A Novel Method for Topic Linkages Between Scientific Publications and Patents

Shuo Xu
Beijing University of Technology, Beijing, 100124, China. E-mail: pzczxs@gmail.com; xushuo@bjut.edu.cn

Dongsheng Zhai
Beijing University of Technology, Beijing, 100124, China. E-mail: zhaidongsheng@bjut.edu.cn

Feifei Wang
Beijing University of Technology, Beijing, 100124, China. E-mail: feifeiwang@bjut.edu.cn

Xin An
Beijing Forestry University, Beijing, 100083, China. E-mail: anxin@bjfu.edu.cn

Hongshen Pang
Shenzhen University, Shenzhen, 518060, China. E-mail: phs@szu.edu.cn

Yirong Sun
Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, 510530, China. E-mail: sunyr1206@163.com

It is increasingly important to build topic linkages between scientific publications and patents for the purpose of understanding the relationships between science and technology. Previous studies on the linkages mainly focus on the analysis of nonpatent references on the front page of patents, or the resulting citation-link networks, but with unsatisfactory performance. In the meanwhile, abundant mentioned entities in the scholarly articles and patents further complicate topic linkages. To deal with this situation, a novel statistical entity-topic model (named the CCorrLDA2 model), armed with the collapsed Gibbs sampling inference algorithm, is proposed to discover the hidden topics respectively from the academic articles and patents. In order to reduce the negative impact on topic similarity calculation, word tokens and entity mentions are grouped by the Brown clustering method. Then a topic linkages construction problem is transformed into the well-known *optimal transportation problem* after topic similarity is calculated on the basis of symmetrized Kullback–Leibler (KL) divergence. Extensive experimental results indicate that our approach is feasible to build topic linkages with more superior performance than the counterparts.

Received November 22, 2017; revised October 26, 2018; accepted November 11, 2018

 DOI: 10.1002/asi.24175

Introduction

It is well known that science and technology have different purposes, and ways of viewing and knowing the world. In order to understand the relationships between science and technology, several theoretical models have been developed in the literature from the viewpoint of technology-oriented innovation processes, such as the linear or pipeline model (Carpenter & Narin, 1983; Narin, 1994), and two-branched model (Rip, 1992). In the former model, new discoveries in scientific research lead to technological ideas, which further promotes companies to develop novel technologies and then to industrialize. This model unilaterally emphasized the influence of science on technology development (Tijssen, 2001) and ignored that technology often shaped science in important ways (Bhattacharya, Kretschmer, & Meyer, 2003). The more rational latter model views the multifaceted relationships between science and technology as two parallel branches of activities, which have many interdependencies and cross-connections, but whose internal linkages are much

stronger than their cross ones, like a pair of dancers (Narin & Noma, 1985) or two strands of DNA (Brooks, 1994). In the meanwhile, as the innovation cycle shortens, cross-linkages are becoming stronger and stronger.

Scientific publications and patents are usually considered good barometers or indicators of basic scientific research and technical development, respectively (Calero-Medina & Noyons, 2008; Dubarić, Giannoccaro, Bengtsson, & Ackermann, 2011; Xu et al., 2017). Therefore, there should be some hidden interactive and exclusive relations between the scientific publications and patents. Moreover, according to multitask learning theory (Caruana, 1997; Xu, An, Qiao, & Zhu, 2014), it may be advantageous to analyze simultaneously both information resources from an interested research field, since this field can be observed from multiple views. Along with the evolution of the theories and models that delineate the science–technology interaction, abundant literature to quantify the science–technology interaction (Albert, 2016; Klitkou & Gulbrandsen, 2010; Xu, Zhu, Qiao, Shi, & Gui, 2012) have shown that the linkages between scientific publications and patents, pioneered by Narin and his colleagues (Narin, Hamilton, & Olivastro, 1997; Narin & Olivastro, 1992; Narin & Olivastro, 1998), are indeed helpful to detect the technology opportunity (Albert, 2016; Lee et al., 2011), understand university–industry–government interaction (Leydesdoref & Meyer, 2007; Tian, 2015), measure innovation (Jibu, 2011), construct high-quality roadmaps (Kostoff & Schaller, 2001), and so on.

In fact, previous studies on linkage between the scientific publications and patents mainly focused on the analysis of nonpatent references (NPRs) on the front pages of patents. The main problems with the previous studies are three-fold: (i) The differences among countries in terms of how NPRs are cited (Michel & Bettels, 2001); (ii) Only about 30–40% of patents contain NPRs (Callaert, Looy, Verbeek, Debackere, & Thijs, 2006); (iii) The contents of patents and cited papers are not always related (Meyer, 2000). In our opinion, the research framework described in Shibata, Kajikawa, and Sakata (2010, 2011) and Xu et al. (2012) can effectively bypass these problems. In addition, the scholarly articles and patents often contain many entity mentions with different categories, such as drugs, formulas, and proteins in the biomedical literature. Figure 1 shows titles and abstracts of scholarly articles and patent samples, annotated explicitly with entity mentions by *BRAT* (Stenetorp et al., 2012), which is an intuitive web-based tool for text annotation supported by natural language processing (NLP) technology. It further complicates topic linkages, and the method by Xu et al. (2012) did not explicitly take the entity mentions into consideration, so Xu et al. (2012) will be generalized for this scenario in this study.

This work still follows the three-step procedure (cf. the last paragraph in *Linkage Between Scientific Publications and Patents* subsection, below) described in Xu et al. (2012), but with the following contributions: (i) A novel statistical entity-topic model, armed with the collapsed Gibbs sampling inference algorithm, is proposed to detect the hidden topics respectively from the scientific publications and patents; (ii) In order to reduce the negative impact on topic similarity calculation from the different statements in the academic articles and patents, word tokens, and entity mentions are grouped by the Brown clustering method (Brown, deSouza, Mercer, Pietra, & Lai, 1992; Liang, 2005). And then topic linkages construction problem is transformed into the well-known *optimal transportation problem* (Hillier & Lieberman, 1995; Rachev & Ruschendorf, 1998) after topic similarity is calculated on the basis of symmetrized Kullback–Leibler (KL) divergence. Finally, the experimental results and comparisons in different settings make clear that it is feasible to construct topic linkages between the scientific publications and patents mentioning many entities.

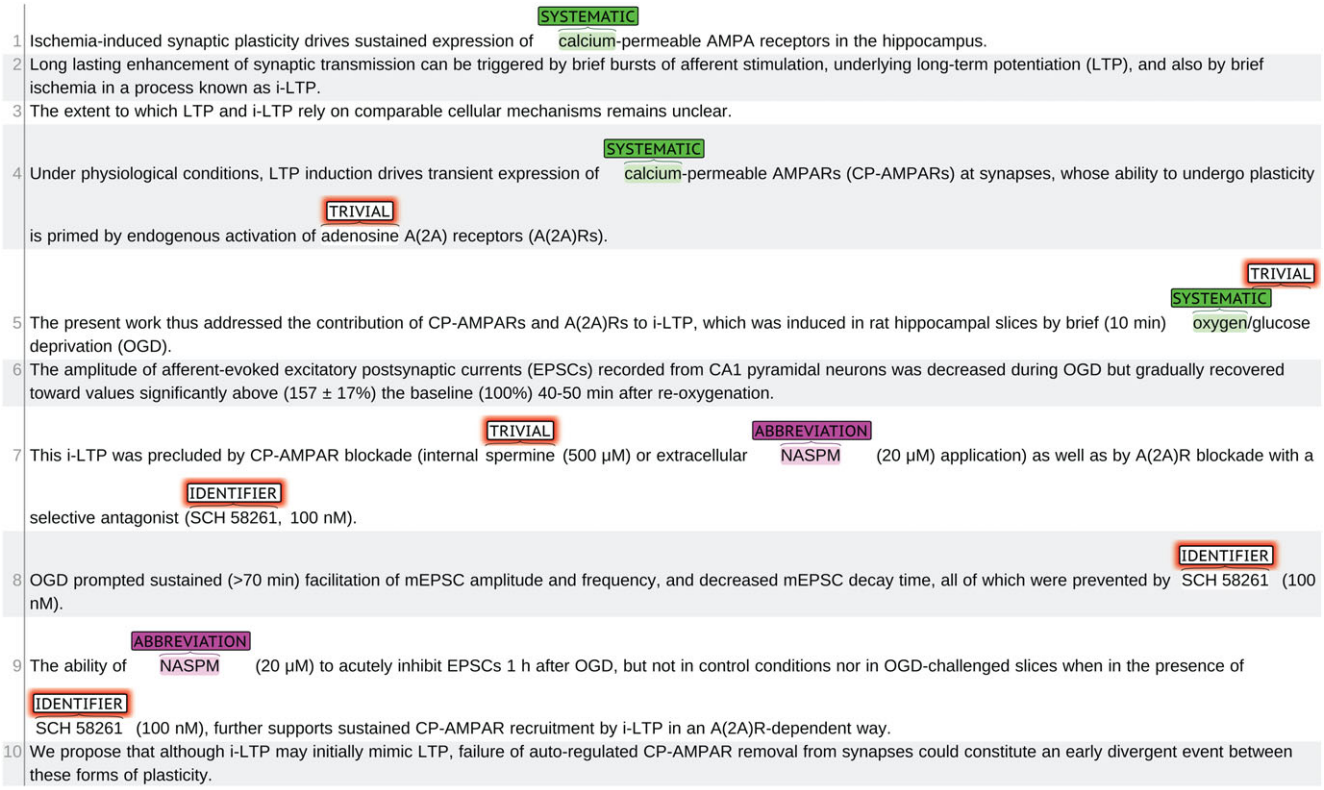
Literature Review

Linkage Between Scientific Publications and Patents

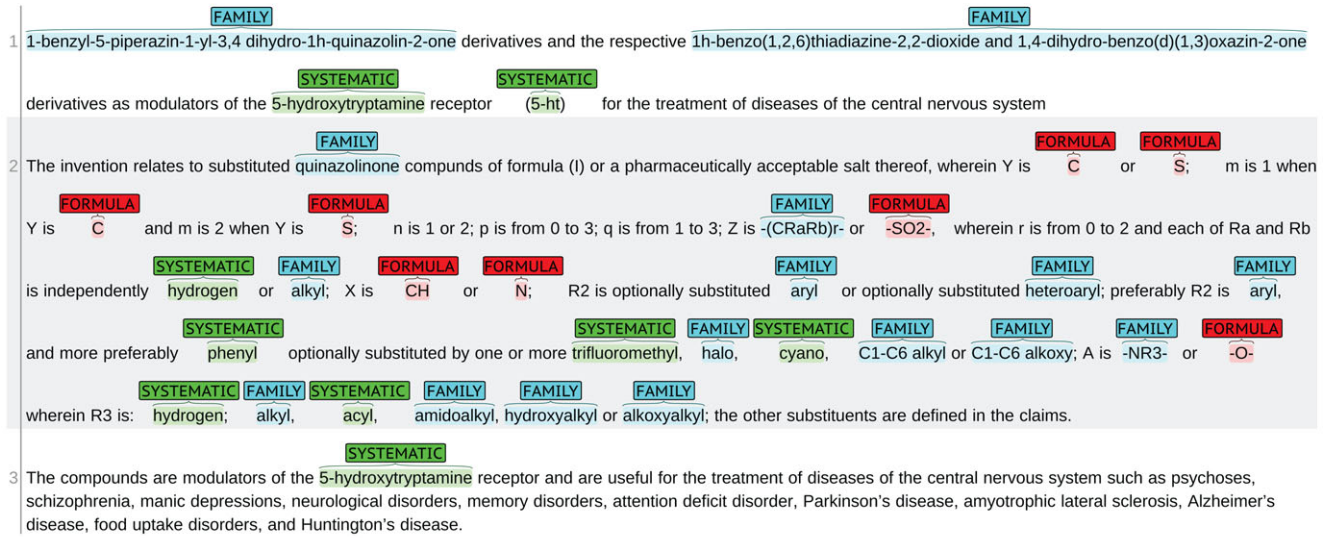
Patents can be viewed as materializations of technologies (Tijssen, 2001). As one of the few useful sources of comprehensive empirical information, patent documents describe the technical details of the resulting inventions and may contain explicit references to the scholarly articles in the form of a nonpatent citation. These NPRs offer a proxy measure for the industrial relevance of research (Tijssen, 2001). The NPR-based approach for analyzing science–technology linkages dates to the pioneering work by Narin and colleagues at Computer Horizons (CHI) (Hicks, Tomizawa, Saitoh, & Kobayashi, 2004; Narin et al., 1997; Narin & Olivastro, 1992; Narin & Olivastro, 1998), who eventually linked all science-based NPRs to specific articles in *Science Citation Index (SCI)* journals starting with the 1983 patent year (Narin & Olivastro, 1998). Meyer also conducted a great amount of work on the basis of NPRs (2000, 2002), much of which concentrates on the area of nanotechnology (Meyer, 2001). Although the number of patents cited by scientific publications is much lower relative to other citation types, the citations of patents in the scientific literature were explored by Glänzel and Meyer (2003). The chemistry-related subfields dominate the citations from articles to patents, which disables this technique to be applicable to other fields.

