Scientometrics (2020) 124:1191–1224
https://doi.org/10.1007/s11192-020-03508-3

Incorporating citation impact into analysis of research trends

Minchul Lee[1] · Min Song[1]

Received: 1 November 2019 / Published online: 18 May 2020
© Akadémiai Kiadó, Budapest, Hungary 2020

Abstract
In the past decades, there have been a number of proposals to apply topic modeling to research trend analysis. However, most of previous studies have relied primarily on document publication year and have not incorporated the impact of articles into trend analysis. Unlike previous trend analysis using topic modeling, we incorporate citation count, which can be viewed as the impact of articles, into trend analysis to shed a new light on the understanding of research trends. To this end, we propose the Generalized Dirichlet multinomial regression (g-DMR) topic model, which improves the DMR topic model by replacing a linear inner product in topic priors, $\exp\left(\boldsymbol{x}_d \cdot \lambda_t\right)$, with a more general form based on topic distribution function (TDF), $\exp\left(\mathrm{f}\left(\boldsymbol{x}_d\right)\right) + \varepsilon$. We use multidimensional Legendre Polynomial as TDF to capture publication year and the number of citations per publication simultaneously. In DMR model, since metadata could affect the document-topic distribution only monotonically and continuous values such as publication year and citation count need to be discretized, it is difficult to view the dynamic change of each topic. But the g-DMR model can handle various orthogonal continuous variables with arbitrary order of polynomial, so it can show more dynamic topic trends. Two major experiments show that the proposed model is better suited for topic generation with consideration of citation impact than DMR does for the trend analysis in the field of Library and Information Science in general and Text Mining in particular.

Keywords Topic model · NLP · Text mining · DMR · Trend analysis · Publication analysis

Introduction

Exploring trends in scientific research is an area of interest for many researchers. Recently, many papers have been published in an electronic form; thus, methods using text mining techniques have been increasingly applied for bibliometrics (Hawkins 2001). One of these text mining techniques is topic modeling, which can reduce dimensionality of vast text data interpretably and produce semantically coherent topics. There have been several attempts

✉ Min Song
min.song@yonsei.ac.kr

Minchul Lee
bab2min@yonsei.ac.kr

[1] Department of Library and Information Science, Yonsei University, Seoul, Korea

to utilize topic modeling to extract important research topics over time (Blei et al. 2003; Blei and Lafferty 2006; Wang and McCallum 2006; Hall et al. 2008; Mimno and McCallum 2012).

Topic modeling captures the probability of each topic's occurrence. In time-variant topic modeling, looking at frequent topics by period is one of the best ways to see how the research field has evolved. However, the question is whether the topics with high probabilities of appearance are the most important topics. Some topics may appear frequently in papers with low citations, while others may appear rarely but in highly cited papers. Therefore, to identify important research topics, it is necessary to consider not only the probability of occurrence but also the number of citations of the papers in which topics appear. While existing time-variant topic models have not taken citations into consideration, topic modeling has been extended to bibliometric measures based on citation counts with topical information (Mann et al. 2006; Kim et al. 2018; Wang et al. 2013). However, these approaches are limited due to the following two reasons. First, most of previous studies are feasible for trend analysis of scientific publications that are cited reasonably often, but they do not consider time variant factors (Kim et al. 2018; Mann et al. 2006; Wang et al. 2013). Second, a few prior studies that considered time variant factors such as Dirichlet multinomial regression (DMR) (Mann et al. 2006) and dynamic topic model (DTM) (Blei and Lafferty 2006) require extra steps to discretize continuous values like publication year and citation count to make the results of topic modeling meaningful. The results section confirms these problems.

To tackle these aforementioned issues, we propose a new generalized Dirichlet Multinomial Regression (g-DMR) topic model that replaces a linear inner product in topic priors with the general form based on Topic Distribution Function (TDF). The proposed model enables continuous and multidimensional analysis by incorporating metadata into topic model non-linearly as well as dynamically. We suggested multidimensional Legendre Polynomial as TDF for research trend analysis. Using g-DMR with Legendre Polynomial, the publication year and number of citations are cohesively combined into one topic model. To evaluate the utility of the proposed model, we applied it to two fields: Library Information Science (LIS) and Text Mining.

