The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/0737-8831.htm

Received 30 June 2017
Revised 20 November 2017
20 November 2017
29 December 2017
Accepted 29 December 2017

Discovering research topics from library electronic references using latent Dirichlet allocation

Debin Fang, Haixia Yang and Baojun Gao
Economics and Management School, Wuhan University, Wuhan, China, and
Xiaojun Li
College of Accounting, Yunnan University of Finance and Economics, Kunming, China

Abstract

Purpose – Discovering the research topics and trends from a large quantity of library electronic references is essential for scientific research. Current research of this kind mainly depends on human justification. The purpose of this paper is to demonstrate how to identify research topics and evolution in trends from library electronic references efficiently and effectively by employing automatic text analysis algorithms.

Design/methodology/approach – The authors used the latent Dirichlet allocation (LDA), a probabilistic generative topic model to extract the latent topic from the large quantity of research abstracts. Then, the authors conducted a regression analysis on the document-topic distributions generated by LDA to identify hot and cold topics.

Findings – First, this paper discovers 32 significant research topics from the abstracts of 3,737 articles published in the six top accounting journals during the period of 1992-2014. Second, based on the document-topic distributions generated by LDA, the authors identified seven hot topics and six cold topics from the 32 topics.

Originality/value – The topics discovered by LDA are highly consistent with the topics identified by human experts, indicating the validity and effectiveness of the methodology. Therefore, this paper provides novel knowledge to the accounting literature and demonstrates a methodology and process for topic discovery with lower cost and higher efficiency than the current methods.

Keywords Academic libraries, Big data, Accounting research, Latent Dirichlet allocation (LDA), Topic model, Topic trends

Paper type Research paper

1. Introduction

In the era of big data, the increasing availability of electronic libraries helps scholars obtain literature easier. The huge volume of electronic references, however, presents scholars with the dilemma of information overload. Facing the large quantity of electronic references recommended by the retrieval systems based on the title, abstract, or key words of scientific works, scholars have to spend so much time determining the most related works. It is also difficult for scholars entering a new research field to identify the trends in research and the hot topics.

Many works have been devoted to discovering the research topics and trends of a specific field. For example, Cushing (1989), Dunbar and Weber (2014), and Farcas (2015) are all studies of this kind in the field of accounting. Historically, this kind of task mainly depends on human justification, a process that is both time-consuming and labor-intensive. As such, employing automatic text analysis algorithms and tools to identify the research topics and their trends more efficiently is necessary.

Latent Dirichlet allocation (LDA) (Blei *et al.*, 2003; Blei, 2012; Griffiths and Steyvers, 2004) is a kind of probabilistic topic model aiming to discover the latent topics from text corpus.

Conflicts of interest: the authors declare no conflict of interest.

The authors would like to thank all the supports from the National Natural Science Foundation Programs of China (NSFC) (71771182, 71673210, 71725007).

Library Hi Tech
Vol. 36 No. 3, 2018
pp. 400-410
© Emerald Publishing Limited
0737-8831
DOI 10.1108/LHT-06-2017-0132

The basic idea of LDA is that a document is a distribution of a topic, and a topic is a distribution of terms. Using LDA, this paper first determines 32 significant research topics from the abstracts of 3,737 articles in six top accounting journals and then identifies seven hot and six cold topics (Griffiths and Steyvers, 2004) in the period from 1992 to 2014 based on the evolution analysis.

The contributions of this paper are twofold. First, this paper contributes to the bibliometric literature by indicating how big data analytics for unstructured electronic library text helps lower costs and raise efficiency for bibliometric analysis. Second, this paper also provides novel knowledge to the accounting literature by discovering the research topics and identifying the hot and cold topics. The topics discovered by LDA are highly consistent with the topics identified by human experts (Dunbar and Weber, 2014), indicating the validity and effectiveness of our methodology.

The rest of the paper is organized as follows. Section 2 is devoted to the literature review. Section 3 shows the data and methodology, including the research data, data processing, and the theory of LDA. Section 4 presents the results for topic discovery and the evolution of trends of topics. Section 5 discusses the results of the paper and presents its implication. Section 6 concludes the paper.