As a matter of fact, whichever patent-to-paper citations or paper-to-patent citations, the links between the scientific publications and patents are noisy because the subject of citation behavior varies from authors and inventors to reviewers and examiners. For the purpose of the identification and ultimately reduction of this noise, the different citing motivations of examiners and inventors were investigated by Li, Chambers, Ding, Zhang, and Meng (2014). They found that the nonself-citations by inventors are quite noisy and cannot indicate the linkages between science and technology and that self-citations by inventors are more appropriate for understanding the linkages.

DOI: 10.1002/asi



(a) Title (line 1) and abstract of a paper sample with PMID = 23041538.



(b) Title (line 1) and abstract of a patent sample with no. = EP1708713B1.

FIG. 1. Scientific publication and patent samples annotated with entity mentions. (**a**) Title (line 1) and abstract of a paper sample with PMID = 23041538. (**b**) Title (line 1) and abstract of a patent sample with no. = EP1708713B1. [Color figure can be viewed at wileyonlinelibrary.com]

Lexical- or topic-based approaches have also been used to establish the linkages between scientific articles and patents. Bassecouolard and Zitt (2004) established the correspondence tables between patent classes and scientific disciplines. A similar research framework, followed by Shibata et al. (2010, 2011) and Xu et al. (2012), is adopted to construct topic linkages between the scientific publications and patents: (i) to extract respective topics in the scientific publications and patents; (ii) to calculate topic similarities; (iii) to construct topic linkages. The main difference between them lies in topic extraction methods: a citation-link approach is utilized in Shibata et al. (2010, 2011), and a text-mining approach is used in Xu et al. (2012). Additionally, in the linkage construction step, only manual examination is involved in Shibata et al. (2010, 2011), but the optimal transportation problem

DOI: 10.1002/asi

solver is introduced in Xu et al. (2012). In fact, as pointed out in Shibata et al. (2010, 2011), the main problem of the citation-link approach is that many patents are eliminated from the giant component and could not be included in the analyzed data. Although the unconnected component inclusion technique can increase the coverage of patents, on average, up to 57.9% of patents inside the original network could be connected (Takano, Mejia, & Kajikawa, 2016). Hence, this work prefers to the text-mining approach.

Statistical Entity-Topic Models

The topic model is a family of generative probabilistic models for discovering the main themes from a collection of documents (Blei, 2012). Examples of topic models include Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), the Dynamic Topic Model (DTM) (Blei & Lafferty, 2006), the Hierarchical Dirichlet Process (HDP) (Teh, Jordan, Beal, & Blei, 2006), and many others. Several successful applications, which involve the field of bibliometrics/scientometrics (Nichols, 2014; Suominen & Toivanen, 2016; Yau, Porter, Newman, & Suominen, 2014), show the power of these models on a wide variety of text collections, such as academic articles, patents, e-mails, webpages, and so on. However, on closer examination it is not difficult to find that these models cannot explicitly consider entities mentioned in the text. Since entity mentions are usually proper nouns consisting of multiple words, the naïve method is to bind the resulting entity mentions as a preprocessing step. To say it in another way, the distinction between words and entities is not made during learning, but post-hoc. Therefore, the predictive performance about entities is not satisfactory (Newman, Chemudugunta, & Smyth, 2006).

In order to deal with this problem, Newman et al. (2006) put forward two statistical entity-topic models: SwitchLDA and CorrLDA2. By comparing these two models with other similar ones, such as CI-LDA (Cohn & Hofmann, 2001) and CorrLDA1 (Blei & Jordan, 2003), Newman et al. (2006) found that the CorrLDA2 model outperforms other models on the entity prediction task. As a matter of fact, entities mentioned in the text usually belong to different categories, such as SYSTEMATIC, FAMILY, FORMULA, and TRIVIAL in Figure 1, but the above models disregard direct category information. Therefore, this work will generalize the CorrLDA2 model to learn the relationship between topics, entity mentions, and categories of entity mentions.

TABLE 1. The involved information for different viewpoints.

Viewpoint	Constraint	Weight	Full word list
multinomial distribution	•	•	•
multi-dimensional vector	○	•	•
ranked list	○	○	•
topic word subset	○	○	○

Topic Similarity Measurement

To construct topic linkages, the similarity between a pair of topics from the scientific publications and patents should be measured. There are a large number of similarity measuring approaches in bibliometrics for topic analysis (Boyack et al., 2011; Leydesdorff, 2008; Zhang et al., 2016). Since the topics are extracted with topic models in this work, many similarity measuring approaches are not applicable to this case, such as TF-IDF cosine, BM25, and so on.

In the topic models, a topic is modeled as a multinomial distribution of words from some basic vocabulary. Therefore, it is very natural for symmetrized KL divergence (Heinrich, 2009; Newman, Asuncion, Smyth, & Welling, 2009) and Jensen–Shannon (JS) divergence (Boyack et al., 2011; Heinrich, 2009; Mimno, Wallach, Naradowsky, Smith, & McCallum, 2009) to be used for measuring the similarity (to be precise, dissimilarity) between topics. In fact, each topic can be also viewed as a multidimensional vector, where each dimension denotes the probability of the resulting word in the interested topic. Thus, cosine similarity (He et al., 2009; Zhang et al., 2016) and dot product (Boyack et al., 2011) can be used here. If all words are sorted in ascending or descending order by topic-word probability, one can interpret each topic as a ranked list of words. Spearman's rank order correlation coefficient (Spearman's ρ) (Press, Teukolsky, Vetterling, & Flannery, 1992) and Kendall's τ (Press et al., 1992) are very applicable for this perspective. In addition, a subset of topic words (words with a probability over a threshold or top words that contribute a cumulative probability mass over a threshold) can be also utilized to denote the specific topic. As for this viewpoint, Jaccard's coefficient (Jain & Dubes, 1988; Shibata et al., 2011) is usually preferred.

It is very easy to see that each metric treats the similarity between topics from different views. The symmetrized KL divergence and JS divergence consider the divergence of two multinomial distributions, and lower divergence would indicate higher similarity between topics. The cosine similarity and dot product measure the angle of two vectors. Spearman's ρ and Kendall's τ consider the ranks of words in a topic, and Jaccard's coefficient focuses on the association between two topic word subsets. From the point of meaning of a topic in the topic models, apart from the symmetrized KL divergence and JS divergence, each metric has some information loss to some extent. For example, the multidimensional vector viewpoint just drops off the constraint of topic-word probability summing to 1; the ranking list viewpoint does not consider the probability weights; and the topic word subset viewpoint may truncate the word list. Table 1 summarizes the involved information for different viewpoints.

Research Framework and Methodology

As shown in Figure 2, our research framework consists of three phases. The first is to detect sentence boundaries,

DOI: 10.1002/asi

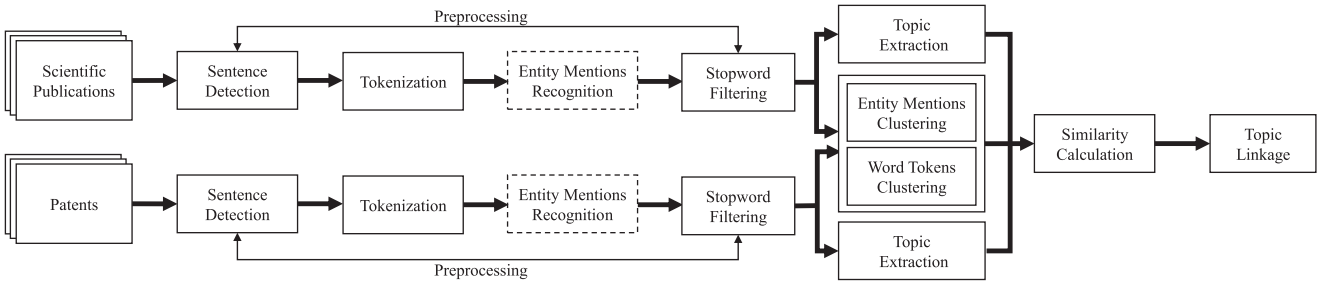


FIG. 2. Research framework of topic linkages aware of entities. (**a**) CorrLDA2 Model. (**b**) CCorrLDA2 Model.

tokenizes each detected sentence, recognizes entity mentions, and then filters stopwords as preprocessing. If entity mentions have been annotated manually or identified in advance, the substep *entity mentions recognition* can just be skipped. At the second phase, hidden topics are extracted from the scientific publications and patents, in which word tokens and entity mentions need to be considered simultaneously. The different statements in academic articles and patents enable many word tokens and entity mentions to only appear in one resource. In order to reduce the negative impact on topic similarity calculation, word tokens and entity mentions are grouped. The final phase is to calculate the similarities between topic pairs from the scientific publications and patents, and then to build topic linkages. We describe the second and third phases in more detail in the following subsections.