To verify the soundness of the proposed approach, we evaluated the quality of the proposed approach in terms of topic coherence and distribution inference. The results of the experiments demonstrated the utility of the proposed approach through a 2×2 matrix of high vs. low impact and hot vs. cold topics. For those who are interested in applying the proposed approach to other research fields, we made the source code and the results of experiments publicly available at https://github.com/bab2min/g-dmr.

In the subsequent sections, we examine related works about models capturing topic trends and discuss their limitations. We review the DMR topic model, generalize it by introducing the TDF, and suggest diverse versions of it. We verify that the proposed model works well through experimental results. Finally, we show the results of trend analysis on the field of LIS and Text Mining using the proposed model.

Related works

Research trend analysis

Most previous studies on research trend analysis have adopted bibliometric analysis methods. Jabeen et al. (2015) analyzed evolution rates and trends in the field of LIS by applying bibliometric techniques to about 18,000 articles published between 2003 and 2012. Sethi and Panda (2012) gauged the growth of the field of LIS by a citation analysis of 1000 articles published from 2000 to 2010. Li et al. (2009) traced the global trend of stem cell research between 1990 and 2007 to understand the characteristics of stem cell research activities by bibliometric techniques. Tran et al. (2018) conducted trend analysis in child maltreatment on Web of Science data with the query term "child maltreatment." By using VOSviewer for network analysis, they identified an increase trend in child maltreatment research, mostly conducted in the United States. Lv et al. (2011) investigated the global scientific production trend of graphene research using bibliometric analysis and visualization analysis with data collected between 1991 and 2010 from the Web of Science.

Another major direction in research trend analysis relies on content analysis, such as topic modeling and keyword analysis. Song et al. (2014) used the PubMed Central full-text database to retrieve articles related to the field of bioinformatics from 2000 to 2011. By applying Latent Dirichlet Allocation (LDA) to three periods (2000–2003, 2004–2007, and 2008–2011), they analyzed topical trends in the field of bioinformatics. Timakum et al. (2018) traced changes in LIS knowledge trends over the past 20 years to grasp developments in LIS research at a fine-grained level. By employing network analysis and topic modeling, they analyzed full-text journal articles from six top-ranked library science journals. Zou (2018) applied LDA for trend analysis to explore the temporal popularity of certain drug safety topics over time using 4347 related articles from 2007 to 2016. In addition, several papers incorporated citation relations into topic model (Wang et al. 2013; Jo et al. 2011). Jo et al. (2011) proposed a hybrid topic modeling algorithm that relied on the temporal order of papers and the document language model to detect the formation of new topics. In doing so, they computed the strength of the relationship between two topics through the total number of shared citations. However, since they ignored the direct relationship of topic dependency, they did not distinguish between differences in topic importance. To overcome this problem, Wang et al. (2013) proposed a probabilistic topic model for literature citation analysis that directly incorporated citation relations.

One recent study applied a deep learning algorithm to research trend prediction. Chen et al. (2018) proposed the new algorithm, the Correlated Neural Influence Model, to predict trending research topics at mutually influenced conferences by sequentially tracking the research topics for each individual conference and embedding all the research topics into the hidden space. In a different paper, the performance of journals over time in a given field was measured by incorporating journal ranking and publication year into the topic model (Song et al. 2017). These authors proposed a journal-time-topic model, an extension of Dirichlet multinomial regression, which we applied to the field of bioinformatics to understand journal contributions to topics in a field and shifts in topic trends.

However, these studies, including Wang et al. (2013), that apply topic modeling to research trend analysis have not yet incorporated both the number of citations and the publication year into the Dirichlet distribution.

Topic modeling

Blei and Lafferty suggested a Dynamic Topic Model (DTM) (2006), which generates documents of each time-slice from a normal distribution across topics, with each time-slice's distribution centroid drawn from the previous one. One of the key features of this model is that it not only generates the topic distribution per document for each period, but the word distribution per topic can also change for each time-slice. But since the document set is divided into N subsets by its publication date to form a Markov chain, the time stamp of documents is bound to be discretized. To solve this problem, continuous DTM was introduced (Wang et al. 2012). They used Brownian motion to model dynamics of the topics. It allows the model to accept a continuous time point of documents and handles a number of time points without a limitation of fixed granularity.

