Identify Topic Relations in Scientific Literature Using Topic Modeling

Hongshu Chen, Ximeng Wang, Shirui Pan, and Fei Xiong

Abstract—**Over the past five years, topic models have been applied to bibliometrics research as an efficient tool for discovering latent and potentially useful content. The combination of topic modeling algorithms and bibliometrics has generated new challenges of interpreting and understanding the outcome of topic modeling. Motivated by these new challenges, this paper proposes a systematic methodology for topic analysis in scientific literature corpora to face the concerns of conducting post topic modeling analysis. By linking the corpus metadata with the discovered topics, we feature them with a number of topic-based analytic indices to explore their significance, developing trend, and received attention. A topic relation identification approach is then presented to quantitatively model the relations among the topics. To demonstrate the feasibility and effectiveness of our methodology, we present two case studies, using big data and dye-sensitized solar cell publications derived from searches in World of Science. Possible application of the methodology in telling good stories of a target corpus is also explored to facilitate further research management and opportunity discovery.**

Index Terms—**Bibliometrics, tech mining, text mining, topic analysis.**

I. Introduction

TECHNOLOGY development today has dramatically accelerated the emergence of data that describes science, technology, and innovation (ST&I) activities, especially scientific literature, and has brought the research on bibliometrics to the age of big data [1]. To improve opportunities for collaboration and funding, the ST&I data from various sources needs to be fused to provide researchers with better understanding of the overall technological landscape, and gain interpretable insights from the existing data [2], [3]. The growth of scientific publications, however, has created information overload [4], whereby researchers, analysts, and decision makers face difficulties in handling, understanding, and analyzing massive textual data [5], [6]. Text mining as a valuable instrument for ST&I, in such circumstances, has attracted widespread interest in augmenting and amplifying the capability of domain experts when dealing with real-world tasks [7].

Text mining in bibliometrics research seeks to identify the commonalities within scientific publications to shape the possible structures and relationships that underlie the data [8]. Much effort has already been devoted to selecting valuable terms and phrases, distinguishing the structures in text, or exploring the similarities between documents to identify meaningful groups, with text clustering, keyword-based morphology, term clumping, and other techniques [9]–[13]. These approaches have been confirmed as performing well in specific fields at specific scales, however, they hold some major concerns. First, although similarity measurement can provide a solid clustering result for document groups, the relations among topics remain confusing [14]. It is even more difficult to understand how information and knowledge are spread by topic relations [15], since frequently used document clustering models are mainly based on the simple assumption that each file is related to only one cluster or one topic. Second, even though many approaches have achieved sound clustering results, correlated postprocessing topic analysis is seldom mentioned systematically. As a result, how to tell good story using clustering products is still under discussion, and demands further research and analysis.

Motivated by these concerns, this paper introduces a topic relation identification methodology after applying topic modeling to massive scientific literature. In past five years, topic model-based approaches have attracted increasing interest in bibliometrics [8], [16], [17]. Wei and Croft [18] have demonstrated that topic models outperform most cluster-based approaches in information retrieval. In particular, latent dirichlet allocation (LDA), one of the most well-known probabilistic topic models, has been applied in analyzing citation networks, time gaps, content comparison, and scientific maps of publications in various areas [19]–[21]. This paper presents a methodology to fully bring the superiority of LDA, providing an interpretable soft clustering of documents, into play and overcomes the drawbacks of the lack of metadata fusions in existing LDA implementations [22]. By comprehensively linking publications' metadata and the discovered topics, we conduct a post topic modeling process to assist the understanding of the clustering products, in which a number

Manuscript received October 16, 2018; revised January 29, 2019 and February 19, 2019; accepted February 27, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61872033, in part by the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant 18YJCZH204, and in part by the Beijing Natural Science Foundation under Grant 4184084. Review of this manuscript was arranged by Deparment Editor Y. Zhang. *(Corresponding author: Fei Xiong.)*

H. Chen is with the School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China (e-mail: Hongsue1114@hotmail.com).

X. Wang is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China, and also with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: wangxm@bjtu.edu.cn).

S. Pan is with the Faculty of Information and Technology, Monash University, Clayton, VIC 3800, Australia (e-mail: shirui.pan@monash.edu).

F. Xiong is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: xiongf@bjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TEM.2019.2903115

of analytic indices are proposed to quantitatively characterize the topic-based weight, trend, and citation for all the topics. Finally, topic-based relations are identified to quantitatively model the correlations and weights among the topics. To demonstrate our approach, we present two case studies with big data and dye-sensitized solar cell (DSSC) publications derived from searches in World of Science (WoS). Topic-based analytic and relation maps are given to visualize the networks of the identified topics.

The main contributions of this paper are summarized as follows: 1) a systematic methodology for post topic modeling analysis on a target corpus is provided; and 2) a topic relation identification approach is presented to model the relations and weights among discovered topics quantitatively.

The rest of this paper is organized as follows. Section II—Related work reviews text clustering, LDA, and topic modeling in scientiometrics research. Section III—Methodology describes the full process of the proposed topic relation identification methodology. Section IV—Case studies and result presents experiments using WoS papers to examine the methodology and then explains how to use it in the context of real scientific literature analysis. Finally, Section V concludes this paper.

II. Related Work

A. Text Clustering and Postclustering Analysis

Text clustering aims to calculate the similarity between documents and reduce dimensionality by grouping massive items into a small number of sets [12]. Citation-based clustering approaches [23], [24], content-based clustering approaches [25], [26], and hybrid methods [27] have been successfully introduced to bibliometrics research with varied performance [28]. Boyack *et al.* [9] compared nine similarity analytical approaches with two million biomedical publications and demonstrated that the probabilistic topic-based model for content similarity measurement proposed by Lin and Wilbur [29] performed best. The results of these approaches were mixed according to Boyack and Klavans [30]. They usually performed well in limited scopes, on a specific dataset, but could not be readily adapted to other real-world datasets [14].

In existing research, the combination of traditional bibliometric approaches and text clustering has become a well-accepted way to interpret the product of a clustering process, while outcomes are explained using graphs, networks, maps, and so forth [31], [32]. Technology roadmapping [33] is one of the best methods for joining textual knowledge and the results of bibliometric analysis. The keyword clusters or meaningful groups of terms are defined as "topics." After topic detection, topic interpretation and topic visualization are the two research strengths of postclustering analysis to present topical features [6], [22], [34]. The most frequently used indicators to reveal these features are topical weight and trend, reflecting whether a topic is comparatively more prominent and will grow or decline in the near future [35]. Based on the concept of weight, different terms are used to characterize the visibility of a topic, for example, prominence, importance, and hotness. In addition, as the fundamental of investigating the patterns of emergence and topic evolution, trend detection also drew great attention [36].

B. Topic Modeling in Bibliometrics Research

Topic models are generative models that measure the possibilities of the "co-occurrence" of topics and documents, and then use their distributions to present concepts. Instead of simply modeling massive textual data based on "keywords and frequency," the semantic meaning of a concept can be better derived by these words and topic distributions, also opening the possibility of better evaluating and understanding text mining outcomes [20], [37]. In practice, LDA is the most frequently used topic modeling approach that uses unsupervised learning to calculate the properties of multinomial observations [38]. It estimates latent topics and the probability of how various documents relate to different topics [39].

Over the last five years, LDA has been applied to bibliometrics as an efficient tool for discovering underlying and potentially useful content. Ding [19] introduced topic-dependent ranks using a combination of a topic model and a weighted PageRank algorithm; Liu and Chen [40] compared latent topics identified from citing abstracts versus citing sentences to the target reference using LDA; Yau *et al.* [8] investigated LDA and its extensions by separating a set of scientific publications into several clusters; De Battisti *et al.* [22] presented an LDA-based postprocessing approach to describe a field of research over time; and Suominen and Toivanen [21] validated an unsupervised learning-based map of science with the assistance of LDA.

