same time, the patent citation data constitute a proportion of this dataset. Finally, we select 450 as the topic number of the A61-Sampling dataset. For the A61-Full and USPTO datasets, we also selected the most suitable topic number of 1000.

Experiments with TF-IDF, ICT, TF-IDF+Word2vec and HTW models are conducted on three datasets. The experimental data of TF-IDF, LSA, TF-IDF+W2V and Doc2vec models on A61-Full and A61-Sampling are obtained from paper (Helmers et al. 2019). The experimental results on A61-Full and A61-Sampling are shown in Table 3. The experimental results on USPTO are shown in Table 4. In Tables 3 and 4, the model names in bold represent the best-performing model, and the numbers in bold represent the best results achieved in this set of experiments.

According to Tables 3 and 4, compared with the topic model(LSA, ICT), HTW performs better because it integrates the word embedding model whose ability to capture

the semantic information from the texts is proved to be very powerful and has been widely used. The HTW model retains the superiority of the word embedding model in semantic information extraction and synonym association. Compared with the embedding method(TF-IDF+W2V, Doc2vec), HTW has also made a preferable improvement because it makes use of the topic distribution feature of the topic model to enhance the difference of polyseme meaning in different contexts. And the topic model is verified to alleviate the problem of polysemy by the topic distribution.

We performed two independent experiments on the A61-Sampling dataset, using the "cited / random" and "relevant / irrelevant" labels as ground truth, respectively. Table 3 shows that HTW performs better in terms of the AUC and AP metrics on the full and sampling datasets. According to the experimental results, compared with the previous topic model LSA with considering text information only, the ICT model that introduces patent network information is more effective in distinguishing relevant patents. However, the topic model is greatly affected by word frequency and cannot filter out high-frequency words. The TF-IDF model considering the inverse document frequency has performed better, even if it only has text information. If we introduce only the multiple patent information into the topic model, such as the ICT model, the results are often inferior to embedding models, such as Word2vec. This means that the embedding model captures deeper semantic information better than the topic model. However, the AP value of the embedding methods on the large dataset A61-Full is lower than that of the topic model. The reason for this result is that the influence of word ambiguity increases with the increase in the number of texts while embedding methods are not capable of distinguishing word ambiguity.

We conducted two groups of experiments on the USPTO dataset, and the experimental results are shown in Table 4. In the first group of experiments(USPTO-Abstract), titles and abstracts are employed as the corpus, while in the second group(USPTO-Claims), title, abstract, and claims are employed as the corpus for training. The results, as shown in Table 4, indicate that HTW is superior to all baselines in the performance of patent recommendation tasks. Comparing the results of the two experiments, it can be seen that although the second group supplements the data of the claim in the corpus, the results of the second group are not significantly improved, and even the effects of some algorithms are reduced. The patent claim is a long text, which contains the details of the technology and is also the core of the patent. It is generally believed that the claim contains more comprehensive and accurate information than the abstract. Unfortunately, none of the algorithms in the experiment seemed to make effective use of the information provided in the claim. This is because of the particularity of the patent, there are too many technical terms in the claim, which makes it difficult to learn the semantics of the patent text. Besides, long text learning is also very challenging research.

All these results prove the validity of the HTW model in the patent recommendation. This is because the HTW model exploits the word embedding for enhanced heterogeneous topic models while considering the heterogeneity of patent data and combining the advantages of topic models and word embedding. The HTW model uses word embeddings to learn the deep semantic information of texts to distinguish word ambiguity and solves the problem of similar words with the help of the topic model. At the same time, HTW introduces inventor and applicant information, which contains rich and useful insights into technological and economic value and can be used to extract more accurate topic distributions for patents.

Target Patent	HTW	TFIDF	ICT	TFIDF+Word2vec
Spinous Process Fusion Devices (34citations)	[1] Spinous Process Fusion Devices and Methods (✓) [2] Adjustable spine distraction implant (✓) [3] Adjustable spinous process spacer device and method of treating spinal disorders (✓) [4] Percutaneous spinous process fusion plate assembly and method (✓) [5] Modular **interspinous** fixation system and method(✓) [6] Spinous process plate and connector assembly and method(✓) [7] Anterior lateral spine cage-plate fixation device and technique (×) [8] Spinous process mounted spinal implant (✓) [9] Spinous process fixation devices and methods of use(✓) [10] Fusion implants and systems for posterior lateral procedures (×)	[1] Spinous Process Fusion Devices and Methods (✓) [2] Orthodontic twin bracket with archwire floor and side wall relief (×) [3] Adjustable spine distraction implant (✓) [4] Spinous process fixation implant (✓) [5] Range of motion table (×) [6] Open body box form interbody fusion cage (×) [7] Lip seal oral device (×) [8] Telescoping flexible bite jumping device (×) [9] Polymer rods for spinal applications (×) [10] Stabilization device for stabilizing bones of a vertebra and rod connector(×)	[1] Spinous Process Fusion Devices and Methods (✓) [2] Stabilization **device** for stabilizing bones of a vertebra and rod connector (×) [3] Orthodontic twin bracket with archwire floor and side wall relief(×) [4] Vertebral rod for spinal osteosynthesis instrumentation and osteosynthesis (✓) [5] Spinous process fixation implant, An implantable spinous process fixation device includes an elongated component (✓) [6] Universal transverse connector **device**(×) [7] Spinous process plate and connector assembly and method(✓) [8] Fusion implants and systems for posterior lateral procedures (×) [9] Open body box form interbody fusion cage (×) [10] Percutaneous spinous process fusion plate assembly and method(✓)	[1] Spinous process fixation implant (✓) [2] Percutaneous spinous process fusion plate assembly and method (✓) [3] Adjustable spine distraction implant (✓) [4] Adjustable spinous process spacer device and method of treating spinal disorders(✓) [5] Spinous Process Fusion Devices and Methods(✓) [6] Spinous process fixation devices and methods of use(✓) [7] System and Methods for Spinous Process Fusion (✓) [8] Interbody device and plate for spinal stabilization and instruments (×) [9] Support insert associated with **spinal** vertebrae (×) [10] Threaded center line cage with winged end cap(×)
Precision	0.8	0.3	0.5	0.7