2. Literature review

Investigating the topics and evolution of a research field is important. To improve the efficiency of this task, researchers of bibliometrics adopt some computer-based technology, such as the VOSviewer clustering technique for overlaying communities and topics (Yan *et al.*, 2012), co-word analysis for library and information science (Ding *et al.*, 2000; Franklin and Jaeger, 2007; Liu *et al.*, 2012; Zong *et al.*, 2013), co-word analysis for biology (Rip and Courtial, 1984; Cambrosio *et al.*, 1993; An and Wu, 2011), co-word analysis for education (Ritzhaupt *et al.*, 2010), and co-citation analysis for PHD theses at the Wuhan University (Gao *et al.*, 2009).

However, the methods mentioned above are supervised, which still require much human labeling for the provision of training data set. Therefore, Hofmann (1999) put forward a topic model, called probabilistic latent semantic indexing (PLSI), to extract information from textual data. Blei *et al.* (2003) put forward a topic model called LDA, which is intended to prevent over-fitting of PLSI. The basic idea of LDA is that each topic is a mixture of topics and each topic is a mixture of words. LDA has been applied in many fields, including pattern detection in images (Agarwal and Triggs, 2008; Coelho *et al.* 2010) and video (Niebles *et al.*, 2008; Wang *et al.*, 2009), automatic essay grading (Kakkonen *et al.*, 2008), and fraud detection (Xing and Girolami, 2007). For different research needs, scholars have performed some expansions to the LDA model, such as the author-topic model (Hwang *et al.*, 2017), dynamic topic models (Blei and Lafferty, 2006), correlated topic models (Blei and Lafferty, 2007), non-Markov continuous-time model (Wang and Mccallum, 2006), and sentence LDA (Bao and Datta, 2014; Jo and Oh, 2011).

Griffiths and Steyvers (2004) analyzed the PNAS data set with LDA to discover topics and to detect hot topics and cold topics. After that, more and more researchers have employed LDA or its expansions to analyze the topics of one field or one journal to understand the evolution of the field or the journal. Table I presents some papers that employ LDA or its expansions to analyze the topics of fields or journals. Table I shows that many researchers from kinds of fields analyzed the topics or evolution based on LDA. The results of their papers show that topic models outperform more traditional, cluster-based approaches in information retrieval. In their studies, EM (Blei and Lafferty, 2007) and the Gibbs sampler (Griffiths and Steyvers, 2004) are often used to estimate the latent variable. Most of these studies use the abstract or the combination of abstract and the title as the corpus.

Table I.
Bibliometrics-based topic model

Literature	Field or journal	Number of samples	Corpus	Estimation method	Model
Griffiths and Steyvers (2004)	*PNAS* papers	28,154	Abstract	Gibbs sampling	LDA
Wang and Mccallum (2006)	NIPS conference papers	2,326	Full article	Gibbs sampling	TOT
Hall *et al.* (2008)	The field of computational linguistics	12,500	Full article	Gibbs sampling	LDA
Wu *et al.* (2010)	The field of bioinformatics	5,639	Abstracts and articles	Gibbs sampling	LDA
Sugimoto *et al.* (2011)	Library and information science	3,121	Title and abstract	Gibbs sampling	LDA
Piepenbrink and Nurmammadov (2015)	Economics or business (600 journals)	5,861	Abstract	Gibbs sampling	LDA
Cohen and Austerweil (2015)	Cognitive science	2,512	Title and abstract	Gibbs sampling	LSA
Yau *et al.* (2014)	Seven different research fields	1,254	Title and abstract	Gibbs sampling	LDA; CTM; HLDA; HDP
Blei and Lafferty (2007)	Science	16,351	Full article	EM	CTM
Sun and Yin (2017)	Transportation journals	17,163	Abstract	Gibbs sampling	LDA
Our paper	Six top accounting journals	3,737	Title and abstract	Gibbs sampling	LDA

In this paper, the authors combined the title and abstract of the same paper from six top accounting journals into a document (corpus). Then, the authors employed LDA to discover topics and analyze the evolution of accounting field. Following Griffiths and Steyvers (2004), the latent variable was estimated using the Gibbs sampler.

3. Methodology

In this section, we first describe the data used in our research as well as our preprocessing approach. Then, we provide an overview of LDA. Table II contains the description of the terms and symbols used in this paper.

3.1 Data source and data preprocessing

In March 2015, we retrieved entries from Thomson Scientific's Web of Science database. We restricted the search to the following journals: *Accounting Review*, *Journal of Accounting Research*, *Journal of Accounting Economics*, *Contemporary Accounting Research*, *Accounting, Organization, and Society*, and *Review of Accounting Studies*. We limited the years of publication to from 1992 to 2014. We saved title (TI), key words (DE), Abstract (AB), journal (SO) and publication year (PY) for later analysis. We obtained 4,720 entries.