C. Latent Dirichlet Allocation and Its Evaluation

In the areas of data mining and machine learning, LDA has been used as a very efficient tool to assist topic discovery in large volumes of textual data: Griffiths and Steyvers [41] applied LDA-based topic modeling to discover the hot topics covered by the scientific journal of PNAS; Yang and his colleagues [42] proposed a topic expertise model for community question answering based on LDA to jointly model topics and expertise; Kim and Oh [43] proposed a framework based on LDA to identify important topics within news archives on the web; Broniatowski and Magee [44] studied knowledge boundaries barriers to knowledge transfer in groups of experts using LDA, and so forth.

The generative process of LDA is denoted by the joint distribution of random variables [38], [45]. As shown in formula 1, the overall documents are denoted as D, the term number of the dth document is presented as N_d and the nth word in document d is $w_{d,n}$, which are all observable. The topic numbers K, the topic proportions for the dth document $\overrightarrow{\vartheta_d}$, the topic assignments Z_d, and the vocabulary distribution for the kth topic $\overrightarrow{\varphi_k}$, however, are all latent random variables. α and β are two hyperparameters that determine the amount of smoothing applied to the topic distributions for each document and the word distributions for

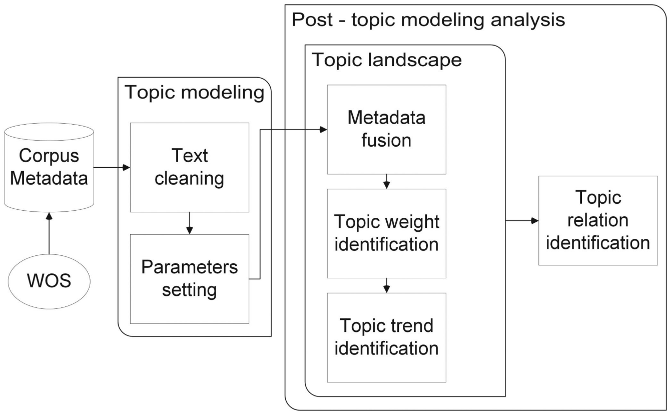

Post - topic modeling analysis
Topic landscape
Topic modeling
Corpus Metadata
WOS
Text cleaning
Parameters setting
Metadata fusion
Topic weight identification
Topic trend identification
Topic relation identification

Fig. 1. Framework of posttopic modeling analysis and topic relation identification method.

each topic.

$$p(\overrightarrow{w_d}, \overrightarrow{z_d}, \overrightarrow{\vartheta_d}, \phi | \overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{n=1}^{N_d} p(w_{d,n} | \overrightarrow{\varphi}_{z_{d,n}})$$
$$p(z_{d,n} | \overrightarrow{\vartheta_d}) p(\overrightarrow{\vartheta_d} | \overrightarrow{\alpha}) p(\phi | \overrightarrow{\beta}). \quad (1)$$

The evaluation of topic models mainly focuses on how to select models and parameters to reveal the precious insights of latent semantic knowledge. This requirement leads the research directly to the question of how to determine the quantity of topics within an LDA implementation. Generally, higher values will increase processing time, while lower values will produce large topical granularity [46]. To the best of our knowledge, likelihood [41] and perplexity [22] are the two main approaches that have been used to decide topic numbers when modeling scientific topics.

III. Methodology

We retrieved the abstracts of the WoS articles within the target technological scope. The corpus also consists of the metadata and citations of those publications, including ISI unique article identifier, publication year, authors, affiliations, and so forth. Each abstract constitutes one text document, while the publication metadata and the citation comprise two single files. After data retrieval, we process the data and conduct topic modeling, to first discover the latent topics of the corpus and then understand the full landscape of the target area by revealing the features reflecting whether a topic is comparatively more visible and has potential to grow. A topic relation identification method is proposed to finally reveal the topical network systematically. An overview of the framework we used for post topic modeling analysis is shown in Fig. 1.

A. Topic Modeling

1) Text Cleaning: LDA follows the assumption bag of words [39]. Every unique term in the target corpus is counted when assigning words to themes, and hence unnecessary or meaningless elements need to be removed before topic modeling. In order to maintain technical vocabularies only, the terms in our target corpus must be cleaned and consolidated. First, textual data is segmented into a unique vocabulary list. All punctuation and nonalphabetic characters are removed, leaving only plain English terms. We then apply a stop words list,[1] publication-related thesauri [12], and high frequency academic words list[2] to remove meaningless terms that provide little or no contribution to the technological topics. Given we want the final model to return unobserved, but potentially useful concepts and topics, not general ideas, we consolidate and exclude all the common words used in that scientific area. The identification for the common words is a combined process of statistical calculation and expert decision making. The top terms appearing in more than 50% of the total records need to be presented to domain experts, seeking their advice to keep, remove, or combine the terms. For example, DSSC publications are used in one of our case studies, following suggestions of experts, we removed the terms of DSSCs, solar, and cell from the corpus and combined another two high-frequency terms, titanium dioxide, into one phrase, titanium dioxide.

2) Parameters Setting: In majority of the cases, we prefer a model to characterize documents by few topics and also assign few terms to each topic for better interpretation. This can be done by lowering the values of α and β in the LDA-based topic model, which will result in more decisive topic associations [46]. In practice, α and β are usually set as 0.5 and 0.01 to provide a fine-grained decomposition of the document collection [41]. Actually, nonparametric LDA models have been developed in the machine learning research, which can determine the number of topics in a probabilistic perspective automatically. However, such number sometimes can be too large for human interpretation. In this paper, we determine the parameter K based on perplexity calculation and also prior knowledge. We run a number of experiments for the number of topics in an interval [10, 50] and compute the perplexity scores to measure how well a trained model fits the dataset with a specific choice of K. The larger the perplexity score is, the higher misrepresentation of the words of the document set it indicates. With the help of expert knowledge, the K value presents comparatively better representation of the words with lower perplexity value is selected.

Perplexity is defined as the reciprocal geometric mean of the likelihood of a test corpus [31], [47] as follows:

$$\text{Perplexity}(D) = \exp - \frac{\sum_{d=1}^{M} \log(p(w))}{\sum_{d=1}^{M} N_d} \quad (2)$$

where N_d is the document length of document d in D, which has M documents, and $\sum \log(p(w))$ represents the likelihood of a corpus given the trained model.

[1] SMART Stopword List: [Online]. Available: http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[2] Academic Vocabulary: [Online]. Available: http://www.nottingham.ac.uk/alzsh3/acvocab/word-lists.htm

B. Post Topic-Modeling Analysis and Topic Relation Identification

Focusing on the major concerns of telling good story in topic analysis, also measuring the relations among topics by assigning one file to different topics, we fuse the corpus metadata with the discovered topics. A number of analytic indices are proposed to quantitatively characterize the topic-based weight, trend, and attention, and a topic relational network is proposed to gain insight on a semantic level.

1) Metadata Fusion: LDA is generally used as a simple topic identification tool working on textual data in the existing bibliometrics research, as it is difficult and time consuming to analyze no matter topic trend or relations since temporal, spacial, or relational features of the corpus are presented in the metadata. As a result, linking the metadata in LDA implementations remains an effective approach to these problems.

In the perspective of topic modeling, topics are the semantically meaningful decompositions of the target corpus; vice versa, every document covers multiple topics with different proportion. We name documents in the corpus with their ISI unique article identifiers, in which a unique key array can be extracted from the metadata. For a document i, the topic with the highest proportion can be denoted as $t_{i_{(1^{\rm st})}}$. This new feature, $t_{i_{(1^{\rm st})}}$ is then assigned to all the articles; a group of corresponding ISI unique identifiers thus can be listed for each topic. Topics are then linked with metadata via the identifiers. It is worth noting that here we are seeing the topic with the highest proportion as a new feature, and for linking the metadata purpose only. We still follow the assumption that each file is related to more than one topic.