There are previous studies analyzing a scholarly impact using topic model Gerrish and Blei (2010). suggested Document Influence Model (DIM) based on DTM. They introduced influence score, which controls how much words in the article affects overall topic drift, to each article of DTM. By using these scores, they could measure the impact of articles on the scholarly trend. Gerow et al. (2018) suggested binding authorship, affiliation and publication venue to DTM and developed an evaluation method assessing their scholarly influence. This research pursued the similar objective with the present study in that it links topic model with the scholarly impact, but it differs from the proposed method greatly in terms of the algorithmic approach. While their work generated topics first, and measured the scholarly impact based on topics, in the present paper, we attempted to discovery topics from both time and scholarly impact. As shown in the Sects. 4 and 5, generating topics from these two dimensions of metadata can yield better quality than using only one of these metadata and can reliably visualize the research trends.

The topics over time (ToT) Model (Wang and McCallum 2006) assumes that a beta distribution related to each topic generates the time stamp of each document. Since the beta distribution is continuous, this model is characterized by its ability to deal with continuous values of each document's time stamp without discretization. However, one of the limitations of ToT is that the topic distribution of ToT must be in line with beta distributions. There are some previous studies analyzing research trend based on ToT model. One of them, Trend Detection Model (Kawamae and Higashinaka 2010), adopted a latent trend class variable into each document. The trend class is defined as topic co-occurrence patterns and it affects a topic distribution and a time stamp of each document. They proved that the introduction of trend class lowers perplexity of the model and improves the accuracy of timestamp prediction.

Also LDA (Blei et al. 2003) can be used to see the change of topics over time. Hall et al. (2008) suggested post hoc calculation for topic probabilities of each year. In general, the LDA model does not have a device for viewing the distribution of topics by time. However, they defined the empirical probability that a paper in a given year was about a certain topic, and the distributions of topic by year were calculated by applying LDA. Alternatively, the Dirichlet-Multinomial Regression model (DMR) (Mimno and McCallum 2012), an expanded version of the LDA for including metadata attached to documents, can be used to analyze trends in research. In the DMR model, the Dirichlet parameter of documents depends on its author, date, or other metadata. That is, the model can reflect the metadata of documents in topic generation.

There were some studies that exploited, not the temporal information, but the citations to analyze research trends. One of them, citation influence model (CIM) (Dietz et al. 2007), introduced a Bernoulli distribution determining whether words are drawn from the topic distribution of the current document or from the topic distribution of the cited document. This allows the topic distribution of the cited document to affect the citing document. They showed that the proposed model can generate a citation graph which includes significances of influence between documents. Also, there was an attempt to detect emerging topics based on the CIM and DIM. Xu and his colleagues introduced the framework for emerging topic detection, which extracts growth, coherence, and influence indicators using DIM and a novelty indicator using CIM (Xu et al. 2019). These indicators are fed to Multi-Task least-squares support vector machine to predict emerging topics. They showed that emerging topics can be detected by combining these indicators.

Meanwhile, structural topic Model (STM) (Roberts et al. 2014), which incorporates metadata of documents into topic prevalence and topic content, has been suggested. Since topic prevalence affects document-topic distributions and topic content affects topic-word distributions, STM provides a general method to analyze the impact of each metadata on topics from the topic structure viewpoint. STM can also accept linear, binary, or categorical variables as metadata by exploiting generalized linear model. Despite its flexibility, it has a limitation that the topic distributions are determined monotonically following a linear combination between its metadata variable and latent variable. The dynamics of topics we tried to inspect may show more complicated patterns with time or scholarly influence than linearly increasing or decreasing patterns. Thus, we needed to develop a new topic model, which is able to capture more dynamic topic patterns with multi-variable metadata than linear, to analyze research trends from the perspectives of both time and number of citations.

Methodology

DMR topic model

Before discussing the g-DMR, we first review the underlying statistical assumption of the DMR model. Mimno and McCallum (2012) proposed a feature vector λ_t for each topic t to generate topics, which are sensitive to metadata of documents. They expanded the generative process of document as below:

(1) For each topic t,

(A) Draw $\boldsymbol{\lambda}_t \sim N\left(0, \sigma^2 I\right)$.
(B) Draw $\boldsymbol{\Phi}_t \sim Dir(\boldsymbol{\beta})$.