Due to our focus on topics of research papers, we limited the types of publications to "articles" or "proceedings papers," which reduced the database to 3,873 entries. As we focused on the publications that can be read by the international community, we concentrated on publications in the original language of English (3,861 publications). Finally, we excluded any papers without abstracts. Thereafter, we obtained 3,737 documents with complete information for the analysis.

With the help of package "tm" (Feinerer *et al.*, 2008) in *R*, we processed the 3,737 documents in the following way:

(1) Step 1: delete the copyright information of publishers at the end abstract (Piepenbrink and Nurmammadov, 2015).

Term or symbol	Description
corpus	The set of all documents
document	A single combination of the title and abstract
vocabulary	The unique words of the entire document
topic	The set of terms
term	Words in the vocabulary
D	The total number of documents (combination of title and abstract)
d	Auxiliary index over documents
N	The number of words
n	Auxiliary index over word
K	The number of topics
j	Auxiliary index over topics
$\boldsymbol{\Phi}$	Multinomial distribution over words
θ	Multinomial distribution over topics
α	Parameter of topic Dirichlet prior
β	Parameter of word Dirichlet prior
$w_{d,n}$	The word n in document d
z_d	The topic assignments for document d
$\boldsymbol{z}_{-d}$	The topic assignments for all documents except document d
$n_{-d}^{(.)}$	A count that does not include the current assignment of zd
$n_j^{(w)}$	A count that word w has been assigned to topic j in the vector assignments z
$C_{d,v}$	Indicates how often the vth term occurs in the dth document

Table II. The description of the terms or the symbols used in this paper

(2) Step 2: combine the title and the abstracts as a document for each paper (Cohen and Austerweil, 2015; Sugimoto *et al.*, 2011; Yau *et al.*, 2014).

(3) Step 3: turn the documents into a plain text document.

(4) Step 4: change the text to lower case and remove all numbers and punctuation characters.

(5) Step 5: change different representations of an accounting term to the same form. For example, change "earnings per share" to "eps," change "generally accepted accounting principles" to "gaap," and so on (Dunning, 1993; Manning and Schütze, 1999).

(6) Step 6: remove English stop words (e.g. *the, and, when, is, at, which, on,* and *in*), which contain little topical content.

(7) Step 7: remove a short list of other words that are not related to topical content (e.g. paper, show, presents, address problem, consider result, find, and note).

(8) Step 8: stem the words.

(9) Step 9: create a document-term matrix (DTM) and remove the words that do not occur in at least three abstracts, which decreases the sparsity of matrix.

After the above steps, a DTM with 3,737 rows and 3,690 columns was created. The rows in this matrix corresponded to the documents and the columns to the terms. The entry $C_{d,v}$ indicates how often the vth term occurred in the dth document. The number of rows is equal to the size of the corpus (the number of documents: 3,737), and the number of columns is equal to the size of the vocabulary (the number of unique words from the corpus: 3,690). Therefore, our DTM consists of 3,737 documents and 3,690 unique words.

3.2 LDA

The basic idea behind LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Each word is assigned to a topic with a certain probability, and words are chosen from the corresponding topic. For example, the paper "Are analysts' earnings forecasts more accurate when accompanied by cash flow forecasts" has the topics "cash flow," "forecast analyst" and "earning management" with high probability; the topic "cash flow" has words about "accrual," "cash," and "flow" with high probability, and the topic "earning management" has words about "earning," "management," and "GAAP" with high probability. The process of generating a corpus is described as follows:

(1) For each topic $j \in \{1, \ldots, K\}$:
   - draw a distribution over vocabulary words Φ_j~*Dirichlet* (β).

(2) For each document $d \in \{1, \ldots, D\}$ in the corpus:
   - draw a vector of topic proportions:

     $\theta_d \sim$ *Dirichlet* (α).
   - For each word $w_{d,n}$ in document d,
     - draw a topic assignment:

       $\boldsymbol{z}_{d,n} \sim$ *Multinomial* (θ_d);
     - draw a word $w_{d,n}$~Multinomial $(\Phi_{z_{d,n}})$.