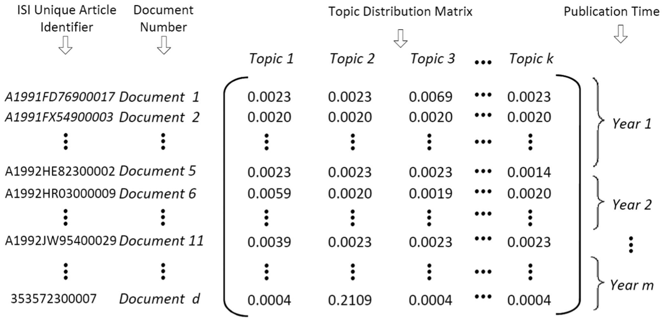
In such perspective, citation information can be linked to topics as well. For topic j, the accumulative citations of the articles with the same $t_{(1^{\rm st})}$, is denoted as $\overrightarrow{C_j}$. According to previous research, the more citations a paper carries, the more relevance it has to a scientific area, and the more contribution it can be considered to have given. Moreover, the more likely it is to be cited itself [48]. $\overrightarrow{C_j}$ is used to define the received attention and potential usefulness of the topic.

2) Topic Weight: A topic can be defined as a semantic concept in a corpus with a specific proportion. Those topics with higher proportion, or we say high popularity topics (i.e., hotspots), are one of our main concerns. We have decomposed D documents to K themes after executing LDA. The proportion of topics comprising the whole corpus is recorded as a topic distribution matrix $\Theta = (\vartheta_{ij})_{D\times K}$. Each column of the matrix indicates how a topic is distributed over different documents. The bigger the total proportion a topic has in the whole corpus, the larger weight it has. Thus, we define the weight of topics as $\mathrm{WI} = (wi_1, wi_2, wi_3, ..., wi_k)$

$$wi_k = \sum_{i=1}^{D} \vartheta_{ik} \tag{3}$$

where wi_k indicates the sum up of column k of Θ.

3) Topic Trend: In addition to the weight, the developing trend of the discovered topics is also very important when gaining the insight to evaluate how a topic is keeping up with the time. We consider the publication year as a temporal label for the document collection, and process them in ascending order to obtain a topic distribution matrix in chronological order, as shown in Fig. 2. We then sum the group of elements in each column that was associated with literature published in the same year, and the total is the annual weight of the corresponding topic.



ISI Unique Article Identifier	Document Number	Topic 1	Topic 2	Topic 3	•••	Topic k	Publication Time
A1991FD76900017	Document 1	0.0023	0.0023	0.0069	•••	0.0023	Year 1
A1991FX54900003	Document 2	0.0020	0.0020	0.0020	•••	0.0020	
⋮	⋮	⋮	⋮	⋮	•••	⋮	
A1992HE82300002	Document 5	0.0023	0.0023	0.0023	•••	0.0014	Year 2
A1992HR03000009	Document 6	0.0059	0.0020	0.0019	•••	0.0020	
⋮	⋮	⋮	⋮	⋮	•••	⋮	
A1992JW95400029	Document 11	0.0039	0.0023	0.0023	•••	0.0023	⋮
⋮	⋮	⋮	⋮	⋮	•••	⋮	Year m
353572300007	Document d	0.0004	0.2109	0.0004	•••	0.0004	

Fig. 2. Example of topic distribution matrix in chronological order.

Specifically, matrix $W_{m\times k}$ represents the annual weight of all estimated k topics that appeared during m years. We then calculate the average proportion of related topics in all articles and applying a linear fit approach to estimate their trends. In a least-squares sense, the average annual weight values of the kth topic can be fitted to a univariate quadratic polynomial

$$\frac{\overrightarrow{w_k}}{Q_{t_i}} = a_k t_i^2 + b_k t_i + c_k \tag{4}$$

where $\overrightarrow{w_k}$ stands for the annual weight, Q_{t_i} indicates the quantity of publication in year t_i $(1 < i \leq m)$. We use the coefficients a_k and b_k to measure the developing trends of the corresponding topic, since a_k controls the speed of increase (or decrease) of the quadratic function, $-b_k/2a_k$ control the axis of symmetry. For instance, if coefficient, a_k, is positive and the symmetry is on the left of *y*-axis, we consider the corresponding topic has a growing trend. The greater a_k is, the faster the growth will be. In summary, the topic trend index is defined as $\mathrm{TI} = (ti_1, ti_2, ti_3, ..., ti_k)$, in which ti_k equals to the coefficients a_k of the quadratic function that we fitted the annual weight to as follows:

$$ti_k = \left(\frac{\overrightarrow{w_k}}{Q_{t_i}} - b_k t - c_k\right)/t^2. \tag{5}$$

4) Topic Relation Identification: Frequently used document clustering models are mainly based on the assumption that each file is related to only one cluster or one topic, yet this assumption is too simple to effectively model a large corpus [18]. For example, it would be imprudent to assign a paper discussing AI-powered systems is placing jobs in banks into just one of groups of AI, bank or job-cut; in addition, these three topics are actually linking to each other, and cannot be seen as three separated factors if we want to gain insight from them. LDA, on the other hand, provides a statistical soft clustering of documents. One document actually associates all the topics with different possibilities. We apply this feature to conduct topic relations identification.

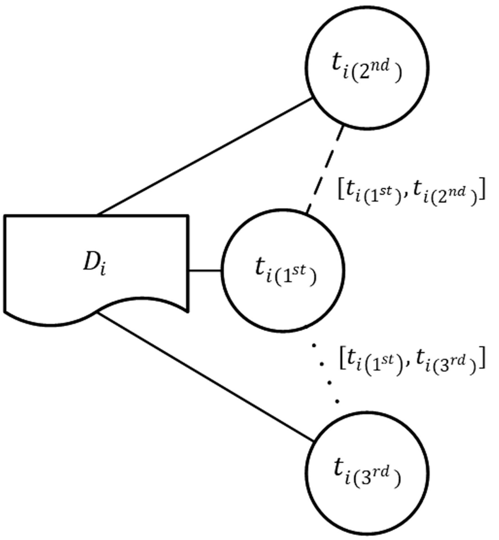

$t_{i(2^{nd})}$
$[t_{i(1^{st})}, t_{i(2^{nd})}]$
D_i
$t_{i(1^{st})}$
$[t_{i(1^{st})}, t_{i(3^{rd})}]$
$t_{i(3^{rd})}$

Fig. 3. Topic# of the three topics that D_i most possibly described and their corresponding topic# pairs.

To present a clear relation structure and reduce the computational complexity, we assume that each paper in the target corpus discusses three topics with the top highest topic proportions; these three topics are related at a topical level. For all the discovered topics, we then define their semantic relations with the co-occurrence statistics between them. In the topic set, this implies that each topic has latent but trackable relations with other two topics with different weights. We create a topic# pair (note that topic# here means the label of topics that were automatically generated by LDA, not the quantity of topics) to denote the relation between two topics; and a topic co-occurrence matrix to present the relations among all the topics.

As shown in Fig. 3, topic# of the three topics that a document, D_i, most possibly described, are captured as $t_{i_{(1^{st})}}, t_{i_{(2^{nd})}}, t_{i_{(3^{rd})}}$, and $t_{i_{(1^{st})}} \neq t_{i_{(2^{nd})}} \neq t_{i_{(3^{rd})}}$. For D_i, focusing on the major topic it covers, we can identify two main topic# pairs $[t_{i_{(1^{st})}}, t_{i_{(2^{nd})}}]$ and $[t_{i_{(1^{st})}}, t_{i_{(3^{rd})}}]$, indicating topic correlation between $t_{i_{(1^{st})}}$ and $t_{i_{(2^{nd})}}$, also $t_{i_{(1^{st})}}$ and $t_{i_{(3^{rd})}}$. After traversing the topic distribution matrix Θ, we can calculate and receive all the topic# pairs and convert these pairs into a topic co-occurrence matrix. Following from the definition, it is not hard to understand that $[t_{i_{(1^{st})}}, t_{i_{(2^{nd})}}]$ has a stronger connection than $[t_{i_{(1^{st})}}, t_{i_{(3^{rd})}}]$, we thus assign different weights to $[t_{i_{(1^{st})}}, t_{i_{(2^{nd})}}]$ and $[t_{i_{(1^{st})}}, t_{i_{(3^{rd})}}]$. In this paper, the relation weight of $[t_{i_{(1^{st})}}, t_{i_{(2^{nd})}}]$ is set as 1, and for $[t_{i_{(1^{st})}}, t_{i_{(3^{rd})}}]$, it is set as 0.5, to reduce the computational consumption. Every document in the target corpus contributes a piece of weight to the relations of the major topics it covers. We sum up the weight values for every unique topic# pair and form a final topic co-occurrence matrix.