(2) For each document d:

(A) For each topic t, let $\boldsymbol{\alpha}_{d,t} = \boldsymbol{x}_d \cdot \boldsymbol{\lambda}_t$.
(B) Draw $\boldsymbol{\Theta}_d \sim \mathrm{Dir}\left(\boldsymbol{\alpha}_d\right)$.
(C) For each word i,

(i) Draw $\mathrm{z}_i \sim \mathrm{Mult}\left(\boldsymbol{\Theta}_d\right)$,
(ii) Draw $\mathrm{w}_i \sim \mathrm{Mult}\left(\boldsymbol{\phi}_{z_i}\right)$,

where σ is the standard deviation of the prior on λ_t; $\boldsymbol{\beta}$ is a concentration parameter in the form of vector, which is generally considered symmetric, for topic-word distribution and $\boldsymbol{\Phi}_t$ is a topic-word distribution vector for topic t; $\boldsymbol{x}_d$ is the metadata of document d, which is represented as a |F|-dimensional vector; F is the set of features; $\boldsymbol{\alpha}_d$ is a parameter vector, which is the collection of $\boldsymbol{\alpha}_{d,t}$, for topic distribution derived from $\boldsymbol{\lambda}_t$ and $\boldsymbol{x}_d$; $\boldsymbol{\Theta}_d$ is a document-topic distribution vector for document d; and the other parameters are the same as denoted in the LDA model (Blei et al. 2003). Thus, this model can accept the metadata of documents and infer topic probabilities from the metadata.

$$P(\mathbf{z}) = \left(\frac{\Gamma(Ka)}{\Gamma(\alpha)^T}\right)^D \frac{\prod_{d,t}\Gamma(n_{d,t})}{\prod_d \Gamma(\sum_t n_{d,t}+T\alpha)}.$$

The likelihood of topic t, which is above in the LDA, expands as follows in the DMR:

$$\mathrm{P}(\mathbf{z}) = \int \mathrm{P}(z|\lambda)\mathrm{P}(\lambda)\mathrm{d}\lambda,$$

$$\mathrm{P}(\boldsymbol{\lambda}) = \prod_{t,k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\lambda_{t,k}^2}{2\sigma^2}\right),$$

$$P(\mathbf{z}|\boldsymbol{\lambda}) = \prod_d \frac{\Gamma(\sum_t \boldsymbol{\alpha}_{d,t})}{\Gamma(\sum_t (n_{d,t}+\boldsymbol{\alpha}_{d,t}))} \prod_{d,t} \frac{\Gamma(n_{d,t}+\boldsymbol{\alpha}_{d,t})}{\Gamma(\boldsymbol{\alpha}_{d,t})}.$$

As this equation is intractable, Mimno and McCallum (2012) trained the model using collapsed Gibbs sampling and adjusted the parameter λ using an L-BFGS optimizer.

This model proposes a method to incorporate metadata into topic model. However, it has a limitation that topic distributions for documents are generated monotonically for metadata because it is from a linear combination of $\boldsymbol{x}_d$ and λ_t. This means that the probability of a topic consistently increases or decreases as each feature, which consists of a vector $\boldsymbol{x}_d$, increases or decreases. Therefore, this model is not suitable for capturing dynamic changes in the topic distributions.

Generalized DMR (g-DMR)

In DMR topic model, $\boldsymbol{\alpha}_{d,t}$ is defined as $\boldsymbol{x}_d \cdot \lambda_t$, and it causes the limitation of monotonicity for metadata. We propose a g-DMR topic model to overcome the limitations of the DMR topic model, replacing $\boldsymbol{\alpha}_{d,t}$ as more general function.

$$\boldsymbol{\alpha}_{d,t} = \exp\left(f_t\left(\boldsymbol{x}_d\right)\right) + \varepsilon, \tag{1}$$

f_t receives the metadata vector of a document and returns the weight of the topic t of the document. Theoretically, the value of the exp function cannot be zero, but it often becomes nearly zero in some f_t functions due to the limitations of computation precision. So, we added non-zero ε, the smoothing parameter that prevents $\boldsymbol{\alpha}_{d,t}$ from getting too close to zero. We introduce various f_t with the name Topic Distribution Function (TDF) that can be used depending on the purpose of the analysis. Some of the TDFs that can be useful are shown in Table 1, and DMR model can be seen as the special case of g-DMR with linear TDF.