Fig. 4 A case study of recommendation results

Case study

A case-study approach is used to highlight the improvement of HTW in distinguishing the multiple meanings of a word and associating different words with the same meaning. Figure 4 shows an example of the *top* 10 recommendation results for the target patent: "Spinous Process Fusion Devices" in the A61-Sampling dataset. HTW correctly recommends the patent "Modular interspinous fixation system and method" at position 5, although it uses the keywords "interspinous" to express the meaning of "spinous process" and "fixation" to express the meaning of "immobilize" in the abstract. This result demonstrates that HTW can extract the semantics of words and associate different words that express the same meaning. The TF-IDF model is unable to associate the two words.

For the ICT model, high-frequency words have much influence, which results in higher values for these words in the patent-word distribution. If the target patent includes high-frequency words that are not the keywords, the ICT model may still recommend patents that contain these words. This situation will affect the identification of related patents. In this case, the ICT model is clearly influenced by the high-frequency word "device". Therefore, the *top* 10 results of ICT have many irrelevant patents that contain word "device", such as "Stabilization device for stabilizing bones of a vertebra and rod connector" and "Universal transverse connector device".

As shown, the patent "Support insert associated with spinal vertebrae", which is not cited by the target patent, is recommended by the Word2vec model at position 9. The keyword "spinal" in this patent means "of or relating to the spine". The word "spinal" also means "having sharp points like needles" in other contexts, which is the same as the word "spinous". As shown in Fig. 4, Word2vec classifies "spinous" and "spinal" as words with the same meaning, which leads to a high level of similarity between the two patents in the results calculated by Word2vec. However, the phrase "spinous process" is a proper noun in

the medical field that is different from "spinous" and "spinal". The subjects of the two patents are not consistent. This case illustrates that the HTW model has the ability to simultaneously associate words with the same meaning and distinguish different meanings of the same word in different contexts.

Conclusion

In this paper, we present a novel patent recommendation model called HTW, which integrates word embedding representation into a heterogeneous topic model and then exploits the word embedding for an enhanced heterogeneous topic model. The topical aspects of texts, inventors and applicants are simultaneously introduced into HTW, which can better extract the embedding of patents. With the realization of HTW, different words with the same meaning and the multiple meanings of the same word in different contexts can be identified simultaneously.

We evaluated the HTW model with five baseline methods. The experimental results show that the best performance is achieved by the HTW model. The HTW model outperforms the five baseline methods by an average of 3.42% for AUC, 15.8% for AP, and 12.4% for Recall. The experimental results demonstrate that the HTW model can extensively improve the performance compared with the baseline methods on the patent recommendation.

For future work, there are several interesting directions that deserve further exploration, such as patent citation recommendation in patent heterogeneous information networks that integrate semantic and structural information. In addition, we will research how to effectively use the full patent information to prove that the semantic information of the full patent rather than just the abstract information is very important for recommending relevant patents.

Acknowledgements This work is supported by the National Key Research and Development Program of China (2017YFB1401903), National Natural Science Foundation of China (Grants #61876001 and #61673020), the Major Program of the National Social Science Foundation of China (Grant No.18ZDA032), and the Provincial Natural Science Foundation of Anhui Province (#1708085QF156).

References

Arts, S., Hou, J., Carlos Gomez, J. (2019). Text mining to measure novelty and diffusion of technological innovation. In: *17th International conference on scientometrics & informetrics (ISSI2019), vol II*, pp. 1798–1800.

Arts, Sam, Cassiman, Bruno, & Gomez, Juan Carlos. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, *39*(1), 62–84.

Bashir, S., & Rauber, A. (2010). Improving retrievability of patents in prior-art search. In: *European Conference on Information Retrieval*, pp. 457–470. Springer.

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993–1022.