IV. Case Study and Result

To demonstrate the feasibility and effectiveness of our methodology, we present two case studies, for big data and DSSC areas, using the publications derived from searches in WoS. The reliability of the topic modeling result is shown by comparing the content of our topics with the previous applied research for these two areas, which has been assessed by domain experts. Specifically, the result of big data topic analytics is compared with [49]; and the outcome of DSSC topic-based analysis is verified by [12].

A. Topic Analysis for Big Data Publications

1) Data and Parameter Setting: We choose one of the most representative emerging technologies, big data, to present a case study, using its related publications derived from searches in WoS. The big data corpus in this research contains 5450 abstracts of related papers published between 2000 and 2015, covers content from terms describing massive information to the main techniques in big data research, which is collected using the search strategy proposed in [49]. The title and abstract fields are combined and represented with one text file, carrying a unique ISI article identifier. We put the publication metadata in a separated file. Metadata is parsed into publication year, authors, affiliations, and number of citations.

After preprocessing and text cleaning, the big data corpus remains 15 585 terms. The hyperparameters were set to $\alpha = 0.5$ and $\beta = 0.01$. We performed each run with 5000 iterations of Gibbs sampling. We then compute the perplexity score with K values from 10 to 50, in which $K = 35$ is selected as the suitable topic number. This number presents comparatively lower misrepresentation of the words, and better capture of the topics with the consideration of easier interpretation.

2) Topic Modeling and Analysis: For the area of big data, we generate 35 topics from 5450 articles. These topics reveal a number of important tools and techniques, algorithms, and applications of big data research in the past 16 years. As shown in Table I, for presentation convenience, we give each topic a short label, with a topic# in it, followed by the topic name and topic description (please note that topic# here means the label of topics that were automatically generated by LDA, not the rank of topics).

Tools and techniques such as MapReduce and Hadoop (T19-Hadoop), parallelism (T06-Para), and large-scale data collection (T09-LarSca) play a very important role in the existing big data research. In addition, some existing algorithms in artificial intelligence area have been boosting the research and application of big data. These algorithms include machine learning (T28-ML), data mining (T32-DatMin), optimization (T03-Opti), Bayesian probability (T27-Bayes), clustering (T16-Clust), text mining (T29-TexMin), neural network (T34-NeuNet), mathematical modeling (T22-Math), recommendation system (T23-RS), and forecasting (T25-Forcas). They cover a large content of big data research as shown in Table 1. Last but not the least, big data research has been applied to a broad range of application areas, in which bioinformatics (T18-BioInf) is one of the most hot topics. Other applications, such as decision making (T10-DecMak), social network (T17-ScoNet), Internet of Things (T26-IOT), epidemiology (T02-Epidem), and healthcare (T31-Health) have been attracting researchers and analysts' attention significantly. Compare our discovered topics with the previous research that

TABLE I
TOPIC MODELING RESULT FOR BIG DATA RESEARCH

Topic Label	Topic Name	Topic Description
T01-Workfl	Scientific Workflow	scientific, workflow, interdisciplinary, computing, cyberinfrastructure
T02-Epidem	Epidemiology	population, observational, epidemiology, diabetes, prevalence
T03-Opti	Optimization	optimization, partitioning, NoSQL, scalability, genetic-algorithm
T04-Annota	Annotation	annotation, graphical, curation, repository, biological
T05-Genome	Genome Data	genome, DNA, NGS (next-generation sequencing), RNA, microarray
T06-Para	Parallelism	scalability, large-scale, parallelism, processors, multicore
T07-Govern	Data Protection	collection, protection, governance, transparency, reuse
T08-Archte	Architecture	architecture, deployment, enterprise, bandwidth, scalable
T09-LarSca	Large-Scale Data	dataset, large-scale, collections, ecological, biodiversity
T10-DecMak	Decision Making	analytics ,decision-making, GPU, competitive, opportunities
T11-Simula	Simulations	simulations, large-scale, observations, frequencies, calculations
T12-3D	3D Localization	3D, localization, interpolation, automata, instrumentation
T13-Relat	Relationship	relationship, correlation, subgraph, behavioral, differences
T14-BenchM	Benchmarking	historical, productivity, benchmarking, socio, disciplines
T15-GIS	Geographic IS	geographic, information-system, geospatial, popularity, behaviour
T16-Clust	Clustering	clustering, decomposition, dimension, unsupervised, pattern-recognition
T17-ScoNet	Social Network	social-network, opportunities, tweets, facebook, decision-making
T18-BioInf	Bioinformatics	biological, bioinformatics, genomics, proteomics, spectrometry
T19-Hadoop	MapReduce&Hadoop	MapReduce, Hadoop, scalability, cloud-computing, apache
T20-Compres	Compression	compression, recognition, preserving, bandwidth, wavelet
T21-ReaTim	Real-Time	real-time, representation, fuzzy, trajectory, spatio-temporal
T22-Math	Mathematical Modelling	mathematical, modeling, convergence, computation, combination
T23-RS	Recommendation System	recommendation, personalized, uncertainty, collaborative, preference
T24-Visual	Visualizations	visualization, exploratory, biomarkers, simultaneous, phylogenetic
T25-Forcas	Forecasting	forecasting, analytics, forecast, trend, historical
T26-IOT	Internet Of Things	IOT, surveillance, middleware, smartphones, google, ubiquitous
T27-Bayes	Bayesian Probability	regression, probability, bayesian, multivariate, stochastic
T28-ML	Machine Learning	machine-learning, classification, classifier, SVM, large-scale
T29-TexMin	Text Mining	text-mining, volumes, congestion, computing, cyber
T30-NeuIma	Neuroimaging	neuroimaging , large-scale, connectivity, connections, neuroscience
T31-Health	Healthcare	predictive, healthcare, electronic, clinical, diagnostic
T32-DatMin	Data-Mining	data-mining, large-scale, metrics, knowledge-discovery, topological
T33-MatDat	Metadata	metadata, heterogeneous, XML, semistructured, ontologies
T34-NeuNet	Neural-Network	neural-network, artificial, large-scale, OLAP, AI
T35-QOS	QOS	QOS, composition, improvements, propagation, large-scale

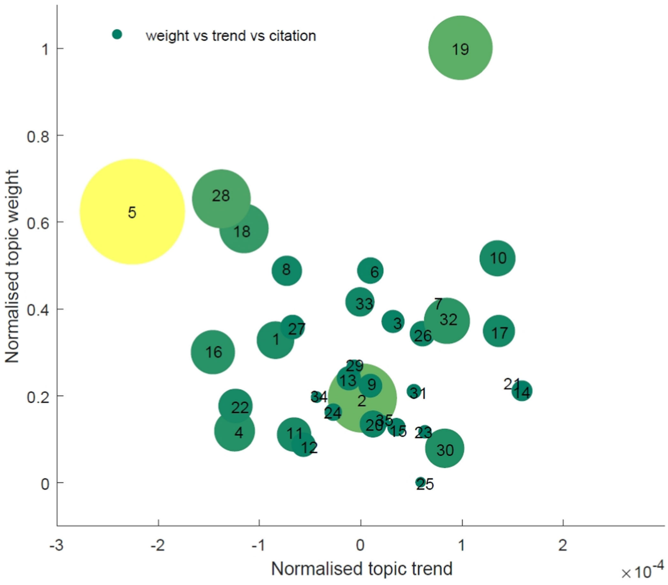


Fig. 4. Topic-based analytic map based on weight, trend, and citation characteristics of big data corpus.