We use Gibbs sampling (Griffiths and Steyvers, 2004) to estimate the latent variables θ_d and Φ_j. At each transition step of the Markov chain, the aspect of the d_{th} document, z_d, is drawn from the conditional probability:

$$\mathrm{P}\left(z_d \middle| \boldsymbol{z}_{-d}, w_d\right) \propto \frac{n_{-d,j}^{(w_d)} + \beta}{n_{-d,j}^{(\cdot)} + V\beta} \frac{n_{-d,j}^{(d)} + \alpha}{n_{-d,\cdot}^{(d)} + K\alpha} \tag{1}$$

The result is quite intuitive: the first ratio expresses the probability of w_d under topic j, and the second ratio expresses the probability of topic j in document d. For any single sample, the approximate probability of word w in topic k is:

$$\Phi_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{\cdot} + V\beta} \tag{2}$$

The approximate probability of topic k in paper d is:

$$\theta_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_{\cdot}^{d} + K\alpha} \tag{3}$$

4. Results

In this part, we first describe the setting of LDA and then discover the accounting topics and topic trends with LDA. We perform all data processing and modeling with the statistical language *R*. We use the package "tm" (Feinerer *et al.*, 2008) for the creation and preprocessing of the corpus and the package "topicmodels" (Hornik and Grün, 2011) for the generation and evaluation of the topic models.

4.1 Specification of LDA

To discover the topics using LDA, the first step is to determine the optimal topic number for these 3,737 documents. Specifically, given the parameter of topic Dirichlet prior α for document-topic distribution and the parameter of word Dirichlet prior β for topic-word distribution, we should choose the appropriate value of specified number of topics K. So far there are several methods for obtaining the optimal number of topics, including perplexity (Blei *et al.*, 2003), empirical likelihood (Li and McCallum, 2006), hierarchical Dirichlet processes (Teh *et al.*, 2006), and marginal likelihood (Newton and Raftery, 1994; Griffiths and Steyvers, 2004; Wallach, 2006) and so on. In this paper, we use the marginal likelihood method, which is approximated by the harmonic mean method (Griffiths and Steyvers, 2004).

To find an appropriate K, we first set the fixed hyper-parameters, the topic distribution Dirichlet parameter $\alpha = 50/K$, and the term distribution of the topics $\beta = 0.1$ according to Griffiths and Steyvers (2004). Then, we change the number of topics K from 2 to 100 topics and calculate the marginal likelihood of each case. We find the number 32 best accounts for our corpus because it maximizes the likelihood of this setting. We need to note that the value of K found through this procedure depends on the setting of parameters α and β, and it is also affected by the corpus and data preprocessing process (Griffiths and Steyvers, 2004).

Once the specified number of topics $K = 32$ is obtained, we run the LDA model by using package "topicmodels" (Hornik and Grün, 2011) in R. Then, we can get the document-topic distribution and the topic-term distribution. Table III shows a sample of the topic distributions of documents. The value for $\theta_{i,\,j}$ indicates the probability of topic j in document i.

4.2 Results for topics distribution and their trends

In this section, we first analyze the research topics from the abstracts and titles of six top accounting journals. Then, we identify the hot and cold topics according to the evolution pattern of document-topic distributions.

As in Table IV, for inspecting the content of each topic, we extract the top ten terms for each topic. We then label each topic by assigning a meaningful name according to these terms and the content of each document. For example, topic 1 is named "audit" because its top ten stemmed terms are "audit," "auditor," "client," "fee," "service," "independ," "fraud," "big," "quality," and "partner." Table V presents the top ten stemmed words for the 32 topics and the label of each topic. From the table, we can see the topics are quite cohesive overall with little intrusion of words within the top ten stemmed terms.

Dunbar and Weber (2014) compile and analyze the reference lists from papers published in nine accounting journals (i.e. *AOS*, *AJPT*, *CAR*, *JAE*, *JAR*, *JATA*, *JMAR*, *RAST*, and *AR*) over the period 1996-2011. They divided the content of the accounting field into nine topics, including five small research disciplines (i.e. auditing, financial reporting, managerial

$\theta_{i,j}$	1	2	3	4	…	29	30	31	32
1	0.1446	0.0677	0.0100	0.0228	…	0.0164	0.0100	0.0485	0.0100
2	0.0259	0.0146	0.0089	0.0089	…	0.0089	0.0146	0.0089	0.0089
3	0.0476	0.0098	0.0098	0.0161	…	0.0098	0.0098	0.0664	0.0098
…	…	…	…	…	…	…	…	…	…
3,724	0.0132	0.0980	0.0132	0.0132	…	0.0132	0.0132	0.0302	0.0132
3,735	0.0153	0.0153	0.0349	0.0153	…	0.0741	0.0153	0.0153	0.0349
3,736	0.0119	0.1951	0.0119	0.0119	…	0.0119	0.0119	0.0119	0.0196
3,737	0.1914	0.0161	0.0367	0.0264	…	0.0264	0.0161	0.0264	0.0264