The constant TDF is the simplest form of TDF, with the topic weight not affected by the metadata of the document at all. Thus, in this case, the model will be the same as the LDA.

Table 1 Examples of topic distribution functions

Name	Formula	Description
Constant TDF	$f_t(x) = \lambda_t$	Topic distributions are not affected by the metadata (equivalent to the LDA model)
Linear TDF	$f_t(\boldsymbol{x}) = \lambda_t \cdot \boldsymbol{x}$	Topic distributions can be different by the one-hot encoded discrete metadata vector or linear metadata vector (equivalent to the DMR model)
Legendre TDF	$f_t(x) = \sum_{i=0}^{I} \lambda_{t,i} L_i(x)$	Topic distributions consist of *I*th order Legendre polynomial approximation (L_i is the *i*th Shifted Legendre polynomial)
2d Legendre TDF	$f_t(\boldsymbol{x}) = \sum_{i=0}^{I} \sum_{j=0}^{J} \lambda_{t,i,j} L_i\left(\boldsymbol{x}_1\right) L_j\left(\boldsymbol{x}_2\right)$	Topic distributions consist of 2-dimensional Legendre polynomial approximation with *I*th and *J*th orders

The linear TDF is the same form as in the DMR model, where the topic distribution of a document can be different by the metadata of the document only monotonically.

There are many methods, such as simple polynomial approximation, Fourier approximation, Hermite, Legendre polynomial approximation, and so on, used to approximate unknown continuous functions. Since a TDF is unknown until the end of training, these methods can be used as the basis of TDFs.

We initially tried to use simple polynomials as the basis of topic distribution function, but it didn't converge properly because of its instability. In more detail, the higher-order terms approach zero in the domain (− 1, 1). Conversely, they diverge in the wider domain than (− 1, 1). Thus, when simple polynomials were set as the basis, then the coefficients of higher order terms were likely to be too large or too small than the one of lower order terms in L-BFGS process, and it failed to optimize the model.

On the other hand, in the case of the Legendre polynomials, their value is bounded by [− 1, 1] within their domain, and they all have zeros. These features help L-BFGS optimization to converge better even if Legendre polynomials have the same order as the simple polynomials. The shifted version of Legendre polynomial has another advantage where since it has only integer coefficients and needs no weight function, it has a computationally lower cost than other polynomial approximation methods such as Hermite polynomial approximation. Due to these strengths, the Legendre polynomial and its variation, Shifted Legendre polynomial (SLP), are widely used to approximate complex equations in physics and engineering (Andrews and Andrews 1992).

Thus, we suggested the SLP as a basis for the TDF. Each order of SLP is defined as the following:

$$L_i(x) = \sum_{k=0}^{i} (-1)^{i+k} \binom{i}{k} \binom{i+k}{k} x^k.$$

The Legendre TDF in Table 1, which has only one variable, is the simplest form of TDF using the SLP approximation. This can be extended as the 2d Legendre TDF to receive 2 variables. The Legendre TDF can also be extended with more variables, but this is not shown in the table. The orders of SLP (I, J in the table) are critical to performance of the TDF. A higher order can make a more complex form of a TDF and approximate data accurately, but it needs a more expensive calculation cost. A lower order of TDF approximation can be performed quickly, but its results may be less accurate. In addition, for 2d Legendre TDF, we draw λ_t from different distributions for $i=0, j=0$ and for the others as follows:

$$\lambda_{t,0,0} \sim N\left(0, \sigma_0^2 I\right)$$
$$\lambda_{t,i\neq 0,j\neq 0} \sim N\left(0, \sigma^2 I\right),$$

where $\sigma_0 > \sigma$. Therefore, the intercept term can have a larger variance and the coefficients of higher order terms may have value closer to zero. Usually, the absolute value of the intercept term is larger than others, and it helps the TDF to converge easily.