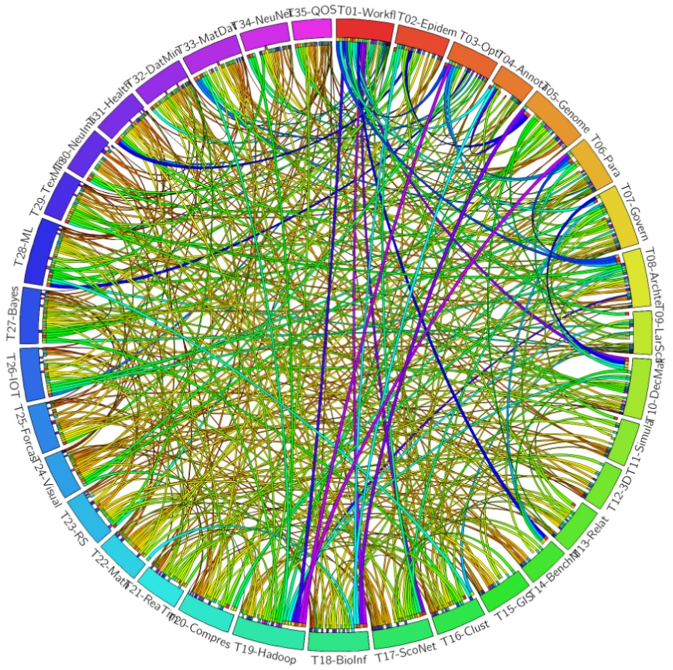


Fig. 5. Co-occurrence map of topics in big data research.

verified by expert panel [13], both "sleeping beauty" topics and major clusters of big data research have been covered; although we provide comparatively smaller size of topics because each one is explained by a group of terms, it actually provide richer information than using the single phrase topics.

We proceed the post topic modeling process to reveal more interesting features of these topics. As shown in Fig. 4, the values of topic weight, trend, and citation indices are visualized on a scatter plot chart using MATLAB. The *x*-axis represents the normalized trend index, which has the values between [−1, 1], explaining how "active" the corresponding topic is; the *y*-axis denotes the normalized weight index, which has the values between [0, 1], providing an insight of how popular this topic is; and the size and the color of the dots represent the accumulative citations, which are normalized to [0, 1]. The warmer the color and the larger its dot, the more existing influence and potential usefulness a topic receives. The numbers on the dots in Fig. 4 correspond with the numbers in the topic labels in Table I. If a topic has a positive trend index value, it means the average proportion of it in all articles has been increasing during the 16 years of research. The topic is then considered growing. In contrast, a negative value of trend index implies the topic has been mentioned less in the corpus averagely.

As shown in the Fig. 4, MapReduce&Hadoop (T19-Hadoop), which covers the content of MapReduce, Hadoop, scalability, cloud computing, and apache, is the most popular and also a very fast developing topic. It also has received quite high citations, showing strong influence on other topics in the big data area. Bio-related and medical applications of big data, genome data (T05-Genome), epidemiology (T02-Epidem), bioinformatics (T18-BioInf), and healthcare (T31-Health), comparatively received much higher citations than other topics. One of the interesting findings is that, although traditionally genome data (T05-Genome) and bioinformatics (T18-BioInf) can be considered as the most successful big data applications, the developing speed of these two topics are dropping, which means the two topics are mentioned less, in the recent published articles than before. In contrast, healthcare (T31-Health) is mentioned more frequently in recent years, even though its citation is still quite low. This may imply the research stress in big data application has changed to some new topics, like public health, internet of things (T26-IOT), and geographic information systems (T15-GIS). Generally, these applications focus on decision making (T10-DecMak) in different scenarios. One of another interesting finding is, machine learning (T28-ML) is one of the hotspots in big data research, and it has received a lot of citations. However, the developing trend of it is downward, which means the terms of machine learning, classification, classifier, and SVM are less active than before.

3) Topic Relation Identification: In this paper, we generate a topic co-occurrence matrix and visualizing it via Circos [50]. As shown in Fig. 5, we use each segment to represent a topic; the topic is marked with the short label given in Table I. The ribbons between segments stand for the semantic relations of two topics. For example, a stronger relationship between the two linked segments is represented by a wider ribbon, which is distinguished by a more intense blue.

We can observe in Fig. 5, that the MapReduce&Hadoop (T19-Hadoop) as one of the hottest topics in the area, has very strong relations with scientific workflow (T01-Workfl), optimization (T03-Opti), parallelism (T06-Para), and architecture (T08-Archte). This result match with the fact that MapReduce and Hadoop are still the two leading tools in big data research [49]. Topic of epidemiology (T02-Epidem) closely correlates with healthcare (T31-Health) and genome data (T05-Genome).

TABLE II
TOPIC MODELING RESULT FOR DSSC RESEARCH

Topic Label	Topic Name	Topic Description
T01-Resist	Resistance	Pt, resistance, electrocatalytic, voltammetry, pedot
T02-DyePho	Dye-Sensitized Photoelectrochemical Cell	dye, titanium-dioxide, sensitised, photo-electrode, dye-adsorption
T03-AbsSpe	Absorption Spectroscopy	absorption, spectroscopy, photoinduced, excitation, fluorescence
T04-Redox	Redox Regeneration	redox, regeneration, acetonitrile, iodide, triiodide
T05-Rsens	Ruthenium Sensitizers	ligand, sensitizers, absorption, Ruthenium(ii), Bipyridyl
T06-NanoP	Nanoparticles	nanoparticles, graphene, nanocomposite, dispersion, fabrication
T07-NanoT	Nanotube	nanotubes, titaniumdioxide, Ti, anodization, fabricated
T08-SolGel	Sol-Gel	sol-gel, mesoporous, titania, anatase, nanocrystalline
T09-TiO2	Titanium-Dioxide	titanium-dioxide, photoanode, scattering, nanofibers, electrospun
T10-MetFre	Metal-Free Sensitizers	sensitizers, photovoltaic, absorption, thiophene, metal-free
T11-NanoC	Nanocrystalline	nanocrystalline,titaniumdioxide, nanoporous, Til4, anatase
T12-RamSpe	Raman Spectroscopy	spectroscopy, x-ray, microscopy, uv-vis, photoelectron, raman
T13-SemCon	Semiconductor Materia	photovoltaic, semiconductor, solid-state, heterojunction, fabrication
T14-FilNano	Film Nanocrystalline	film, Tio2, nanocrystalline, photocurrent, fabricated
T15-Deposition	Deposition	deposition, FTO, film, transparent, ITO, fluorine-doped
T16-ShoCir	Short-Circuit	short-circuit, open-circuit, photocurrent, conversion-efficiency, photovoltage
T17-SolEle	Solid Electrolyte	electrolytes, conductivity, quasi-solid-state, quasi-solid, polyethylene
T18-PhoSpe	Photocurrent Spectroscopy	recombination, spectroscopy, photocurrent, photovoltage, resistance
T19-Absorp	Absorption	electronic, absorption, DFT, time-dependent, excitation
T20-Hydrot	Hydrothermal Synthesis	anatase, nanorods, hydrothermal, rutile, synthesizesis
T21-ConEff	Conversion-Efficiency	conversion-efficiency, photoelectrode, nanoparticles, photovoltaic, photoelectric
T22-Photoc	Photocatalytic Activity	photocatalytic, degradation, titanium-dioxide, photoelectrochemical, methylene
T23-Ultrav	Ultraviolet	ultraviolet, DNA, sensitivity, UV-B, luminescence
T24-Electr	Electrolyte	electrolytes, conductivity, triiodide, imidazolium, concentrations
T25-Porphy	Porphyrin	porphyrin, synthesized, absorption, phthalocyanine, voltammetry
T26-SolSta	Solid-State	solid-state, photovoltaic, fabricated, spiro-OMeTAD, Zn2SnO4
T27-NanoW	Nanowires	nanowires, nanostructures, morphology, nanosheets, photoanode, 3d
T28-Photoe	Photoelectrochemical Performance	quantu-dots, deposition, photovoltaic, recombination, photoelectrochemical
T29-PhoCur	Photon-To-Current	photon-to-current sensitizers wavelength nanocrystalline squaraine
T30-Sno2	Synthesized Sno2	Sno2, fluorescence, bodipy, chromophores, fluorescence

Other interesting correlations include the one between annotation (T04-Annota) and genome data (T05-Genome); the connection between topic of machine learning (T28-ML) and optimization (T03-Opti); the one between data protection (T07-Govern) and decision making (T10-DecMak); the one between genome data (T05-Genome) and bioinformatics (T18-BioInf), which is not that surprising; and the one between parallelism (T06-Para) and metadata (T33-MatDat). It is worth noted that the link between data protection and decision making tells the story of data privacy being highly concerned. Zhang *et al.* [49] provided the evidence of this highlight with big data report in 2014. Using the result from topic modeling, we can get the insight that, the data collection, protection, governance, transparency, and reuse are the main concerns for data privacy.