Table III. The topic distributions of documents

Table IV.
Top ten stemmed words for each topic

No.	Topic	Top ten stemmed words
1	Audit	Audit, auditor, client, fee, service, independ, fraud, big, quality, partner
2	Information and signal	Inform, privat, public, signal, uncertainty, asymmetri, precis, content, environ, equilibrium
3	Performance measure	Perform, measure, evalu, target, base, weight, peer, common, nonfinanci, indic
4	Social practice	Account, economy, practice, histor histori, argu, economi, chang, society, natur
5	Products	Cost, profit, product, activ, competit, effici, base, transfer, alloc, custom
6	Bank	Loss, bank, secur, liabil, loan, rule, legal, gain, litig, deriv
7	Risk	Risk, level, industri, factor, assess, specif, busi, plan, premium, rank
8	Debt	Debt, rate, conservat, firm, interest, credit, coven, bond, financ, conserv
9	Asset valuation	Asset, valuat, income, fair, relev, adjust, book, valu, ratio, differ
10	Model	Model, estim, predict, test, time, empir, section, variabl, cross, specif
11	Profession practice	Social, practice, profession state process polit develop profess culture institut
12	Empirical study	Posit, negat, respons, associ, signific, prior, support, hypothesi, hypothes, initi
13	Announcement	Announc, period, adopt, reaction, post, year, restat, ifr, chang, sec
14	Accounting standards	Standard, statement, finance, sfas, recognit, expens, require, reliabl, recogn, report
15	Capital	Capit, firm, invest, equity, cost, growth, long, dividend, short, real
16	Earning management	Earn, manag, quarter, surpris, gaap, meet, annual, benchmark, year, expect
17	Management experience	Knowledge, process, task, structur, group, experi, complex, function, individu, team
18	Tax	Tax, firm, income, rate, corpor, avoid, shift, acquisit, investing, inventori
19	Intern-extern control	Firm, control, intern, country, extern, sox, cross, list, weak, foreign
20	Investor	Investor, trade, institute, ownership, larg, insid, transact, share, small, fund
21	Firm compensation	Firm, compens, ceo, option, execut, board, sharehold, esop, director, pay
22	Corporate govern	Govern, corpor, regul, economy, regulatori, form, public, mechan, law, polici
23	Management system	Manag, system, control, strategi, budget, strategy, organ, organiz , design, negoti
24	Stock market	Market, price, return, stock, expect, future, fundament, volatile, reflect, aggreg
25	Cash flow	Accrual, cash, future, flow, oper, compon, persist, discretionary, item, revenu
26	Decision (behavior)	Decis, influenc, particip, experi, judgment, make, behavior, experiment, materi, condit
27	Incentive management	Incent, manag, contract, manageri, effort, employe, agent, optim, agenc, monitor
28	Experimental	Base, choic, sampl, method, differ, number, test, data, select, altern
29	Forecast analyst	Forecast, analyst, bias, error, accuracy, recommend, guidance, revis, accur, abil
30	Report quality	Report, finance, quality, company, improve, reput, greater, transpar, frequenc, enhanc
31	Theoretical	Theory, interpret, approach, develop, analysi, question, contribut, perspect, framework, argu
32	Disclosure	Disclosur, firm, news, voluntary, disclos, market, environment, ipo, bad, good

accounting, tax, and the other) and four methodologies (i.e. archival, experimental, theoretical, and other). The topics discovered by LDA are highly consistent with the topics identified by human experts (Dunbar and Weber, 2014), indicating the validity and effectiveness of our methodology. Furthermore, the LDA model is able to discover fine-grained topics and thus draw a quite detailed picture of the research in the accounting field.