Topic Extraction

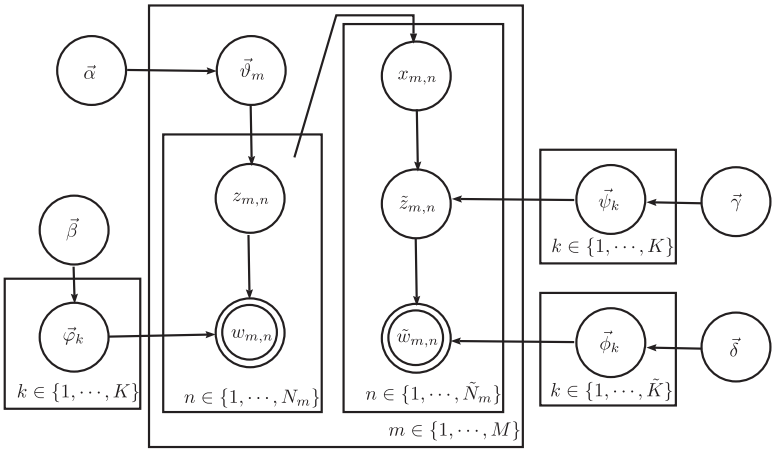
In this subsection, the CorrLDA2 model (Newman et al., 2006) is generalized to learn the relationship between topics discussed in the corpus of scientific publications or patents, entities mentioned in each document, and categories of entity mentions by introducing two sets of hidden random variables, $\vec{y}$ and $\vec{\xi}$. This novel model, renamed CCorrLDA2, degenerates to the CorrLDA2 model when all interested entities belong to one category (that is, $C = 1$). If the corpus mentions no entities at all, both CCorrLDA2 and CorrLDA2 models degenerate to the standard LDA model (Blei et al., 2003). That is to say, the CorrLDA2 and LDA models are two special cases of our model. It is worth noting that the idea in the CCorrLDA2 model is also applicable to other similar models described in the *Statistical Entity-Topic Models* subsection, above. For convenience, the notation is summarized in Table 2, and the graphical model representations of the CorrLDA2 and CCorrLDA2 models are shown in Figure 3. One can also describe the CCorrLDA2 model from the viewpoint of a generative process as follows:

1. For each topic $k \in \{1, \cdots, K\}$, draw $\vec{\varphi}_k \sim \mathrm{Dir}\left(\vec{\beta}\right)$ and $\vec{\xi}_k \sim \mathrm{Dir}\left(\vec{\mu}\right)$, respectively;
2. For each topic $k \in \{1, \cdots, K\}$ and each class $c \in \{1, \cdots, C\}$, draw $\vec{\psi}_{k,c} \sim \mathrm{Dir}\left(\vec{\gamma}\right)$;
3. For each topic $\tilde{k} \in \left\{1, \cdots, \tilde{K}\right\}$, draw $\vec{\phi}_{\tilde{k}} \sim \mathrm{Dir}\left(\vec{\delta}\right)$;
4. For each document $m \in \{1, \cdots, M\}$, draw $\vec{\vartheta}_m \sim \mathrm{Dir}\left(\vec{\alpha}\right)$;
5. For each document $m \in \{1, \cdots, M\}$ and each word token $n \in \{1, \cdots, N_m\}$ in the document m, draw $z_{m,n} \sim \mathrm{Mult}\left(\vec{\vartheta}_m\right)$ and then $w_{m,n} \sim \mathrm{Mult}\left(\vec{\varphi}_{z_{m,n}}\right)$;
6. For each document $m \in \{1, \cdots, M\}$ and each entity mention $\tilde{n} \in \left\{1, \cdots, \tilde{N}_m\right\}$ in the document m, draw a super-topic $x_{m,\tilde{n}} \sim \mathrm{Unif}(z_{m,1}, \cdots, z_{m,N_m})$, $y_{m,\tilde{n}} \sim \mathrm{Mult}\left(\vec{\xi}_{x_{m,\tilde{n}}}\right)$, $\tilde{z}_{m,\tilde{n}} \sim \mathrm{Mult}\left(\vec{\psi}_{x_{m,\tilde{n}}, y_{m,\tilde{n}}}\right)$ and then $\tilde{w}_{m,\tilde{n}} \sim \mathrm{Mult}\left(\vec{\phi}_{\tilde{z}_{m,\tilde{n}}}\right)$.

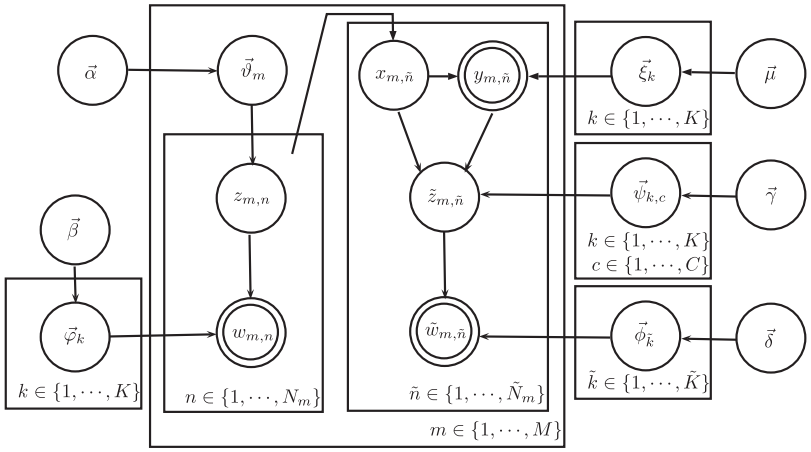
TABLE 2. Notations used in the CorrLDA2 and CCorrLDA2 models.

Symbol	Description
$K, \tilde{K}$	number of word and entity topics, respectively
M, C	number of documents and unique entity classes, respectively
$V, \tilde{V}$	number of unique words and entities, respectively
$N_m, \tilde{N}_m$	number of word tokens and entity mentions in the document m, respectively
$\vec{\vartheta}_m$	multinomial distribution of topics specific to the document m
$\vec{\phi}_k$	multinomial distribution of words specific to the word topic k
$\vec{\varphi}_{\tilde{k}}$	multinomial distribution of entities specific to the entity topic $\tilde{k}$
$\vec{\psi}_{k,c}$	multinomial distribution of entity topics specific to the word topic k and the entity class c
$\vec{\xi}_k$	multinomial distribution of entity classes specific to the word topic k
$z_{m,\,n}$	word topic associated with the n-th word token in the document m
$\tilde{z}_{m,\tilde{n}}$	entity topic associated with the $\tilde{n}$-th entity mention in the document m
$x_{m,\tilde{n}}$	super-topic associated with the $\tilde{n}$-th entity mention in the document m
$y_{m,\tilde{n}}$	class associated with the $\tilde{n}$-th entity mention in the document m
$w_{m,n}, \tilde{w}_{m,\tilde{n}}$	n-th word token and $\tilde{n}$-th entity mention in the document m, respectively
$\vec{\alpha}, \vec{\beta}, \vec{\delta}, \vec{\gamma}, \vec{\mu}$	hyperparameters

DOI: 10.1002/asi



(a) CorrLDA2 Model



(b) CCorrLDA2 Model

FIG. 3. The graphical model representation of (**a**) CorrLDA2 and (**b**) CCorrLDA2 models.

As with many well-known topic models, posterior inference cannot be done exactly in this model. A variety of approximate inference algorithms have appeared in recent years, such as the mean-field variational method (Jordan, Grhahramani, Jaakkola, & Saul, 1999), Markov chain Monte Carlo (MCMC) sampling (Andrieu, de Freitas, Doucet, & Jordan, 2003), and stochastic variational inference (Hoffman, Blei, Wang, & Paisley, 2013). Furthermore, when the hyperparameters are optimized, the performance differences in terms of perplexity (Azzonpardi, Girolami, & van Risjbergen, 2003) among the inference algorithms diminish significantly (Asuncion, Welling, Smyth, & Teh, 2009). In this work, the collapsed Gibbs sampling (a special case of MCMC), armed with the hyperparameter optimization subprocedure (Minka, 2003), is utilized, since it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows a combination of estimates from several local maxima of the posterior distribution.

In the collapsed Gibbs sampling procedure, we need to calculate the posterior distribution, conditional distributions of the hidden random variables ($\vec{z}, \vec{\tilde{z}}$, and $\vec{x}$) given the observations and other hidden variables, $\Pr\left(z_{m,n} | \vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}\right)$ and $\Pr\left(x_{m,\tilde{n}}, \tilde{z}_{m,\tilde{n}} | \vec{\tilde{w}}, \vec{z}, \vec{y}, \vec{x}_{\neg(m,\tilde{n})}, \vec{\tilde{z}}_{\neg(m,\tilde{n})}, \vec{\mu}, \vec{\gamma}, \vec{\delta}\right)$, where $\vec{z}_{\neg(m,n)}$ represents the topic assignments for all tokens except $w_{m,n}$, and $\vec{x}_{\neg(m,\tilde{n})}$, $\vec{\tilde{z}}_{\neg(m,\tilde{n})}$ represents the super-topic and entity topic assignments for all entity mentions except $\tilde{w}_{m,\tilde{n}}$, respectively. After a simple derivation, the posterior distributions can be formally expressed as follows:

$$\Pr\left(z_{m,n} | \vec{w}, \vec{z}_{\neg(m,n)}, \vec{\alpha}, \vec{\beta}\right) \propto \frac{n_{z_{m,n}}^{(w_{m,n})} + \beta_{w_{m,n}} - 1}{\sum_{v=1}^{V}\left(n_{z_{m,n}}^{(v)} + \beta_v\right) - 1} \left(n_m^{(z_{m,n})} + \alpha_{z_{m,n}} - 1\right) \quad (1)$$

DOI: 10.1002/asi

$$\Pr\left(x_{m,\tilde{n}}, \tilde{z}_{m,\tilde{n}} \,|\, \vec{\tilde{w}}, \vec{z}, \vec{y}, \vec{x}_{\neg(m,\tilde{n})}, \vec{\tilde{z}}_{\neg(m,\tilde{n})}, \vec{\mu}, \vec{\gamma}, \vec{\delta}\right)$$
$$\propto \frac{n_m^{\left(x_{m,\tilde{n}}\right)}}{N_m} \frac{n_{\tilde{z}_{m,\tilde{n}}}^{\left(\tilde{w}_{m,\tilde{n}}\right)} + \delta_{\tilde{w}_{m,\tilde{n}}} - 1}{\sum_{\tilde{v}=1}^{\tilde{V}} \left(n_{\tilde{z}_{m,\tilde{n}}\tilde{v}} + \delta_{\tilde{v}}\right) - 1} \frac{n_{x_{m,\tilde{n}}}^{\left(y_{m,\tilde{n}}\right)} + \mu_{y_{m,\tilde{n}}} - 1}{\sum_{c=1}^{C} \left(n_{x_{m,\tilde{n}}}^{(c)} + \mu_c\right) - 1} \frac{n_{x_{m,\tilde{n}},y_{m,\tilde{n}}}^{\left(\tilde{z}_{m,\tilde{n}}\right)} + \gamma_{\tilde{z}_{m,\tilde{n}}} - 1}{\sum_{\tilde{k}=1}^{\tilde{K}} \left(n_{x_{m,\tilde{n}},y_{m,\tilde{n}}}^{\left(\tilde{k}\right)} + \gamma_{\tilde{k}}\right) - 1} \tag{2}$$

where $n_k^{(v)}$ is the number of tokens of word v that are assigned to topic k, $n_m^{(k)}$ represent the number of tokens in document m that are assigned to topic k, $n_k^{(c)}$ is the number of entity mentions with the class c surrounding the super-topic k, $n_{k,c}^{\left(\tilde{k}\right)}$ represents the number of entity mentions with super-topic k and class c assigned to entity topic $\tilde{k}$, and $n\tilde{k}^{(\tilde{v})}$ represents the number of mentions of entity $\tilde{v}$ assigned to entity topic $\tilde{k}$. Equation 1 is the same as that in the standard LDA model (Griffiths & Steyvers, 2004), and the first two terms in Equation 2 are shared with the CorrLDA2 model (Newman et al., 2006). Using the expectation of Dirichlet distribution, one can easily obtain the resulting model parameters as follows:

$$\varphi_{k,v} = \frac{n_k^{(v)} + \beta_v}{\sum_{v=1}^{V} \left(n_k^{(v)} + \beta_v\right)} \tag{3}$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{N_m + \sum_{k=1}^{K} \alpha_k} \tag{4}$$

$$\xi_{k,c} = \frac{n_k^{(c)} + \mu_c}{\sum_{c=1}^{C} \left(n_k^{(c)} + \mu_c\right)} \tag{5}$$

$$\psi_{k,c,\tilde{k}} = \frac{n_{k,c}^{\left(\tilde{k}\right)} + \gamma_{\tilde{k}}}{\sum_{\tilde{k}=1}^{\tilde{K}} \left(n_{k,c}^{\left(\tilde{k}\right)} + \gamma_{\tilde{k}}\right)} \tag{6}$$

$$\phi_{\tilde{k},\tilde{v}} = \frac{n_{\tilde{k}}^{(\tilde{v})} + \delta_{\tilde{v}}}{\sum_{\tilde{v}=1}^{\tilde{V}} \left(n_{\tilde{k}}^{(\tilde{v})} + \delta_{\tilde{v}}\right)} \tag{7}$$

In this work, the number of word topics $K^{(s)}$ and $K^{(t)}$ for scholarly articles and patents are fixed to 50, and the number of entity topics $\tilde{K}^{(s)}$ and $\tilde{K}^{(t)}$ are fixed to 20. In the light of the work by Wallach, Mimmo, and McCallum (2010), the asymmetric Dirichlet priors $\vec{\alpha}$ and symmetric ones β, μ, γ, and δ are utilized, and their initial values are set to $[0.5, 0.5, \cdots, 0.5]^{\mathrm{T}}$, 0.01, 0.1, 0.5, and 0.01, respectively. All the hyperparameters are optimized with the fixed-point iteration by the Dirichlet-multinomial/Pólya distribution directly (cf. eq. 55 in Minka, 2003). The collapsed Gibbs sampling is run for 2,000 iterations, including 500 for the burn-in period.