Cao, Juan, Zhang, Yong-Dong, Li, Jintao, & Tang, Sheng. (2008). A method of adaptively selecting best lda model based on density. *Chinese Journal of Computers*, *31*(31), 1780–1787.

Chen, Lixin. (2017). Do patent citations indicate knowledge linkage? the evidence from text similarities between patents and their citations. *Journal of Informetrics*, *11*(1), 63–79.

Chen, Baitong, Ding, Ying, & Ma, Feicheng. (2018). Semantic word shifts in a scientific domain. *Scientometrics*, *117*(1), 211–226.

Choi, Hayoung, Seunghyun, Oh, Choi, Sungchul, & Yoon, Janghyeok. (2018). Innovation topic analysis of technology: The case of augmented reality patents. *IEEE Access*, *6*, 16119–16137.

Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., & Harshman, Richard. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Ganguly, D., Leveling, J., Magdy, W., & Jones, G. J. (2011). Patent query reduction using pseudo relevance feedback. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1953–1956. ACM.

Golestan Far, M., Sanner, S., Bouadjenek, M. R., Ferraro, G., & Hawking, D. (2015). 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*, pp. 803–806. ACM.

Griffiths, Thomas L., & Steyvers, Mark. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Helmers, Lea, Horn, Franziska, Biegler, Franziska, Oppermann, Tim, & Müller, Klaus-Robert. (2019). Automating the search for a patent's prior art with a full text similarity search. *PloS ONE*, *14*(3), e0212103.

Krestel, R., & Smyth, P. (2013). Recommending patents based on latent topics. In:*Proceedings of the 7th ACM conference on Recommender systems*, pp. 395–398.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In: *International conference on machine learning*, pp. 1188–1196.

Li, Shaobo, Jie, Hu, Cui, Yuxin, & Jianjun, Hu. (2018). Deeppatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics*, *117*(2), 721–744.

Li, Guancheng, Lai, Ronald, D'Amour, Alexander, Doolin, David M., Sun, Ye, Torvik, Vetle I., et al. (2014). Disambiguation and co-authorship networks of the us patent inventor database (1975–2010). *Research Policy*, *43*(6), 941–955.

Liu, Y., Liu, Z., Chua, T. S., & Sun, M. (2015). Topical word embeddings. In: *Twenty-ninth AAAI conference on artificial intelligence*.

Lupu, M., Piroi, F., & Stefanov, V. (2017). An introduction to contemporary search technology. In: *Current challenges in patent information retrieval* (pp. 47–73). Springer, Berlin, Heidelberg

Mahdabi, P., & Crestani, F. (2012). Learning-based pseudo-relevance feedback for patent retrieval. In: *Information retrieval facility conference*, pp. 1–11. Springer.

Mahdabi, P., Keikha, M., Gerani, S., Landoni, M., & Crestani, F. (2011). Building queries for prior-art search. In: *Information retrieval facility conference*, pp. 3–15. Springer.

Mikolov, T., Sutskever, I., Chen, K., Corrado, Greg S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*, pp. 3111–3119.

Oh, S., Lei, Z., Lee, W. C., Mitra, P., & Yen, J. (2013). Cv-pcr: a context-guided value-driven framework for patent citation recommendation. In: *Proceedings of the 22nd ACM international conference on information & knowledge management*, pp. 2291–2296. ACM.

Shalaby, W., & Zadrozny, W. (2018). Toward an interactive patent retrieval framework based on distributed representations. In: *The 41st International ACM SIGIR conference on research & development in information retrieval*, pp. 957–960. ACM.

Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. *Knowledge and Information Systems*,. https://doi.org/10.1007/s10115-018-132.

Singh, Jagendra, & Sharan, Aditi. (2016). Relevance feedback-based query expansion model using ranks combining and word2vec approach. *IETE Journal of Research*, *62*(5), 591–604.

Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., Gao, B., Huang, M., Xu, P., Li, W., et al. (2012). Patentminer: topic-driven patent analysis and mining. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1366–1374. ACM.

Tannebaum, W., & Rauber, A. (2015). PatNet: a lexical database for the patent domain. In: *European conference on information retrieval* (pp. 550–555). Springer, Cham.

Verma, M., & Varma, V. (2011). Applying key phrase extraction to aid invalidity search. In: *Proceedings of the 13th international conference on artificial intelligence and law*, pp. 249–255. ACM.

Wang, F., & Lin, L. (2017). Exploiting semantic knowledge base for patent retrieval. In: *2017 13th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)*, pp. 2195–2200. IEEE.

Wang, F., Lin, L., Yang, S., & Zhu, X. (2013). A semantic query expansion-based patent retrieval approach. In:*2013 10th International conference on fuzzy systems and knowledge discovery (FSKD)*, pp. 572–577.

Xue, X., & Croft, W. B. (2009). Automatic query generation for patent search. In: *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 2037–2040. ACM.

Zhang, Longhui, Liu, Zheng, Li, Lei, Shen, Chao, & Li, Tao. (2018). Patsearch: An integrated framework for patentability retrieval. *Knowledge and Information Systems*, *57*(1), 135–158.