And since SLP has a domain ranged [0, 1], the metadata of all documents should be normalized into [0, 1] as follows:

$$x_d^{'} = \frac{x_d - \min_{d'} x_{d'}}{\max_{d'} x_{d'} - \min_{d'} x_{d'}}.$$

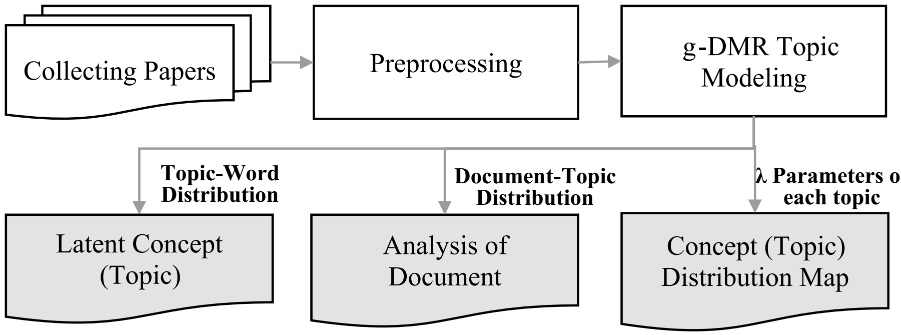


Fig. 1 Process of trend analysis using g-DMR

Table 2 Description of datasets

	Dataset I	Dataset II
Keywords	Text mining	LIS
Year of publication	2000–2017	1997–2017
Number of documents	10,907	54,703
Distribution of No. of CITATIONS	2 ~ 13,584	2 ~ 2561
Average of No. of citations (Median)	21.65 (6)	29.71 (8)
Download date	December 02, 2018	

Trend Analysis using g-DMR

Once λ is inferred through the model fitting process, the topic probability distribution at a specific point can be calculated by the following equation:

$$P(t|\boldsymbol{x}) = \frac{\exp\left(f_t(\boldsymbol{x})\right) + \varepsilon}{\sum_{t'}\left(\exp\left(f_t(\boldsymbol{x})\right) + \varepsilon\right)}, \tag{2}$$

where t is a topic and x is a metadata value. If x is constructed by the axis of the publication year and the number of times of citation, we can infer the probability of each topic at a specific value. Based on this, the research trend of the topic can be examined.

Figure 1 shows the process of trend analysis proposed in this paper. The first step is collecting all the academic papers in the fields of 'Text Mining' and 'LIS'. For LIS, we downloaded all available articles in the field of Library and Information Science provided in Scopus. For Text Mining, since there was no such explicit field in Scopus, we collected all articles that contain a keyword 'text mining'. The collected datasets include abstract, year of publication, and the number of citations. The year of publication, and the number of citations can be used to analyze which topics were hot or cold in the field by time series analysis.

Table 2 shows the details of the data collection for experiments. We downloaded two sets of scholarly publications from Scopus. One dataset consists of publications in the field of Library and Information Science (LIS) and the other dataset consists of publications in the field of Text Mining.

The following pre-processing steps were applied: (1) stemming the documents, (2) removing stopwords, and (3) refining features. We selected Porter Stemmer rather than lemmatization techniques. It is because the text we used is the abstract section of publications whose length is typically short. Thus, lemmatizing the text results in the sparser representation of the text than what stemming generates. Also, a lemmatizer results in similar words with different POS tags belonging to the same topics (e.g. lemmatization, lemmatize, lemmatizing). The distribution of the number of citations has a typical long-tail distribution, which consists of a small number of upper rankers and numerous lower rankers. Since the numbers of citations were highly skewed, we took its relative rank each year, not its citation count. The relative rank is determined as a value between 0 and 1. The most cited articles in the year are given a value of 1, and the least cited articles are given a value of zero. And a paper with a median number of citations is given a 0.5.

These preprocessed documents go through the g-DMR topic model process, resulting in three types of output, which are topic–word distribution, document–topic distribution, and λ parameters of each topic. The topic–word distribution is used to represent the semantic meaning of each topic. To calculate the topic distribution for a specific year and number of citation, λ parameters can be used. This allows us to draw a topic distribution map. The document–topic distribution will not be used for trend analysis, but we can use it to analyze documents semantically.

Evaluation

We performed experiments on two datasets, academic papers in the fields of LIS and Text Mining, to evaluate the new model. For the experiments, we implemented the g-DMR code in the C++programming language based on the PCGS (Yan et al. 2009), which is capable of multithreaded performance to handle a large collection of text efficiently. In addition, to accelerate the Gibbs sampling process and the L-BFGS optimization using GPU, OpenCL was used. L-BFGS[1] was used to estimate parameters.