Word Tokens and Entity Mentions Clustering

Although both academic publications and patents belong to text media, there still exists some heterogeneity between them in terms of purpose, statement, quality, and so on. The purpose of an academic article is to communicate scientific findings to the relevant research community and general public. Patents are legal documents utilized to prevent other agents from commercializing the technical processes or devices they describe. In order to guarantee the quality, peer review is usually used to filter, improve, and curate scientific articles. The review on patents is driven by the legal requirements only. A major point is the absence of overlapping with existing patent documents and other publicly available materials. In other words, if you or someone else published an article about your process or device, you cannot patent it anymore. The other way seems nonproblematic if the technical processes were patented, although most journals do not want to republish existing literature.

The heterogeneity between the scholarly articles and patents enables many word tokens and entity mentions to appear in only one resource. Table 3 shows two examples of respective scholarly articles and patents. The tokens with gray color correspond to stopwords, and the bolded ones are shared by scholarly articles and patents. This means that whichever similarity calculation method is

TABLE 3. Two examples of respective scholarly article and patent.

PMID	Title after tokenization
9925120	**cholesterol** - **lowering** effects of **dietary** fiber: a meta - analysis
21776465	mechanisms underlying the **cholesterol** - **lowering** properties of soluble **dietary fiber** polysaccharides
Patent No.	Title after tokenization
EP1526857A1	**cholesterol** - reducing agent made of **dietary fiber** and **cholesterol** - reducing substances
US6180660	**cholesterol** - **lowering** therapy

DOI: 10.1002/asi

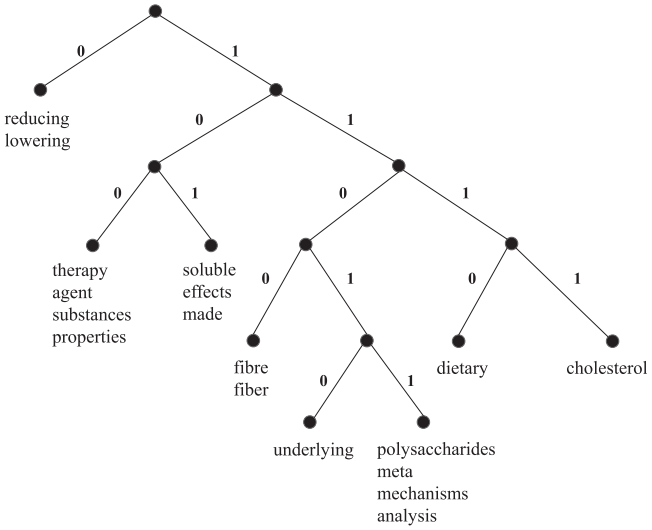


FIG. 4. A binary tree with eight clusters for the word tokens in titles of scholarly articles and patents from Table 3.

utilized in the next step, many word tokens and entity mentions (that is, those that only appear in one resource) will be discarded directly, thus having no contribution for similarity calculation. In our opinion, this may be one of the main reasons for the unsatisfactory performance of Shibata et al. (2010), Shibata et al., 2011), and Xu et al. (2012)): many topic pairs with top similarity values are not regarded to be semantically similar by experts. In more detail, only two pairs of topics were regarded to be semantically similar. Therefore, in order to reduce the negative impact on topic similarity calculation, word tokens and entity mentions should be clustered before it. Here, by the negative impact, we mean that only shared word tokens and entity mentions are involved in topic similarity calculation.

Many unsupervised word clustering methods are proposed in the literature, such as the Brown clustering method (Brown et al., 1992; Liang, 2005), word embedding (Collobert & Weston, 2008; Mnih & Andriy, 2009), spectral feature alignment (Pan, Ni, Sun, Yang, & Chen, 2010), and so on. Here the Brown clustering method is used. A binary tree can be obtained by running the Brown clustering method, where each leaf node defines a cluster of words or entities. The root node represents a single cluster containing all tokens or entities, and interior nodes represent intermediate size clusters containing all of the tokens or entities that they dominate. In other words, nodes lower in the binary tree correspond to smaller clusters, while higher nodes correspond to larger clusters. According to the Huffman coding method (Huffman, 1952), a particular token or entity can be assigned a binary string by following the traversal path from the root to its leaf, assigning a 0 for each left branch, and a 1 for each right branch.

Figure 4 shows the binary tree with eight clusters for the word tokens in titles of scholarly articles and patents from Table 3. Intuitively, the Brown clustering method will merge the tokens or entities with similar contexts into the same cluster, such as *reducing* and *lowering*, and *fiber* and

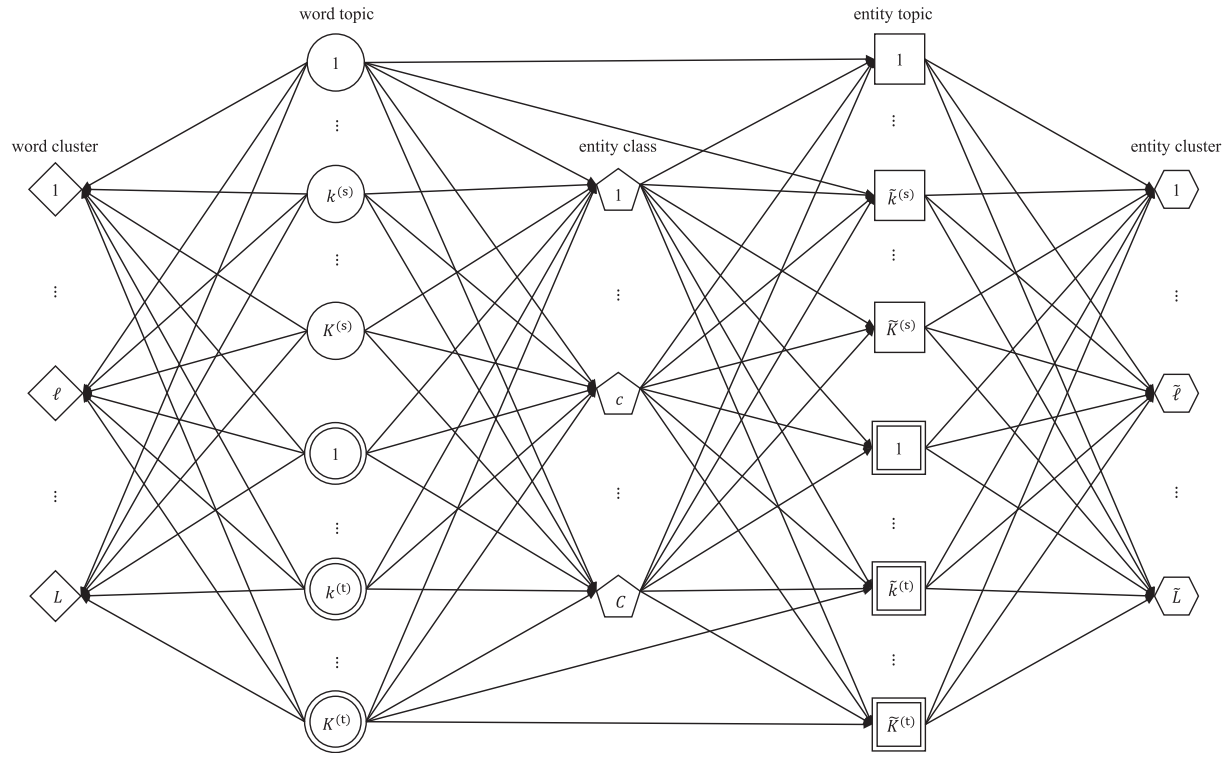
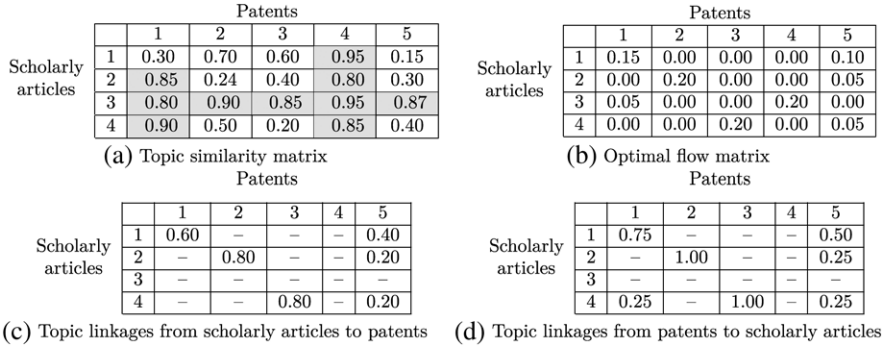


FIG. 5. Network structure consisting of conditional distributions between word clusters, word topics, entity classes, entity topics, and entity clusters.

DOI: 10.1002/asi



Patents

Scholarly articles	1	2	3	4	5
1	0.30	0.70	0.60	0.95	0.15
2	0.85	0.24	0.40	0.80	0.30
3	0.80	0.90	0.85	0.95	0.87
4	0.90	0.50	0.20	0.85	0.40

(a) Topic similarity matrix

Patents

Scholarly articles	1	2	3	4	5
1	0.15	0.00	0.00	0.00	0.10
2	0.00	0.20	0.00	0.00	0.05
3	0.05	0.00	0.00	0.20	0.00
4	0.00	0.00	0.20	0.00	0.05

(b) Optimal flow matrix

Patents

Scholarly articles	1	2	3	4	5
1	0.60	–	–	–	0.40
2	–	0.80	–	–	0.20
3	–	–	–	–	–
4	–	–	0.80	–	0.20

(c) Topic linkages from scholarly articles to patents

Patents

Scholarly articles	1	2	3	4	5
1	0.75	–	–	–	0.50
2	–	1.00	–	–	0.25
3	–	–	–	–	–
4	0.25	–	1.00	–	0.25

(d) Topic linkages from patents to scholarly articles

FIG. 6. An illustration example about topic linkages. (**a**) Linkages in Shibata et al. (2010). (**b**) Our linkages.

fiber. Following the main idea of the Brown clustering method, the more similar the prefix of the token's or entity's Huffman codes, the more similar are the tokens or entities. For instance, the token *cholesterol* (1111) in Figure 4 is more similar than the token *dietary* (1110) with the token *polysaccharides* (11011). After this step, high-dimensional word or entity space will be considerably reduced to low-dimensional cluster space. In this study, the implementation of the Brown clustering method by Liang[1] is adopted. The word tokens and entity mentions are clustered separately to L = 1,000 groups $\{\mathcal{P}_1,\cdots,\mathcal{P}_L\}$ and $\tilde{L}$ = 500 groups $\{\tilde{\mathcal{P}}_1,\cdots,\tilde{\mathcal{P}}_{\tilde{L}}\}$.