After discovering the research topics using LDA, we then further identify the hot and cold topics by investigating whether the mean assignment to articles (i.e. document-topic distributions) is increasing or decreasing over time (Griffiths and Steyvers, 2004). By doing so, we build an AR (1) (autoregressive) model (Gao *et al.*, 2018) by regressing the average topic assignments of each topic in year t on its own previous values, as shown in the following equation:

$$\theta_{j,t} = a_j + b_j \theta_{j,t-1} + \varepsilon_j \tag{4}$$

Table V. Slope of the percentage trend of each of the 32 topics for all articles from 1992 to 2014

No.	Topic	Hot ($p = 0.001$)	Cold ($p = 0.001$)	Balanced ($p = 0.001$)
15	Capital	0.032***		
19	Intern-extern control	0.060***		
20	Investor	0.041***		
21	Firm compensation	0.056***		
29	Forecast analyst	0.073***		
30	Report quality	0.052***		
32	Disclosure	0.046***		
4	Social practice		−0.100***	
5	Products		−0.040***	
10	Model		−0.054***	
17	Management experience		−0.069***	
26	Decision (behavior)		−0.047***	
28	Experimental		−0.039***	
1	Audit			0.028**
2	Information and signal			0.016
3	Performance measure			0.005
6	Bank			−0.021
7	Risk			0.012.
8	Debt			0.035*
9	Asset valuation			−0.017.
11	Profession practice			−0.067**
12	Empirical study			−0.005
13	Announcement			0.040**
14	Accounting standards			−0.003
16	Earning management			0.040**
18	Tax			−0.005
22	Corporate govern			0.017*
23	Management system			−0.057**
24	Stock market			0.004
25	Cash flow			0.023
27	Incentive management			0.003
31	Theoretical			−0.058**

Notes: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

We are interested in the value of the coefficient b_j. A positive (negative) value indicates its proportion in the corpus is increasing (decreasing) over time, which identifies a topic as "hot" ("cold"). The results are presented in Table V. From Table V, we can find that 13 of the 32 topics have a coefficient significantly different from 0 at the 0.001 significance level, among which the coefficients of seven topics are positive and six topics are negative. We define the topics with positive coefficient significance at the 0.001 significance level as hot topics, and topics with negative coefficient significance at the 0.001 significance level as cold topics. Therefore, we identified seven hot topics (i.e. topic 15, 19, 20, 21, 29, 30 and 32) and six cold topics (i.e. topic 4, 5, 10, 17, 26, 28). We need to note that the cold topics were not topics that lacked prevalence in the corpus but the topics that presented a strong decrease in popularity over time.

5. Discussion and implications

In the big data era, the great volume of electronic references provides more accessible information to the scholars in each field. On the other hand, however, the big electronic reference data also put the scholar into a dilemma, that is, they have to spend more time on identifying the research topics and their evolution trends. Therefore, how to understand the research topics and their evolution trends is a critical challenge.

Cultivating the big data of electronic library references and the advances of machine learning techniques, this study identified the research topics of the six top accounting journals and their evolution trends using LDA. This work helps researchers grasp the current research situation and the future trends of accounting field, thus providing them a guideline of which areas they should focus on. For example, the topic "inter-extern control(19)" has got more attention in the past years. Therefore, they should pay more attentions to these increasing topics.

LDA is an unsupervised machine learning algorithm that does not need any prior knowledge. The unsupervised nature of LDA means that it is very suitable to extract insightful knowledge from data, without the intervention of human. This paper demonstrates the flows and steps of using LDA for bibliometric analysis, thus providing a map and reference for the topic analysis of research topics of fields other than accounting. The topics identified in this paper are highly consisted with the results of Dunbar and Weber (2014), indicating the validity of effectiveness of the application of LDA for bibliometric research when meeting upon big text data.

6. Conclusion and future work

Cultivating the boom of big data technology and taking the accounting references as examples, we employ an LDA model to discover research topics and the evolution of topic patterns from electronic library references. The number of topics identified in our model is 32, which is larger than what is found in previous research (Dunbar and Weber, 2014). Therefore, this research draws a fine-grained picture for the accounting literature. Based on the document-topic assignment matrix, we obtain seven hot topics and six cold topics at a significance level of 0.001. Our results can help young researchers conduct enhanced topic-based literature searches and understand the topic trends in the accounting field.

However, there are also some limitations in our work. First, the LDA model assumes that documents are independent of each other, but some papers have strong relationships with each other (previous studies can have a big effect on later studies). In the LDA model, the positions of individual words are neglected when inferring the topic. As discussed in Wallach (2006), this property may not always be appropriate. In the future, the Corrected Topic Model (Blei and Lafferty, 2007) and the dynamic topic model (Blei and Lafferty, 2006) might be better solutions to address these problems. Second, the data for this research is just the abstracts of six top accounting journals from Web of Science. To get more comprehensive results, electronic references of other accounting journals can also be included.