B. Topic Analysis for DSSC Publications

1) Data and Parameter Setting: To further demonstrate the feasibility of our methodology, we present another case study using publications of DSSC. In the past 20 years, DSSC is a promising technology that can add functionality and lower costs to enhance the value proposition of solar power generation, thus this area is investigated by many researchers for technological forecasting and analysis purpose [51]. Our DSSC corpus derived from WoS database contains 12 435 abstracts of related papers published between 1991—the year that DSSC were first announced in *Nature*—and 2014. The data derived from a multistep Boolean search algorithm and applied via the WoS database [52]. Same as the previous case study, each abstract represents one text document and carries a unique ISI article identifier. The metadata for all publications are stored in a single file.

After preprocessing and text cleaning, this corpus remains 41 214 terms. The hyperparameters were also set to $\alpha = 0.5$ and $\beta = 0.01$. We performed each run with 5000 iterations of Gibbs sampling and compute the perplexity score with K values from 10 to 50, in which $K = 30$ is selected as the suitable topic number.

2) Topic Modeling and Analysis: After topic modeling, we discovered 30 latent semantic topics in the target corpus. The topic distribution matrix is presented as a separate file named “theta,” where $\Theta = (\vartheta_{ij})_{12345\times30}$. The topic name and topic description are shown in Table II. For presentation convenience, a short label with a topic# is assigned to each topic (please note that topic# here means the label of topics that were automatically generated by LDA, not the rank of topics). These topics in Table II reveals a number of research objects, techniques, and application of DSSC research during the past 24 years. We compared our result with the previous study in [12]. All the top keywords in their clusters are caught by our topics. For example, our topic Conversion-Efficiency (T21-ConEff) matches with the cluster in their research, which has the main content of photoelectric property and higher conversion efficiency; their cluster “Sol gel” is our eighth topic, T08-SolGel, in which we have provide more details about this topic, mesoporous, titania, anatase, and nanocrystalline; our topic, Ruthenium Sensitizers (T05-Rsens) matches with their fifth cluster, which has almost the same content.

We proceed our post topic modeling, mapped the values of topic weight, trend, and citation indices to a scatter plot chart using MATLAB, as shown in Fig. 6. Similar as the previous case study, the *x*-axis represents the trend index, the *y*-axis denotes the weight index, and the size and the color of the dots indicates the citations they received. For visual and comparison convenience, we scaled the indices values of topic-based weight and citation to a range between [0, 1] and scaled the values of topic-based trend index to [−1, 1]. Normalized trend values larger than 0 indicate the topics have a growing trend; on the contrary, values smaller than 0 describe a declining trend.

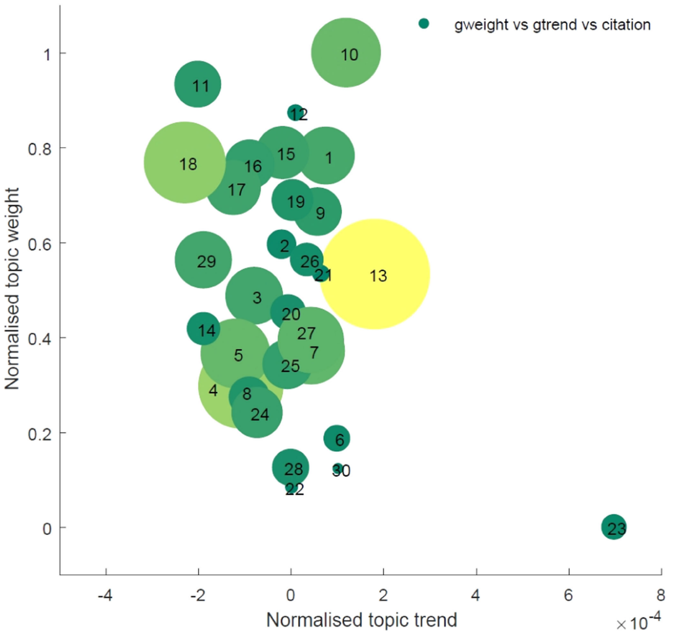


Fig. 6. Topic-based analytic map based on weight, trend, and citation characteristics of DSSC corpus.

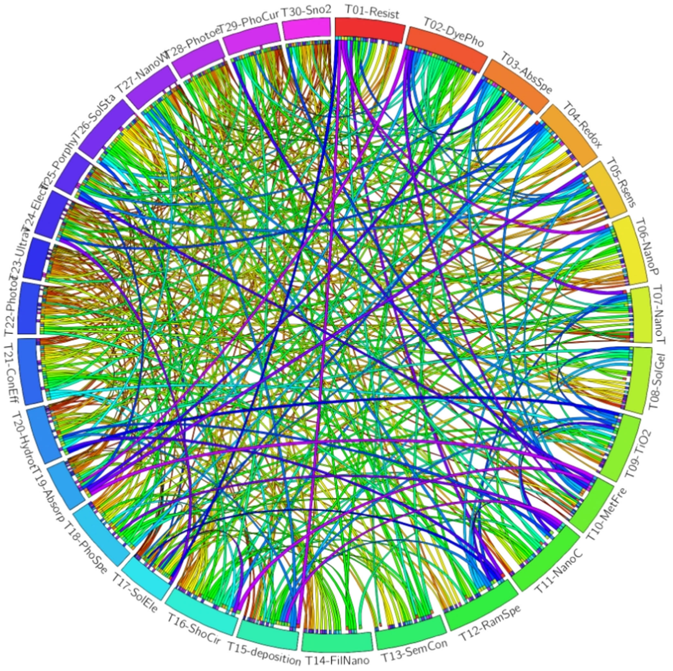


Fig. 7. Co-occurrence map of topics in DSSC research.

In Fig. 6, semiconductor materia (T13-SemCon) is the topic received highest citations, indicating it has largest existing influence and potential usefulness. This topic is also a fast growing one. It is mentioned more in recent year publications than in older papers. In addition, one of our interesting findings is that the upward trend of topic ultraviolet (T23-Ultrav) is growing very fast, although the proportion of this topic is very low. This can be verified using Zhang *et al.*’s research [12], in which cluster “ultraviolet” has the lowest coverage in their clusters. This may suggest that the topic is an emerging and fast developing one. Metal-free sensitizers (T10-MetFre) is one of the hottest topics in the DSSC research, which is highly possible keep being popular since its trend indices is positive. In contrast, although topic of Photocurrent Spectroscopy (T18-PhoSpe) is also a current research strength, it has a dropping trend that indicates it is mentioned less frequently in the recent publications.

3) Topic Relation Identification: After exploring the features of the discovered topics for DSSC research, we quantitatively model the relations among these topics, and build a topic relation network. Fig. 7 visualizes the topic relation matrix via Circos.