Topic Similarity

From Figure 3, one can see that the definition of a topic in the CCorrLDA2 model is different from that in the standard LDA model. The CCorrLDA2 model makes a clear distinction between a word topic and an entity topic. The word topic in the CCorrLDA2 model is the same as the topic in the standard LDA model. More specifically, each word topic is surrounded by many entities with different classes, and then the entities with some class are then further divided into several entity topics. Put another way, the word topic serves as the super topic of the resulting entity topics. Figure 5 illustrates the network structures between word clusters (rhombus nodes), word topics (single circle nodes for the scientific publications and double ones for the patents), entity classes (pentagon nodes), entity topics (single square nodes for the scientific publications and double ones for the patents), and entity clusters (hexagon nodes), in which the arrows mean the resulting conditional distributions. This work does not draw all the arrows from word topics to entity topics for clarity. From Figure 5, it is not difficult to see that word clusters, entity classes, and entity clusters can act as *pivots* between the super topic $k^{(s)}$ from the scientific publications and the super topic $k^{(t)}$ from the patents. Hence, the similarity can be defined formally as follows:

$$\begin{aligned}\mathrm{sim}\big(k^{(s)},k^{(t)}\big) &= (1-\lambda-\rho)\mathrm{sim}\Big(\Pr\big([\mathcal{P}_\ell]_{\ell=1}^{L}|k^{(s)}\big),\Pr\big([\mathcal{P}_\ell]_{\ell=1}^{L}|k^{(t)}\big)\Big)\\ &+\lambda\mathrm{sim}\Big(\Pr\Big([c]_{c=1}^{C}|k^{(s)}\Big),\Pr\Big([c]_{c=1}^{C}|k^{(t)}\Big)\Big)\\ &+\rho\mathrm{sim}\Big(\Pr\Big([\tilde{\mathcal{P}}_{\tilde{\ell}}]_{\tilde{\ell}=1\tilde{L}}|k^{(s)}\Big),\Pr\big([\tilde{\mathcal{P}}_{\tilde{\ell}}]\tilde{\ell}\tilde{L}|k^{(t)}\big)\Big)\end{aligned} \tag{8}$$

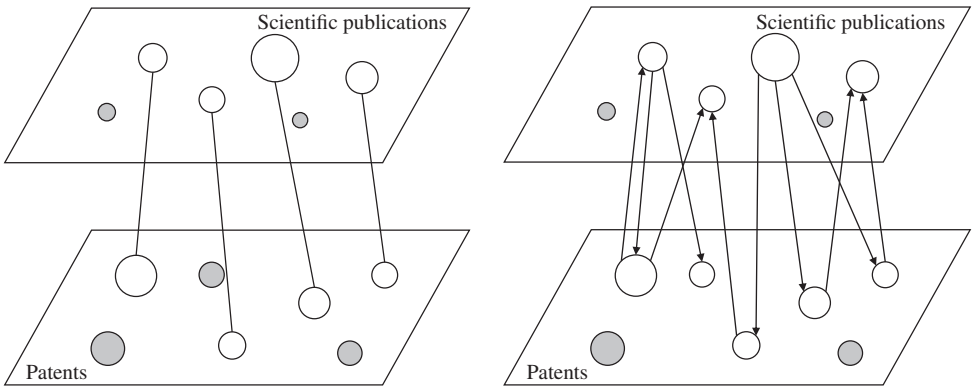
Here, $\Pr(\mathcal{P}_\ell|\cdot)=\sum_{v\in\mathcal{P}_\ell}\varphi_{\cdot,v}$, $\Pr(c|\cdot)=\xi_{\cdot,c}$, and $\Pr\big(\tilde{\mathcal{P}}_{\tilde{\ell}}|\cdot\big)=\sum_{\tilde{v}\in\tilde{P}_{\tilde{\ell}}}\Pr(\tilde{v}|\cdot)$ with $\Pr(\tilde{v}|\cdot)=\sum_c\sum_{\tilde{k}}\Pr(c|\cdot)\ \Pr\big(\tilde{k}|\cdot,c\big)\ \Pr\big(\tilde{v}|\tilde{k}\big)=\sum_c\sum_{\tilde{k}}\xi_{\cdot,c}\psi_{\cdot,c,\tilde{k}}\phi_{\tilde{k},\tilde{v}}$. The three terms of the right-hand side in Equation 8 correspond to the similarity between word topics, the similarity between entity classes, and the similarity between entity topics, respectively. The weights λ and ρ can regulate the degree of importance of each term. In this study, λ and ρ are set to $\frac{1}{3}$. That is, the same important degree for these three terms is assumed. It is worth mentioning that the similarity (Equation 8) is transformed into the distance by: $\mathrm{d}(k^{(s)}, k^{(t)}) = \max(\mathrm{sim}(k^{(s)}, k^{(s)}), \mathrm{sim}(k^{(t)}, k^{(t)})) - \mathrm{sim}(k^{(s)}, k^{(t)})$ (Chen, Ma, & Zhang, 2009) in this work, but the similarity (Equation 8) (dissimilarity in this case, to be more precise) remains unchanged if the JS or symmetrized KL divergence is involved.

Topic Linkages

Similar to Xu et al. (2012), if one can see topics in the scientific publications and the patents as sources and sinks, respectively, or vice versa, and topic similarities as distances between sources and sinks, the topic linkages construction problem can be transformed into the well-known *optimal transportation problem* (Hillier & Lieberman, 1995; Rachev & Ruschendorf, 1998). The question answered by the optimal transportation problem is: what is the cheapest way to move a set of masses from sources to sinks? Here, the cost is defined as the total *mass* × *distance* moved. In order to facilitate understanding, one can think of the sources as factories and the sinks as warehouses to make the problem concrete. In this work, the sources are assumed to ship exactly as much mass as the sinks are expecting.

[1] https://github.com/percyliang/brown-cluster

DOI: 10.1002/asi



(a) Linkages in Shibata et al. (2010) (b) Our linkages

FIG. 7. The illustration of topic linkages between scientific publications and patents. (**a**) Scientific publications. (**b**) Patents.

By running the CCorrLDA2 model, the two topic sets $\mathcal{T}^{(s)} = \left[k^{(s)}\right]_{k=1}^{K^{(s)}}$ and $\mathcal{T}^{(t)} = \left[k^{(t)}\right]_{k=1}^{K^{(t)}}$ can be obtained from the scientific publications and the patents with respective associated nonnegative weights $p_{k^{(s)}} = \Pr\left(k^{(s)}\right)$ and $q_{k^{(t)}} = \Pr\left(k^{(t)}\right)$. In this study, Pr($k$) is set to $\frac{\sum_v n_k^{(v)}}{\sum_k \sum_v n_k^{(v)}}$, which represents the importance of the corresponding topic. The optimal transportation distance between $\mathcal{T}^{(s)}$ and $\mathcal{T}^{(t)}$ is formally defined as $d\left(\mathcal{T}^{(s)}, \mathcal{T}^{(t)}\right) = \sum_{k^{(s)}=1}^{K^{(s)}} \sum_{k^{(t)}=1}^{K^{(t)}} \left[f^*_{k^{(s)},k^{(t)}} \times \mathrm{d}\left(k^{(s)}, k^{(t)}\right)\right]$, where the optimal flow $F^* = \left[f^*_{k^{(s)},k^{(t)}}\right]_{K^{(s)} \times K^{(t)}}$ between $\mathcal{T}^{(s)}$ and $\mathcal{T}^{(t)}$ is the solution of the following linear programming, which guarantees an optimal solution.

$$\min_{F \in \mathbb{R}^{K^{(s)} \times K^{(t)}}} d\left(\mathcal{T}^{(s)}, \mathcal{T}^{(t)}\right) \tag{9}$$

$$\text{s.t. } f_{k^{(s)},k^{(t)}} > 0, 1 \le k^{(s)} \le K^{(s)}, 1 \le k^{(s)} \le K^{(t)} \tag{10}$$

$$\sum_{k^{(t)}=1}^{K^{(t)}} f_{k^{(s)},k^{(t)}} = p_{k^{(s)}}, 1 \le k^{(s)} \le K^{(s)} \tag{11}$$

$$\sum_{k^{(s)}=1}^{K^{(s)}} f_{k^{(s)},k^{(t)}} = q_{k^{(t)}}, 1 \le k^{(t)} \le K^{(t)} \tag{12}$$

$$\sum_{k^{(s)}=1}^{K^{(s)}} \sum_{k^{(t)}=1}^{K^{(t)}} f_{k^{(s)},k^{(t)}} = 1 \tag{13}$$

Once the optimal flow matrix is known, it is very easy to construct topic linkages from nonzero entries in the optimal flow matrix. Let's take Figure 6 as an example. The scientific publications and patents are posited as the mixtures of $K^{(s)} = 4$ and $K^{(t)} = 5$ topics, respectively. The distance matrix between these topics is given in Figure 6a. The grayed cells (greater than or equal to 50% percentile of the distances) mean the resulting topics are so different that it is impossible for them to link to each other. The associated nonnegative weights $p_{k^{(s)}}$ and $q_{k^{(t)}}$ are assumed to be $\frac{1}{K^{(s)}} = 0.25$ and $\frac{1}{K^{(t)}} = 0.2$, respectively. The resulting optimal flow matrix F^* is shown in Figure 6b, from which topic 1 from the scientific publications links to topic 1 and topic 5 from the patents with the respective linkage strength $\frac{f_{1,1}}{p_1} = 0.6$ and $\frac{f_{1,5}}{p_1} = 0.4$ (Figure 6c), and topic 5 from the patents links to topic 1, topic 2, and topic 4 from the scientific publications with the respective linkage strength $\frac{f_{1,1}}{q_5} = 0.50$, $\frac{f_{2,1}}{q_5} = 0.25$ and $\frac{f_{4,1}}{q_5} = 0.25$ (Figure 6d). Topic 3 from the scientific publications and topic 4 from the patents can be seen as the resource-specific topics.

In summary, each topic from one resource can link to multiple topics from the other resource, and our topic linkages are asymmetric, which are different from Shibata et al. (2010). More specifically, Figure 7 illustrates the main differences between the linkages in Shibata et al. (2010) and our linkages, in which gray topics in each layer mean the resource-specific topics. In Shibata et al. (2010), a statement of "*a* links to *b*" and a statement of "*b* links to *a*" should be of equal value. However, our topic linkages are more similar to hyperlinks from a hypertext file or document to another location or file: "*a* hyperlinks to *b*" does not mean "*b* hyperlinks to *a*." In our opinion, these bidirectional linkages are more in line with the actual situation. Because if topic *a* connects to topic *b*, it is usually assumed that topic *a* is semantically related to topic *b*.