Reference

Agarwal, A. and Triggs, B. (2008), "Multilevel image coding with hyperfeatures", *International Journal of Computer Vision*, Vol. 78 No. 1, pp. 15-27.

An, X.Y. and Wu, Q.Q. (2011), "Co-word analysis of the trends in stem cells field based on subject heading weighting", *Scientometrics*, Vol. 88 No. 1, pp. 133-144.

Bao, Y. and Datta, A. (2014), "Simultaneously discovering and quantifying risk types from textual risk disclosures", *Management Science*, Vol. 60 No. 6, pp. 1371-1391.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55 No. 4, pp. 77-84.

Blei, D.M. and Lafferty, J.D. (2006), "Dynamic topic models", *Proceedings of the 23rd International Conference on Machine Learning in Pittsburgh, ACM, Pennsylvania, PA and New York, NY*, pp. 113-120.

Blei, D.M. and Lafferty, J.D. (2007), "Correction: a correlated topic model of science", *The Annals of Applied Statistics*, Vol. 1 No. 1, pp. 17-35.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Cambrosio, A., Limoges, C., Courtial, J.P. and Laville, F. (1993), "Historical scientometrics? Mapping over 70 years of biological safety research with co-word analysis", *Scientometrics*, Vol. 27 No. 2, pp. 119-143.

Coelho, L.P., Peng, T. and Murphy, R.F. (2010), "Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing", *Bioinformatics*, Vol. 26 No. 12, pp. i7-i12.

Cohen, P.U. and Austerweil, J.L. (2015), "Analyzing the history of cognition using topic models", *Cognition*, Vol. 135, pp. 4-9.

Cushing, B.E. (1989), "A Kuhnian interpretation of the historical evolution of accounting", *Accounting Historians Journal*, Vol. 16 No. 2, pp. 1-41.

Ding, Y., Chowdhury, G.G. and Foo, S. (2000), "Incorporating the results of co-word analyses to increase search variety for information retrieval", *Journal of Information Science*, Vol. 26 No. 6, pp. 429-451.

Dunbar, A.E. and Weber, D.P. (2014), "What influences accounting research? A citations based analysis", *Social Science Electronic Publishing*, Vol. 29 No. 1, pp. 1-60.

Dunning, T. (1993), "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, Vol. 19 No. 1, pp. 61-74.

Farcas, T.V. (2015), "An overlook into the accounting history evolution from a Romanian point of view – a literature review", *Knowledge Horizons – Economics*, Vol. 7 No. 3, pp. 14-27.

Feinerer, I., Hornik, K. and Meyer, D. (2008), "Text mining infrastructure in R", *Journal of Statistical Software*, Vol. 25 No. 5, pp. 1-54.

Franklin, R.E. and Jaeger, P.T. (2007), "A decade of doctorates: an examination of dissertations written by African American women in library and information studies", *Journal of Education for Library and Information Science*, Vol. 48 No. 3, pp. 187-201.

Gao, B., Li, X., Liu, S. and Fang, D. (2018), "How power distance affects online hotel ratings: the positive moderating roles of hotel chain and reviewers' travel experience", *Tourism Management*, Vol. 65, pp. 176-186.

Gao, S.J., Yu, W.Z. and Luo, F.P. (2009), "Citation analysis of PhD thesis at Wuhan University, China", *Library Collections Acquisitions & Technical Services*, Vol. 33 No. 1, pp. 8-16.

Griffiths, T.L. and Steyvers, M. (2004), "Finding scientific topics", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101 No. S1, pp. 5228-5235.

Hall, D., Jurafsky, D. and Manning, C.D. (2008), "Studying the history of ideas using topic models", *Proceedings of the Conference on Empirical Methods in Natural Language Processing in Honolulu, Hawaii, Association for Computational Linguistics*, *Stroudsburg, PA*, pp. 363-371.

Hofmann, T. (1999), "Probabilistic latent semantic indexing", *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Berkeley, CA, ACM*, *New York, NY*, Vol. 42, pp. 50-57.

Hornik, K. and Grün, B. (2011), "Topicmodels: an R package for fitting topic models", *Journal of Statistical Software*, Vol. 40 No. 13, pp. 1-30.