TABLE III
MAJOR TOPIC RELATIONS OF DSSC RESEARCH

Source Label	Target Label	MAA	MAR	Affiliation of MAR
T01-Resist	T17-SolEle	Chinese Acad Sci	Lan, Z	Huaqiao Univ
	T04-Redox	Univ London Imperial Coll Sci Technol & Med	Kusama, H	Natl Inst Adv Ind Sci & Technol
	T07-NanoT	Univ Erlangen Nurnberg	Shankar, K	Penn State Univ
	T12-RamSpe	Chinese Acad Sci	Mali, S S	Shivaji Univ
T02-DyePho	T03-AbsSpe	Chinese Acad Sci	Katoh, R	Natl Inst Adv Ind Sci & Technol
	T10-MetFre	E China Univ Sci & Technol	Ooyama, Y	Hiroshima Univ
	T19-Absorp	King Khalid Univ	Irfan, A	King Khalid Univ
T03-AbsSpe	T19-Absorp	King Khalid Univ	Irfan, A	King Khalid Univ
T04-Redox	T17-SolEle	Chinese Acad Sci	Lan, Z	Huaqiao Univ
	T24-Electr	Chinese Acad Sci	Singh, P K	Sogang Univ
T09-TiO2	T20-Hydrot	Chinese Acad Sci	Kang, S H	Seoul Natl Univ
	T08-SolGel	Yonsei Univ	Park, J T	Yonsei Univ
T10-MetFre	T25-Porphy	Kyoto Univ	Giribabu, L	Indian Inst Chem Technol
	T29-PhoCur	Swiss Fed Inst Technol	Matsui, M	Gifu Univ
	T30-Sno2	Hiroshima Univ	Ooyama, Y	Hiroshima Univ
T11-NanoC	T09-TiO2	Chinese Acad Sci	Zhang, Q F	Univ Washington
	T20-Hydrot	Chinese Acad Sci	Kang, S H	Seoul Natl Univ
T12-RamSpe	T17-SolEle	Chinese Acad Sci	Lan, Z	Huaqiao Univ
	T20-Hydrot	Chinese Acad Sci	Kang, S H	Seoul Natl Univ
	T07-NanoT	Univ Erlangen Nurnberg	Shankar, K	Penn State Univ
	T11-NanoC	Chinese Acad Sci	Jin, Y S	Kyungwon Univ
T25-Porphy	T19-Absorp	King Khalid Univ	Irfan, A	King Khalid Univ

Each segment in the figure indicates a topic, named with the short label given in Table II. The ribbons between segments shows their semantic relations. A wider ribbon with a more intense blue represents a stronger relationship between the two linked segments.

As shown in Fig. 7, most of the strongest semantic correlations are between resistance (T01-Resist), dye-sensitized photoelectrochemical cell (T02-DyePho), absorption spectroscopy (T03-AbsSpe), redox regeneration (T04-Redox), titanium dioxide (T09-TiO2), metal-free ssensitizers (T10-MetFre), nanocrystalline (T11-NanoC), Raman spectroscopy (T12-RamSpe), and porphyrin (T25-Porphy). Table III shows the detailed relations between these topics, in which we also add the metadata linkage outcome to present the author and affiliation with the highest rank, for each highly correlated topics. The most active affiliations is denoted by MAA, and the most active researchers is labeled by MAR.

We can consider that publication clusters under two strongly related topics are semantically associated with each other, thus affiliations and researchers of the two clusters share potential collaboration opportunities on the corresponding topics, although these topics may be from disparate subject categories. As shown in the target label column, these topics can be seen as possible cooperation directions for an affiliation or a researcher who is working on DSSC topics and seeking for possible collaborations. For example, for topic metal-free sensitizers (T10-MetFre), the researchers working on DSSC research with the E China University of Science & Technology have potential collaboration opportunities with the researchers in areas of porphyrin (T25-Porphy), photon-to-current (T29-PhoCur), and synthesized Sno2 (T30-SNo_2) which have the representative affiliations as Kyoto University, Swiss Federal Institute of Technology, and Hiroshima University. In addition, the research papers of the most active researchers are also potentially useful for researchers seeking engagement and innovation enlightenment in the joined area. In real cases, factors effecting the cooperation between affiliations and researchers will be far more complicated than just considering research topic connection, but the result of our proposed methodology can be used as a useful support to the decision-making process.

C. Discussion

Theoretically, this paper improves traditional scientific literature clustering models, extends the existing LDA-based approach, and may well lead to a higher degree of heavily geared, cross-disciplinary research. The identified topic relations provide a statistical prospective on discovering semantic level connections. Practically, potential collaboration identification is one of the possible applications of our proposed methodology. In today's increasingly competitive and collaborative research environment, the ability to choose collaborators wisely is a key component of success. However, organizational and discipline boundaries can make this process particularly difficult. The proposed method attempts to provide suggestions for research collaboration at a topic-based level, especially for early-career researchers, e.g., to answer the questions like, "what topics associate with my research but previously I don't know them," or "Who might collaborate with me within my university if I want to do the related interdisciplinary research?" or "Any recommendation for possible partners for a funding application on a particular topic?" We are able to identify different groups of researchers and affiliations that are working on closely related topics internally and externally, to gain possible collaboration opportunity cross-the-board for further research or funding application.

Not only for academia this method can assist decision making, for industry users and governmental decision makers, it also can be used as an efficient tool to discover topical relations and the full picture of the target area, to reveal the possible direction and the insight of development and to recognise potential collaborators and competitors.

V. Conclusion

Since its introduction to the field of bibliometrics, LDA has been used as an efficient tool for key content extraction and thematic analysis. It will continue to be emphasized because of its ability to estimate the possibilities for the "co-occurrence" of topics, and provide semantically meaningful outcomes. These advances, however, also intensify the need for methodologies that provide productive and useful post-LDA analysis, and fully utilizing the comparatively complex products of topic modeling may create bottlenecks for the enhancement and expansion of statistical modeling in bibliometrics research.

This paper outlined a complete process for analyzing technological landscapes, including pre-topic modeling preparation and post-clustering analysis. The comprehensive overview it generated assisted researchers in identifying focal themes and the relations among those themes when using LDA to explain a scientific literature corpus. In addition, the topic-based analytic maps and topic-based relational network proposed in this paper delivered an informative presentation that quantitatively measured the discovered topics allowing indepth analysis. In the case studies, we discussed one of the possible applications of topic-based analysis method using big data and DSSC publication datasets. By using the topic relation identification methodology, we could identify potential collaboration opportunities internally and externally. Although in real cases, factors effecting the cooperation between affiliations and researchers were more complicated than just considering research topic connection, the result of our proposed methodology could serve as a support to the decision-making process.

During the experiments, we found that how to trace temporal topic changing and evolving automatically is one of the main tasks in exploring more application of our existing topic-based analysis methodology. Topic matching and topic variegation tracking will be addressed in our future research.

References

[1] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, "Big data: The management revolution," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.

[2] A. L. Porter and S. W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage*. Hoboken, NJ, USA: Wiley, 2004.

[3] I. Ketata, W. Sofka, and C. Grimpe, "The role of internal capabilities and firms' environment for sustainable innovation: Evidence for Germany," *R&D Manage.*, vol. 45, no. 1, pp. 60–75, 2015.

[4] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2018.2854000.

[5] S. W. Cunningham, A. L. Porter, and N. C. Newman, "Special issue on tech mining," *Technol. Forecasting Social Change*, vol. 8, no. 73, pp. 915–922, 2006.

[6] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Inf. Process. Manage.*, vol. 43, no. 5, pp. 1216–1247, 2007.

[7] R. N. Kostoff, D. R. Toothman, H. J. Eberhart, and J. A. Humenik, "Text mining using database tomography and bibliometrics: A review," *Technol. Forecasting Social Change*, vol. 68, no. 3, pp. 223–253, 2001.

[8] C. K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, 2014.

[9] K. W. Boyack *et al.*, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PloS One*, vol. 6, no. 3, 2011, Art. no. e18029.

[10] B. Yoon and Y. Park, "A systematic approach for identifying technology opportunities: Keyword-based morphology analysis," *Technol. Forecasting Social Change*, vol. 72, no. 2, pp. 145–160, 2005.

[11] G. Cascini and D. Russo, "Computer-aided analysis of patents and search for TRIZ contradictions," *Int. J. Prod. Develop.*, vol. 4, no. 1–2, pp. 52–67, 2006.

[12] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, "Term clumping for technical intelligence: A case study on dye-sensitized solar cells," *Technol. Forecasting Social Change*, vol. 85, pp. 26–39, 2014.

[13] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Scientific evolutionary pathways: Identifying and visualizing relationships for scientific topics," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 8, pp. 1925–1939, 2017.

[14] Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, and J. Lu, "Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research," *Technol. Forecasting Social Change*, vol. 105, pp. 179–191, 2016.

[15] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, pp. 45–59, 2016.

[16] J. M. Cotelo, F. L. Cruz, F. Enríquez, and J. Troyano, "Tweet categorization by combining content and structural knowledge," *Inf. Fusion*, vol. 31, pp. 54–64, 2016.