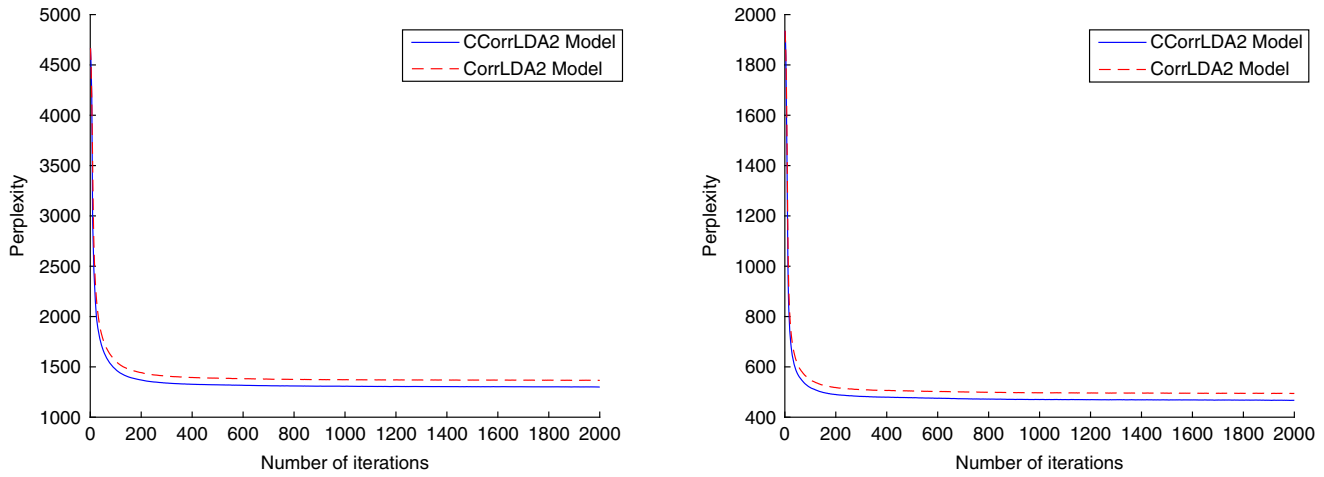
In practice, whichever method is used to extract the respective topics from two different resources, it is often very difficult to make sure that each topic from one resource is semantically identical to some topic from another resource. In most cases, a topic from one resource is semantically covered by several topics from another resource, and vice versa. Let's take Figure 6 as an example. Topics 1 and 5 from the patents express 60% and 40% meaning of topic 1 from the scientific publications, respectively. Topics 1, 2, and 4 from the scientific publications

DOI: 10.1002/asi

TABLE 4. Statistics for the data set.

	Scientific publications		Patents				
	Word tokens	Unique words	Word tokens	Unique words	Intersection	Union	Rate (%)
	1,128,450	41,221	673,932	24,848	14,225	51,844	27.43
	Scientific Publications		Patents				
Entity Class	Entity mentions	Unique entities	Entity mentions	Unique entities	Intersection	Union	Rate (%)
1	13,118	1,781	1,042	343	120	2,004	5.99
2	11,935	3,155	23,919	8,458	494	11,119	4.44
3	12,028	1,821	4,359	1,002	161	2,662	6.05
4	1,824	574	224	119	11	682	1.61
5	589	489	281	224	0	713	0.00
6	19,138	6,152	18,764	5,936	658	11,430	5.76
7	25,610	3,764	17,096	3,163	1,007	5,920	17.01
∑	84,242	17,627	65,685	18,444	2,747	33,324	8.24

1: ABBREVIATION; 2: FAMILY; 3: FORMULA; 4: IDENTIFIER; 5: MULTIPLE; 6: SYSTEMATIC; 7: TRIVIAL.



(a) Scientific publications (b) Patents

FIG. 8. The perplexities of the CCorrLDA2 and CorrLDA2 models on (**a**) scientific publications and (**b**) patents. (a) A topic example in the scientific publications. (b) A topic example in the patents. [Color figure can be viewed at wileyonlinelibrary.com]

express 50%, 25%, and 25% meaning of topic 5 from the patents, respectively. Therefore, similar to hyperlinks for the surfers, our linkages will directly benefit topic navigation (Zhang & Zhang, 2008) when the scholars are exploring the heterogeneous information resources.

Experimental Results & Discussions

Data Set

In order to build topic linkages between the scientific publications and the patents, two corpora, the CHEMDNER (chemical compound and drug named entity recognition) corpus (Krallinger et al., 2015) and the CHEMDNER-patents corpus (Pérez-Pérez et al., 2017), are used in this study. They were originally used for CHEMDNER challenges in BioCreative IV and V, respectively. In total, CHEMDNER corpus includes 10,000 scholarly articles, and CHEMDNER-patents corpus consists of 14,000 patents. The seven categories (ABBREVIATION, FAMILY, FORMULA, IDENTIFIER, MULTIPLE, SYSTEMATIC, and TRIVIAL) of entity mentions in the title and abstract of each document are annotated by domain experts. These two corpora are further divided into three sets: *train*, *development,* and *test*, but entity annotations in the test set of the CHEMDNER-patents corpus are not provided publicly. Therefore, only the train and development sets for this corpus are utilized here. See Krallinger et al. (2015) and Pérez-Pérez et al. (2017) for more details of selection criteria of the scientific publications and the patents, entity annotations, and splitting into train, development, and test sets.

This study conducts the same preprocessing steps (sentence detection and tokenization) as Xu, An, Zhu, Zhang, and Zhang (2015). Please check Xu et al. (2015) for more details. As for stopword filtering, an English stopword list from NLTK (Natural Language Toolkit) [2] is used here, but

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

DOI: 10.1002/asi

Word Topic 11	
Word	Prob. (%)
cancer	8.19
cells	4.73
cell	4.52
tumor	3.72
NUMBER	2.00
breast	1.86
human	1.78
growth	1.62
lines	1.22
prostate	1.00

Entity Class (%)				
TRIVIAL	ABBREVIATION	FAMILY	IDENTIFIER	FORMULA
70.01	12.88	10.67	5.87	0.57

100.00 100.00 99.99

Entity Topic 2	
Entity	Prob. (%)
tyrosine	4.30
cisplatin	3.47
serine	2.74
rapamycin	2.53
cysteine	2.50
lysine	2.09
paclitaxel	1.70
doxorubicin	1.30
glucose	1.28
tamoxifen	1.25

Entity Topic 6	
Entity	Prob. (%)
gsh	3.20
bpa	2.82
atp	2.61
dox	2.23
tcdd	2.15
camp	1.79
mda	1.69
peg	1,59
nadph	1.46
mtt	1.20

Entity Topic 18	
Entity	Prob. (%)
estrogen	6.24
androgen	5.31
amino acid	4.13
amino acids	3.00
nucleotide	2.79
steroid	2.44
sugar	2.32
polyphenols	2.11
phenolic	1.79
fattty acids	1.64

(a) A topic example in the scientific publications

Word Topic 43	
Word	Prob. (%)
cancer	17.63
tumor	6.43
cell	5.08
cells	4.77
lung	2.16
breast	2.03
anti	1.96
human	1.87
growth	1.75
treating	1.69

Entity Class (%)					
FAMILY	SYSTEMATIC	TRIVIAL	ABBREVIATION	MULTIPLE	FORMULA
40.28	28.01	19.98	8.79	1.95	0.98

99.98 99.96 99.99

Entity Topic 14	
Entity	Prob. (%)
cephalosporin	1.95
tetracycline	1.32
nucleoside	1.07
aminoglycoside	0.94
azole	0.82
macrolide	0.76
isoflavone	0.69
glucoside	0.69
diterpenoid	0.69
pyridine	0.63

Entity Topic 4	
Entity	Prob. (%)
calcium	2.58
amine	2.15
potassium	1.95
sodium	1.80
phosphate	0.94
acetylcholine	0.94
platinum	0.86
acyl	0.66
5-ht	0.66
oxygen	0.62

Entity Topic 16	
Entity	Prob. (%)
curcumin	2.51
chukrasone b	2.23
sarcaboside b	2.18
sarcaboside a	1.94
chukrasone a	1.85
salvianolic acid a	1.71
eryngiolide a	1.56
houttuynoid b	1.56
incarviatone a	1.28
houttuynoid a	1.28

(b) A topic example in the patents

FIG. 9. Two topic examples discovered by the CCorrLDA2 model from (**a**) scientific publications and (**b**) patents. (a) Scientific publications (b) Patents.

DOI: 10.1002/asi

expanded with some punctuation (such as @, %, and so on). All numbers in the scholarly articles and patents are replaced with a special word *NUMBER*. Since the entity mentioned in CHEMDNER and CHEMDNER-patents corpora have been annotated in advance, *entity mentions recognition* in preprocessing of Figure 2 is skipped directly. Table 4 shows the statistics for the used data sets. The upper half part of Table 4 is for word tokens and unique words, and the lower half par for entity mentions and unique entities. Due to different statements in the scientific publications and patents, the intersections of unique words (27.43%) and unique entities (8.24%) are very low. It is noted that the words and entities are assumed to be uncorrelated in the CCorrLDA2 and other similar models, although one can incorporate word and entity correlation knowledge into our model by following the idea in the Markov Random Field (MRF) regularized LDA model (Xie, Yang, & Xing, 2015). Therefore, it is very necessary to cluster word tokens and entity mentions to alleviate the problem from synonyms before topic similarity calculation.

Topic Extraction and Examples

The perplexity, originally used in language modeling (Azzonpardi et al., 2003), is a standard measure for estimating the performance of a probabilistic model. This measure is defined as the exponential of the negative normalized predictive likelihood under the model, and a lower value indicates a better modeling performance. Figure 8 depicts the perplexity values of the CCorrLDA2 and CorrLDA2 models. Note that the CorrLDA2 model does not consider the entity class, this consideration is taken into post-hoc (that is, not during learning). Both CCorrLDA2 and CorrLDA2 models converges after around 200 iterations, but the CCorrLDA2 model performs better on topic extraction than the CorrLDA2 model. Here, the convergence is defined as the situation where the resulting perplexity does not decrease any more over the number of iterations.

Figure 9 illustrates two topic examples learned by the CCorrLDA2 model from the scientific publications and patents, respectively. These topics are extracted from a single sample at the 2,000th iteration of the collapsed Gibbs sampler. Each word topic is illustrated with (i) the top 10 most relevant words for the word topic; (ii) the most likely classes of the entities surrounding the resulting word topic; (iii) the most relevant entity topic for the each main entity class; and (iv) the top 10 most relevant entities for each entity topic. It is not difficult to see that both of these two topics are related to *cancer*. This indicates that there are indeed some hidden topics, which can be interlinked, from the scientific publications and the patents. Nonetheless, there are still some differences between these two topics. For instance, the entities surrounding the word topic 11 in the scientific publications mainly belong to the classes TRIVIAL (70.01%), ABBREVIATION (12.88%), and FAMILY (10.67%), and the entities surrounding the word topic 43 in the patents to the classes FAMILY (40.28%), SYSTEMATIC (28.01%), TRIVIAL (8.79%), and ABBREVIATION (7.78%). This makes clear that the statements in the patent documents are more formal than those in the article documents. In addition, the entities surrounding the specific word topic mostly belong to only one entity topic (cf. the weights from entity classes to entity topics in Figure 9).