Kakkonen, T., Myller, N., Sutinen, E. and Timonen, J. (2008), "Comparison of dimension reduction methods for automated essay grading", *Educational Technology & Society*, Vol. 11 No. 3, pp. 275-288.

Hwang, S.Y., Wei, C.P., Lee, C.H. and Chen, Y.S. (2017), "Co-authorship network-based literature recommendation with topic model", *Online Information Review*, Vol. 41 No. 3, pp. 318-336.

Jo, Y. and Oh, A.H. (2011), "Aspect and sentiment unification model for online review analysis", *Proceedings of the 4th ACM International Conference on Web Search and Data Mining in Hong Kong, ACM*, *New York, NY*, pp. 815-824.

Li, W. and McCallum, A. (2006), "Pachinko allocation: DAG-structured mixture models of topic correlations", *Proceedings of the 23rd International Conference on Machine Learning in Pittsburgh, PA, ACM*, *New York, NY*, pp. 577-584.

Liu, G.Y., Hu, J.M. and Wang, H.L. (2012), "A co-word analysis of digital library field in China", *Scientometrics*, Vol. 91 No. 1, pp. 203-217.

Manning, C.D. and Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Newton, M.A. and Raftery, A.E. (1994), "Approximate Bayesian-inference with the weighted likelihood bootstrap", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 56 No. 1, pp. 3-48.

Niebles, J.C., Wang, H.C. and Fei-Fei, L. (2008), "Unsupervised learning of human action categories using spatial-temporal words", *International Journal of Computer Vision*, Vol. 79 No. 3, pp. 299-318.

Piepenbrink, A. and Nurmammadov, E. (2015), "Topics in the literature of transition economies and emerging markets", *Scientometrics*, Vol. 102 No. 3, pp. 2107-2130.

Rip, A. and Courtial, J.P. (1984), "Co-word maps of biotechnology – an example of cognitive scientometrics", *Scientometrics*, Vol. 6 No. 6, pp. 381-400.

Ritzhaupt, A.D., Stewart, M., Smith, P. and Barron, A.E. (2010), "An investigation of distance education in north American research literature using co-word analysis", *International Review of Research in Open and Distance Learning*, Vol. 11 No. 1, pp. 37-60.

Sugimoto, C.R., Li, D., Russell, T.G., Finlay, S.C. and Ding, Y. (2011), "The shifting sands of disciplinary development: analyzing North American library and information science dissertations using latent Dirichlet allocation", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 1, pp. 185-204.

Sun, L. and Yin, Y. (2017), "Discovering themes and trends in transportation research using topic modeling", *Transportation Research Part C: Emerging Technologies*, Vol. 77, pp. 49-66.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006), "Hierarchical Dirichlet processes", *Journal of the American Statistical Association*, Vol. 101 No. 476, pp. 1566-1581.

Wallach, H.M. (2006), "Topic modeling: beyond bag-of-words", *Proceedings of the 23rd International Conference on Machine Learning in Pittsburgh, PA, ACM, New York, NY*, pp. 977-984.

Wang, X. and Mccallum, A. (2006), "Topics over time: a non-Markov continuous-time model of topical trends", *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, ACM, New York, NY*, pp. 424-433.

Wang, X., Ma, X. and Grimson, W.E. (2009), "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models", *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 31 No. 3, pp. 539-555.

Wu, H., Wang, M., Feng, J. and Pei, Y. (2010), "Research topic evolution in 'bioinformatics'", *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in Philadelphia, PA, ACM, New York, NY*, pp. 424-433.

Xing, D. and Girolami, M. (2007), "Employing latent Dirichlet allocation for fraud detection in telecommunications", *Pattern Recognition Letters*, Vol. 28 No. 13, pp. 1727-1734.

Yan, E., Ding, Y. and Jacob, E.K. (2012), "Overlaying communities and topics: an analysis on publication networks", *Scientometrics*, Vol. 90 No. 2, pp. 499-513.

Yau, C.K., Porter, A., Newman, N. and Suominen, A. (2014), "Clustering scientific documents with topic modeling", *Scientometrics*, Vol. 100 No. 3, pp. 767-786.

Zong, Q.J., Shen, H.Z., Yuan, Q.J., Hu, X.W., Hou, Z.P. and Deng, S.G. (2013), "Doctoral dissertations of library and information science in China: a co-word analysis", *Scientometrics*, Vol. 94 No. 2, pp. 781-799.

Corresponding author
Baojun Gao can be contacted at: gaobj@whu.edu.cn

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**