[17] C. De Maio, G. Fenza, V. Loia, and M. Parente, "Time aware knowledge extraction for microblog summarization on twitter," *Inf. Fusion*, vol. 28, pp. 60–74, 2016.

[18] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 178–185.

[19] Y. Ding, "Topic-based pagerank on author cocitation networks," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 3, pp. 449–466, 2011.

[20] H. Chen, G. Zhang, D. Zhu, and J. Lu, "Topic-based technological forecasting based on patent data: A case study of Australian patents from 2000 to 2014," *Technol. Forecasting Social Change*, vol. 119, pp. 39–52, 2017.

[21] A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human assigned subject classification," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 10, pp. 2464–2476, 2016.

[22] F. De Battisti, A. Ferrara, and S. Salini, "A decade of research in statistics: A topic model approach," *Scientometrics*, vol. 103, no. 2, pp. 413–433, 2015.

[23] L. Waltman and N. J. V. Eck, "A systematic empirical comparison of different approaches for normalizing citation impact indicators," *J. Informetrics*, vol. 7, no. 4, pp. 833–849, 2013.

[24] K. W. Boyack and R. Klavans, "Including cited nonsource items in a large-scale map of science: What difference does it make?" *J. Informetrics*, vol. 8, no. 3, pp. 569–580, 2014.

[25] E. P. Jiang, "Content-based spam email classification using machine-learning algorithms," in *Text Mining: Applications and Theory*. Hoboken, NJ, USA: Wiley, 2010, pp. 37–56.

[26] E. Yan and Y. Zhu, "Identifying entities from scientific publications: A comparison of vocabulary- and model-based methods," *J. Informetrics*, vol. 9, no. 3, pp. 455–465, 2015.

[27] M. D. Cao and X. Gao, "Combining contents and citations for scientific document classification," in *Proc. Australas. Joint Conf. Artif. Intell.*, 2005, pp. 143–152.

[28] P. Ahlgren and C. Colliander, "Document–document similarity approaches and science mapping: Experimental comparison of five approaches," *J. Informetrics*, vol. 3, no. 1, pp. 49–63, 2009.

[29] J. Lin and W. J. Wilbur, "Pubmed related articles: A probabilistic topic-based model for content similarity," *BMC Bioinf.*, vol. 8, no. 1, 2007, Art. no. 423.

[30] K. W. Boyack and R. Klavans, "Cocitation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?" *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2389–2404, 2010.

[31] P. Glenisson, W. Glänzel, F. Janssens, and B. De Moor, "Combining full text and bibliometric information in mapping scientific disciplines," *Inf. Process. Manage.*, vol. 41, no. 6, pp. 1548–1572, 2005.

[32] L. Waltman, N. J. V. Eck, and E. C. M. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *J. Informetrics*, vol. 4, no. 4, pp. 629–635, 2010.

[33] R. Phaal, C. J. Farrukh, and D. R. Probert, "Technology roadmapping—A planning framework for evolution and revolution," *Technol. Forecasting Social Change*, vol. 71, no. 1–2, pp. 5–26, 2004.

[34] R. C. Basole, H. Park, and R. O. Chao, "Visual analysis of venture similarity in entrepreneurial ecosystems," *IEEE Trans. Eng. Manag.*, to be published, doi: 10.1109/TEM.2018.2855435.

[35] R. Klavans and K. W. Boyack, "Research portfolio analysis and topic prominence," *J. Informetrics*, vol. 11, no. 4, pp. 1158–1174, 2017.

[36] K. W. Boyack, R. Klavans, H. Small, and L. Ungar, "Characterizing the emergence of two nanotechnology topics using a contemporaneous global micromodel of science," *J. Eng. Technol. Manage.*, vol. 32, no. 32, pp. 147–159, 2014.

[37] H. Chen, Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Modeling technological topic changes in patent claims," in *Proc. Portland Int. Conf. Manage. Eng. Technol.*, 2015, pp. 2049–2059.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[39] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[40] S. Liu and C. Chen, "The differences between latent topics in abstracts and citation contexts of citing papers," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 3, pp. 627–639, 2013.

[41] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci.*, vol. 101, pp. 5228–5235, 2004.

[42] L. Yang *et al.*, "CQArank: Jointly model topics and expertise in community question answering," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 99–108.

[43] D. Kim and A. Oh, "Topic chains for understanding a news corpus," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, 2011, pp. 163–176.

[44] D. A. Broniatowski and C. L. Magee, "The emergence and collapse of knowledge boundaries," *IEEE Trans. Eng. Manage.*, vol. 64, no. 3, pp. 337–350, Aug. 2017.

[45] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook Latent Semantic Anal.*, vol. 427, no. 7, pp. 424–440, 2007.

[46] H. Gregor, "Parameter estimation for text analysis," Fraunhofer IGD, Tech. Rep. Version 2.9, 2009.

[47] A. H. Huang, R. Lehavy, A. Y. Zang, and R. Zheng, "Analyst information discovery and interpretation roles: A topic modeling approach," *Manage. Sci.*, vol. 64, no. 6, pp. 2833–2855, 2018.

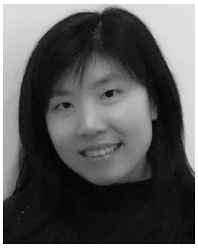
[48] F. Radicchi and C. Castellano, "Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts," *J. Informetrics*, vol. 6, no. 1, pp. 121–130, 2012.

[49] Y. Zhang, Y. Huang, A. L. Porter, G. Zhang, and J. Lu, "Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study," *Technol. Forecasting Social Change*, to be published, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162517315081

[50] M. I. Krzywinski *et al.*, "Circos: An information aesthetic for comparative genomics," *Genome Res.*, vol. 19, pp. 1639–1645, 2009.

[51] Y. Huang, A. L. Porter, Y. Zhang, X. Lian, and Y. Guo, "An assessment of technology forecasting: Revisiting earlier analyses on dye-sensitized solar cells (dsscs)," *Technol. Forecasting Social Change*, to be published, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162517318541

[52] Y. Guo, T. Ma, A. L. Porter, and L. Huang, "Text mining of information resources to inform forecasting innovation pathways," *Technol. Anal. Strategic Manage.*, vol. 24, no. 8, pp. 843–861, 2012.



Hongshu Chen received the Ph.D. degree in management science and engineering from the Beijing Institute of Technology, Beijing, China, in 2015, and the second Ph.D. degree in software engineering from the University of Technology Sydney, Ultimo, NSW, Australia, in 2016.

She is currently an Assistant Professor with the School of Management and Economics, Beijing Institute of Technology. Her research interests include bibliometrics, information systems, and text analytics.

Ximeng Wang received the B.E. degree in communication engineering from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2011, and the M.E. degree in software engineering in 2013 from Beijing Jiaotong University, Beijing, China, where he is currently working toward the Ph.D. degree.

Since 2017, he has been a Joint Ph.D. Student with the University of Technology Sydney, Ultimo, NSW, Australia. His current research interests include recommender systems, complex networks, and data mining.

Shirui Pan (M'16) received the Ph.D. degree in computer science from the University of Technology Sydney, Ultimo, NSW, Australia.

He is currently a Lecturer with the Faculty of Information Technology, Monash University, Melbourne, VIC, Australia. Prior to that, he was a Research Fellow and then a Lecturer with the School of Software, University of Technology Sydney. His research interests include data mining and machine learning.

Dr. Pan has authored/co-authored more than 50 research papers in top-tier journals and conferences, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON CYBERNETICS, IEEE International Conference on Data Engineering, Association for the Advancement of Artificial Intelligence, International Joint Conference on Artificial Intelligence, and IEEE International Conference on Data Mining.

Fei Xiong received the B.E. and Ph.D. degrees in communication and information systems from Beijing Jiaotong University, Beijing, China, in 2007 and 2013, respectively.

He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. From 2011 to 2012, he was a Visiting Scholar with Carnegie Mellon University. He has authored/co-authored more than 60 papers in refereed journals and conference proceedings. He was a recipient of National Natural Science Foundations of China and several other research grants. His current research interests include the areas of web mining, complex networks, and complex systems.