Topic Similarity and Linkages

To evaluate the performance of topic similarity measures reviewed in the *Topic Similarity Measurement* subsection, above, the procedure proposed by Kim and Oh (2011) was utilized here. Specifically, starting from a set of

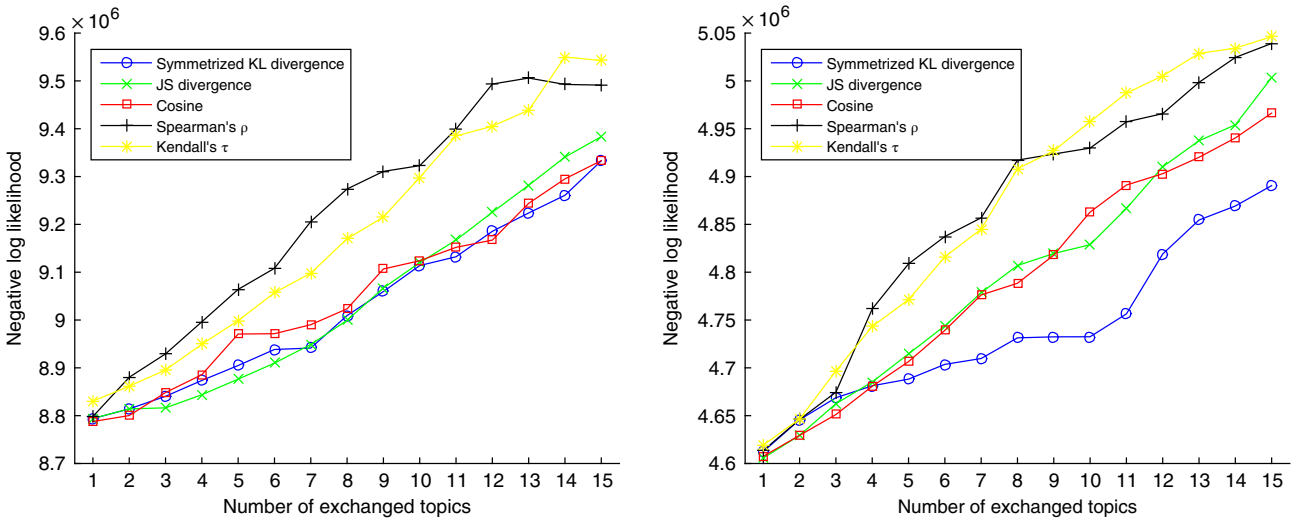
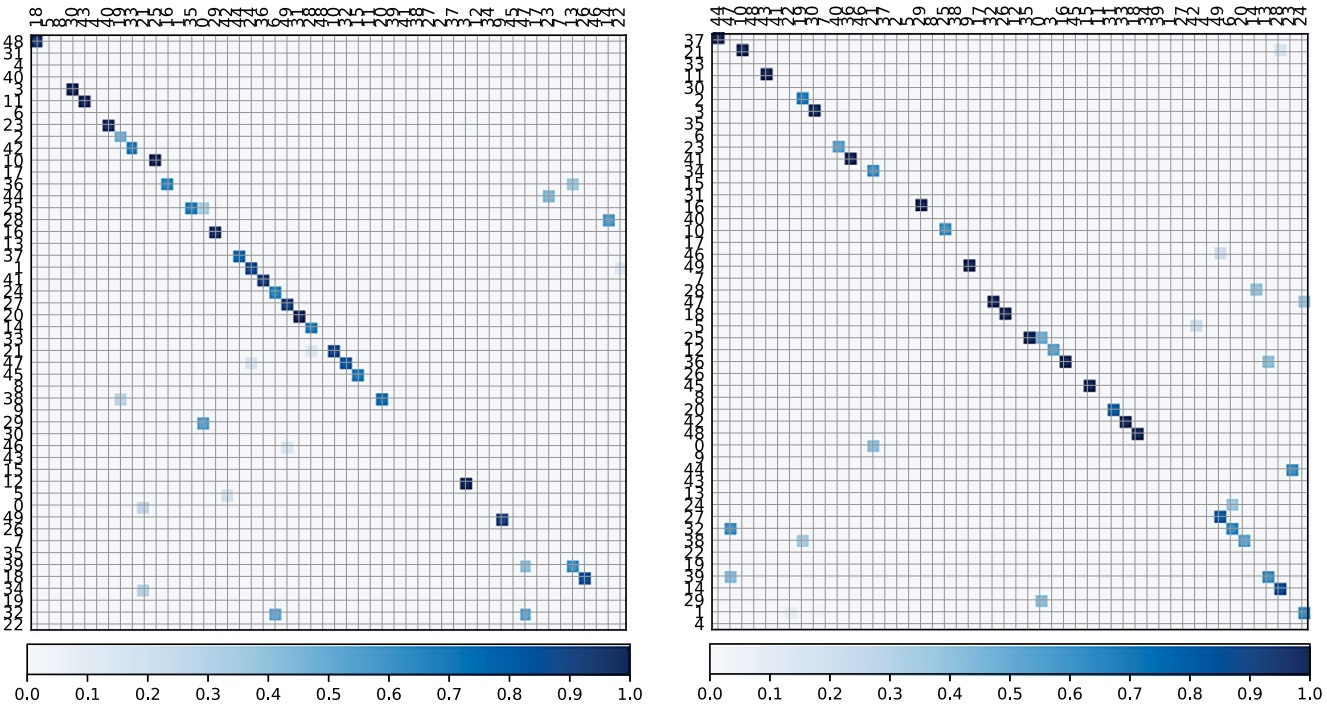


FIG. 10. The comparison of negative log likelihood for five similarity metrics. (**a**) From scientific publications to patents (**b**) From patents to scientific publications. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi



(a) From scientific publications to patents

(b) From patents to scientific publications

FIG. 11. The linkage strength map between scientific publications and patents. [Color figure can be viewed at wileyonlinelibrary.com]

topics extracted for scientific publications/patents, we exchange top similar topics from the patents/academic articles according to the above metrics to form the resulting modified sets of topics. Then the negative log likelihoods of the respective corpus, which measures how well the model explains the corpus, are calculated with the modified sets of topics, and are given in Figure 10. A better similarity metric gives a lower negative log likelihood.

As Figure 10 shows, symmetrized KL divergence produces the lowest negative log likelihood scores on the patent corpus, especially when the number of exchanged topics is greater than 4, but has similar scores with JS divergences and cosine similarity on the scientific articles. Spearman's ρ and Kendall's τ give consistently the worst performance. Another way to say this is that symmetrized KL divergence performs the best among the five metrics, which is different from the observations by Kim and Oh (2011), but is in line with the observations by Xu et al. (2012). Hence, symmetrized KL divergence is embedded in the topic similarity calculation (cf. Equation 8).

Figure 11 shows the linkage strength map between the scientific publications and the patents (filtered by ≥90% percentile of the distances). The horizontal and vertical axes correspond to topics from the scientific publications and patents, respectively. The darker square marks mean the more linkage strength. From Figure 11, one can see that whichever linkage strength from the academic articles to the patents or from the patents to academic articles, a sparse diagonal structure occurs by rearranging the order of topics. This indicates that most topics from one resource are mainly linked to only

TABLE 5. Comparisons of topic linkages.

Model	Clustering	5	4	3	2	1	Σ
CCorrLDA2	√	18 (45.00%)	9 (22.50%)	7 (17.50%)	5 (12.50%)	1 (2.50%)	40
CCorrLDA2	×	6 (37.50%)	3 (18.75%)	1 (6.25%)	5 (31.25%)	2 (12.50%)	16
CorrLDA2	√	15 (35.71%)	8 (19.05%)	9 (21.43%)	7 (16.67%)	3 (7.14%)	42
CorrLDA2	×	5 (29.41%)	3 (17.65%)	2 (17.76%)	4 (23.53%)	3 (17.65%)	17
LDA	√	14 (30.43%)	10 (21.74%)	9 (19.57%)	8 (17.39%)	5 (10.87%)	46
LDA	×	5 (26.32%)	2 (10.53%)	4 (21.05%)	4 (21.05%)	4 (21.05%)	19

DOI: 10.1002/asi

one topic from the other resource, such as topic 11 in the scholarly articles and topic 43 in the patents (see Figure 9 for more detail). Of course, it is possible for one topic from one resource to link to multiple topics from the other resource. For instance, topic 38 in the scientific publications links to topics 19 and 20 in the patents, and topic 19 in the patents links to topics 2 and 38 in the scientific publications. If there are no square marks in some of the rows in Figure 11a or columns in Figure 11b, the topics corresponding to resulting rows or columns are resource-specific, such as topic 7 in the scientific publications and topic 2 in the patents.

In order to evaluate quantitatively the performance of our approach, six topic linkage sets are obtained with different settings. Please refer to the first two columns in Table 5 for more detail on the setting, where *clustering* means whether the word tokens and entity mentions are clustered or not before the topic similarity calculation. As mentioned in the *Topic Extraction* subsection, above, the standard LDA model is a special case of our CCorrLDA2 model when there are no entities mentioned in the corpus at all. Therefore, the setting in the last row of Table 5 is equivalent to that in Xu et al. (2012). In addition, the CorrLDA2 model does not take the entity classes into account and the LDA model does not consider the entities and resulting categories during learning; these considerations are taken into post-hoc. Three participants, who are all graduate students with a background in biochemistry, were recruited from our university. All topic linkages were shuffled randomly, examined manually one by one, and then rated with a 1 to 5 scoring value (1, poor; 2, fair; 3, average; 4, good; and 5, excellent) by the three participants. The participants did not know which group each topic linkage belongs to in advance. After all topic linkages were rated completely by the three participants, scores for each topic linkage were averaged and rounded to the nearest integer. The results are summarized in Table 5.

From Table 5, several phenomena can be observed as follows. (i) When the word tokens and entity mentions clustering is disabled, the number of built topic linkages is much less than that from the resulting clustering-enabled counterpart with whichever model. That is to say, many topics, which should be linked together, are left unlinked due to different statements from the scholarly articles and patents. (ii) When the clustering is enabled, the percentages of topic linkages with the rate ≥3 are 85.00%, 76.19%, and 71.74% for CCorrLDA2, CorrLDA2, and LDA models, respectively. But when the clustering is disabled, the corresponding percentages (about 60%) are very close. This indicates that it is very important to cluster word tokens and entity mentions ahead for the topic linkage construction. (iii) Under the same conditions, our CCorrLDA2 model outperforms the other models, and the standard LDA model is a distant third among the three. This makes it clear that entity mentions, especially resulting categories, are vitally important for building topic linkages between the scientific publications and patents.

Conclusion

Scientific publications and patents are usually considered good barometers of basic research and technical development, respectively. In order to understand the relationships between science and technology very well, this work devotes to building topic linkages between the scholarly articles and patents. Previous studies on linkages mainly focused on the analysis of NPRs on the front page of patents or the resulting citation-link networks, but with unsatisfactory performance. In the meanwhile, abundant entities are usually mentioned in the academic articles and patents, which further complicates topic linkages.

In order to deal with this situation, a novel statistical entity-topic model, the CCorrLDA2 model, is proposed to discover the hidden topics from the scientific publications and patents, and the collapsed Gibbs sampling algorithm is utilized for inferring the model parameters. The idea in the CCorrLDA2 model is also applicable to other similar models. Furthermore, the heterogeneity between the scholarly articles and patents enables many word tokens and entity mentions to appear in only one resource, which will result in a negative impact on topic similarity calculation. Therefore, the Brown clustering method is used to group word tokens and entity mentions before calculating the topic similarity, and then the topic linkages construction problem is transformed into the well-known optimal transportation problem.

In fact, our topic linkages are more similar to hyperlinks. That is to say, each topic from one resource can link to multiple topics from the other resource, and our topic linkages are asymmetric. Extensive experimental results indicate that our approach is feasible to build topic linkages with more superior performance than the compared ones. This investigation will benefit the topic navigation between different information resources (Zhang & Zhang, 2008), technological fronts detection (Shibata et al., 2011), business opportunity identification (Lee & Lee, 2017), and so on. A case study will be described in another article.

Although it is very natural to generalize the nonjoint method between two information resources to build topic linkages between multiple resources, a joint method or model still remains open. Another possible direction is to combine our lexical-based approach with a citation-based one (Shibata et al., 2010, 2011; Xu et al., 2018), which may further improve the performance. In addition, due to no benchmark data sets publicly available with known topic linkages until now, it is not trivial to compare very fairly two specific solutions.

Acknowledgments

This research received financial support from the Social Science Foundation of Beijing Municipality under grant number 17GLB074; National Science Foundation of China under grant number 71403255; Science and Technology Project of Guangdong Province under grant number 2017A030303065;

and Natural Science Foundation of Beijing Municipality under grant number 9174029. We thank the anonymous reviewers for valuable comments.

References

Albert, T. (2016). Measuring technology maturity: Operationalizing information from patents, scientific publications and the web. Wiesbaden, Germany: Springer Gabler.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M.I. (2003). An introduction to MCMC for machine learning. Machine Learning, 50(1–2), 5–43.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y.W. (2009). On smoothing and inference for topic models. In Proceedings of the 25th International Conference on Uncertainty in Artificial Intelligence (pp. 27–34). Arlington, VA: AUAI Press.

Azzonpardi, L., Girolami, M., & van Risjbergen, K. (2003). Investigating the relationship between language model perplexity and IR precision-recall measures. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 369–370). New York: ACM.

Bassecouolard, E., & Zitt, M. (2004). Patents and publications: The lexical connection. In H.F. Moed, W. Glänzel, & U. Schoch (Eds.), Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of s&t systems (pp. 665–694). Dordrecht, The Netherlands: Springer.

Bhattacharya, S., Kretschmer, H., & Meyer, M. (2003). Characterizing intellectual spaces between science and technology. Scientometricis, 58(2), 369–390.

Blei, D.M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Blei, D.M. & Jordan, M.I. (2003). Modeling annotated data. In Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 127–134). New York: ACM.

Blei, D.M. & Lafferty, J.D. (2006). Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning (pp. 113–120). New York: ACM.

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

Boyack, K.W., Newman, D., Duhon, R.J., Klavans, R., Patek, M., Biberstine, J.R., & Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuraries of nine text-based similarity approaches. PLoS One, 6(3), e18029.

Brooks, H. (1994). The relationship between science and technology. Research Policy, 23(5), 477–486.

Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., & Lai, J.C. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18(4), 467–479.

Calero-Medina, C., & Noyons, E.C.M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics, 2(4), 272–279.

Callaert, J., Looy, B.V., Verbeek, A., Debackere, K., & Thijs, B. (2006). Traces of prior art: An analysis of nonpatent references found in patent documents. Scientometrics, 69(1), 3–20.

Carpenter, M.P., & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. World Patent Information, 5(3), 180–185.

Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75.

Chen, S., Ma, B., & Zhang, K. (2009). On the similarity metric and the distance metric. Theoretical Computer Science, 410(24–25), 2365–2376.

Cohn, D., & Hofmann, T. (2001). The missing link - a probabilistic model of document content and hypertext connectivity. In In advances in neural information processing systems 13. Cambridge, MA: MIT Press.

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning (pp. 160–167). New York: ACM

Dubarić, E., Giannoccaro, D., Bengtsson, R., & Ackermann, T. (2011). Patent data as indicators of wind power technology development. World Patent Information, 33(2), 144–149.

Glänzel, W., & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. Scientometrics, 58(2), 425–428.

Griffiths, T.L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl. 5228–5235).

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, C.L. (2009). Detecting topic evolution in scientific literature: How can citations help? In Proceedings of the 18th ACM International Conference on Information and Knowledge Management (pp. 957–966). New York: ACM

Heinrich, G. (2009). Parameter estimation for text analysis technical report version 2.9. Darmstadt, Germany: vsonix GmbH and University of Leipzig.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. Research Evaluation, 13(2), 76–86.

Hillier, F., & Lieberman, G.J. (Eds.). (1995). Introduction to mathematical programming. New York: McGraw-Hill.

Hoffman, M.D., Blei, D.M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. Journal of Machine Learning Research, 14, 1303–1347.

Huffman, D.A. (1952). A method for the construction of minimum-redundancy codes. Proceedings of the I.R.E., 40(9), 1098–1101.

Jain, A.K., & Dubes, R.C. (1988). Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall.

Jibu, M. (2011). An analysis of the achievements of JST operations through scientific patenting: Linkage between patents and scientific papers. In Proceedings of the Conference on Science and Innovation Policy (pp. 1–7). Washington, DC: IEEE.

Jordan, M., Grhahramani, Z., Jaakkola, T.S., & Saul, L.K. (1999). An introduction to variational methods for graphical models. Machine Learning, 37(2), 183–233.

Kim, D. & Oh, A. (2011). Topic chains for understanding a news corpus. In Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (pp. 163–176). Berlin, Heidelberg: Springer.

Klitkou, A., & Gulbrandsen, M. (2010). The relationship between academic patenting and scientific publishing in Norway. Scientometrics, 82(1), 93–108.

Kostoff, R.N., & Schaller, R.R. (2001). Science and technology roadmaps. IEEE Transactions on Engineering Management, 48(2), 132–143.

Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., … Valencia, A. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics, 7(Suppl. 1), S2.

Lee, M., & Lee, S. (2017). Identifying new business opportunities from competitor intelligence: An integrated used of patent and trademark databases. Technological Forecasting and Social Change, 119, 170–183.

Lee, M., Lee, S., Kim, J., Seo, D., Kim, P., Jung, H., … Sung, W.-K. (2011). Decision-making support service based on technology opportunity discovery model. In T.-H. Kim, et al. (Eds.), FGIT-UNESST 2011 (Vol. 264, pp. 263–268). Berlin, Heidelberg: Springer.

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. Journal of the Association for Information Science and Technology, 59(1), 77–85.

Leydesdoref, L., & Meyer, M. (2007). The scientometrics of a triple helix of university-industry-government relations (introduction to the topical issue). Scientometrics, 70(2), 207–222.

Li, R., Chambers, T., Ding, Y., Zhang, G., & Meng, L. (2014). Patent citation analysis: Calculating science linkage based on citing motivation. Journal of the Association for Information Science and Technology, 65(5), 1007–1017.

Liang, P. (2005). Semi-supervised learning for natural language Unpublished master's thesis, Cambridge, MA: MIT.
Meyer, M. (2000). Does science push technology? Patents citing scientific literature. Research Policy, 29(3), 409–434.
Meyer, M. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. Scientometrics, 51(1), 163–183.
Meyer, M. (2002). Tracing knowledge flows in innovation systems. Scientometrics, 54(2), 212.
Michel, J., & Bettels, B. (2001). Patent citation analysis: A closer look at the basic input data from patent search reports. Scientometrics, 51(1), 185–201.
Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., & McCallum, A. (2009). Polylingual topic models. In Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 880–889). Stroudsburg, PA: Association for Computational Linguistics.
Minka, T.P. (2003). Estimating a Dirichlet distribution (Tech. Rep.). Cambridge, MA: MIT.
Mnih, A., & Andriy, G. (2009). A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), Advances in neural information processing systems 21 (pp. 1081–1088). Cambridge, MA: Curran Associates.
Narin, F. (1994). Patent bibliometrics. Scientometrics, 30(1), 147–155.
Narin, F., Hamilton, K.S., & Olivastro, D. (1997). The increasing linkage between U.S. technology and public science. Research Policy, 26(3), 317–330.
Narin, F., & Noma, E. (1985). Is technology becoming science? Scientometrics, 7(3–6), 369–381.
Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. Research Policy, 21(3), 237–249.
Narin, F., & Olivastro, D. (1998). Linkage between patents and papers: An interim EPO/US comparison. Scientometrics, 41(1–2), 51–59.
Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2009). Distributed algorithms for topic models. Journal of Machine Learning Research, 10, 1801–1828.
Newman, D., Chemudugunta, C., & Smyth, P. (2006). Statistical entity-topic models. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 680–686). New York: ACM.
Nichols, L.G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics, 100(3), 741–754.
Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., & Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the 19th International Conference on World Wide Web (pp. 751–760). New York: ACM.
Pérez-Pérez, M., Rabal, O., Pérez-Rodríguez, G., Vazquez, M., Fdez-Riverola, F., Oyarzabal, J. ... Krallinger, M. (2017). Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: The CEMP and GPRO patents tracks. In Proceedings of the BioCreative V.5 Challenge Evaluation Workshop (pp. 11–18).
Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (1992). Numerical recipes in C: The art of scientific computing (2nd ed.). New York: Cambridge University Press.
Rachev, S.T. (1998). In L. Ruschendorf (Ed.), Mass transportation problems: Volume I: Theory (probability and its applications). New York, NY: Springer.
Rip, A. (1992). Science and technology as dancing partners. In P. Kroes & M. Bakker (Eds.), Technological development and science in the industrial age: New perspectives on the science-technology relationship (pp. 231–270). Dordrecht, Netherlands: Kluwer Academic Publishers.
Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology — Case study of a solar cell. Technological Forecasting and Social Change, 77(7), 1147–1155.
Shibata, N., Kajikawa, Y., & Sakata, I. (2011). Detecting potential technological fronts by comparing scientific papers and patents. Foresight, 13(5), 51–60.
Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). BRAT: A web-based tool for NLP-assisted text annotation. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102–107). Stroudsburg, PA: Association for Computational Linguistics.
Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. Journal of the Association for Information Science and Technology, 7(10), 2464–2476.
Takano, Y., Mejia, C., & Kajikawa, Y. (2016). Unconnected component inclusion technique for patent network analysis: Case study of internet of things-related technologies. Journal of Informetrics, 10(4), 967–980.
Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476), 1566–1581.
Tian, Y. (2015). From publishing to patenting: Survey construction of Swedish academics' motivations. Unpublished master's thesis. Sweden: University of Gothenburg.
Tijssen, R.J.W. (2001). Global and domestic utilization of industrial relevant science: Patent citation analysis of science–technology interactions and knowledge flows. Research Policy, 30(1), 35–54.
Wallach, H.M., Mimmo, D., & McCallum, A. (2010). Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), Advances in neural information processing systems 22 (pp. 1973–1981). Cambridge MA: MIT Press.
Xie, P., Yang, D., & Xing, E.P. (2015). Incorporating word correlation knowledge into topic modeling. In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies (pp. 725–734). Association for Computational Linguistics.
Xu, H.-Y., Yue, Z.-H., Wang, C., Dong, K., Pang, H.-S., & Han, Z. (2017). Multi-source data fusion study in scientometrics. Scientometrics, 111(2), 773–792.
Xu, S., An, X., Qiao, X., & Zhu, L. (2014). Multi-task least-squares support vector machines. Multimedia Tools and Applications, 71(2), 699–715.
Xu, S., An, X., Zhu, L., Zhang, Y., & Zhang, H. (2015). A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. Journal of Cheminformatics, 7(Suppl 1), S11.
Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. Scientometrics, 117, 61–84. [Epub ahead of print.]. https://doi.org/10.1007/s11192-018-2841-4.
Xu, S., Zhu, L., Qiao, X., Shi, Q., & Gui, J. (2012). Topic linkages between papers and patents. In Proceedings of the 4th International Conference on Advanced Science and Technology (pp. 176–183). Daejeon, South Korea: Science and Engineering Research Support soCiety (SERSC).
Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. Scientometrics, 100(3), 767–786.
Zhang, C. & Zhang, Q. (2008). Topic navigation generation using topic extraction and clustering. In Proceedings of the International Symposium on Knowledge Acquisition and Modeling.
Zhang, Y., Shang, L., Huang, L., Porter, A.L., Zhang, G., Lu, J., & Zhu, D. (2016). A hybrid similarity measure method for patent portfolio analysis. Journal of Informetrics, 10(4), 1108–1130.