To perform the g-DMR topic model with the metadata, year of publication x, and relative citation ranks y, the following 2d Legendre TDF was used:

$$f_t(x, y) = \sum_{i=0}^{I} \sum_{j=0}^{J} \lambda_{t,i,j} L_i(x) L_j(y).$$

It is important to select appropriate hyper-parameters for the performance improvement of the model. Normally, β is set to a value less than 1, which makes the topic-word distribution sparse. It is possible to inference β as an optimal value (Blei et al. 2003), but β is usually set between 0.01 and 0.1. We setβ =0.01 for both models same as in DMR (Mimno and McCallum 2012). The parameters σ and σ_0 affect the variation of the topic distribution by the metadata. Since the smaller σ and σ_0 generates the more uniform the overall distribution by the metadata, too small σ and σ_0 values make it hard to differentiate topics. On the other hand, too large σ or σ_0 disturbs the convergence of the entire model because it maximizes the distribution difference and makes the Gibbs sampling process in the model

[1] https://www.chokkan.org/software/liblbfgs/.

Table 3 Perplexities by I and J for each dataset in $K=40$

	$I=0$	1	2	3	4	5
Dataset I (text mining)						
$J=0$	2049.90	1983.60	1993.60	1995.19	1995.59	1968.24
1	2068.78	1988.46	**1959.60**	**1957.45**	1981.96	1975.77
2	2010.97	2011.96	1993.88	**1957.20**	1972.47	1987.38
3	2015.13	1971.19	1992.55	1979.05	**1966.18**	1986.05
4	2030.75	1988.90	**1966.75**	2005.45	1986.89	2016.61
Dataset II (LIS)						
$J=0$	2510.73	2454.95	2456.08	**2445.34**	**2454.30**	2470.64
1	2467.23	2474.34	**2449.14**	2456.82	2461.77	2481.55
2	2496.10	2500.02	**2452.58**	2475.80	2471.94	2488.43
3	2491.95	2480.46	2466.96	2455.38	**2448.60**	2462.77
4	2472.29	2489.08	2457.23	2472.50	2466.11	2478.06

unstable. When setting$>\sigma 3$ for DMR and$>\sigma_0 3$ for g-DMR, both models often failed to converge in our data. Thus, we set $\sigma = 3$ for DMR and $\sigma_0 = 3$ for g-DMR. In g-DMR, σ determines the distribution of the coefficient of the higher order term. Thus, this value should be less than σ_0 to help converge the entire model. we tried using various values and found an appropriate value of 0.25 forσ_0. Finally, the smoothing parameter ε does not play an important role in the convergence of the model. This is only to prevent the exponential term in Eq. (1) from getting close to zero, and we used a very small value of 10^{-10} for only g-DMR.

Topics of LIS and text mining extracted by g-DMR

We trained each model with 800 iterations where initial 200 iterations were the burn-in period. Thus, optimization of the parameters λ was disabled in the burn-in period. After initial 200 iterations, we optimized the parameters every 20 iterations For optimization, all λ parameters were drawn from its random distribution and the optimal values were found by using L-BFGS. This was done totally 5 times independently, and the best λ parameters were selected. The evaluation was performed over 10-fold cross validation.

As the value of I and J can change the overall performance of g-DMR, we conducted the preliminary test of perplexity for each I and J value at $K=40$. Table 3 shows perplexities of the result from $I=0$ to 5 and $J=0$ to 4 and the top 5 values are shown in bold. For both datasets, the result of $I=4$ and $J=3$ shows the top performance. Accordingly, we chose these parameters for the following experiments. Interestingly, unlike our expectation, the high I and J values do not always produce the better perplexity. This seems to be due to the fact that the high I and J generate the more complicated space, which makes it difficult to find the global minimum with L-BFGS.

Tables 4 and 5 show the results of g-DMR with the number of topics setting to 40. In particular, as shown in Table 4, the g-DMR results for the field of text mining present various core subjects of text mining including text classification, document clustering and semantics and ontology as important topics. In addition, the results show that the most dominant application areas are related to the biomedical domain including bioinformatics, health informatics, and drug discovery ("Appendix 1").

Table 4 Topic–word results for Top 5 topics from text mining dataset at $K=30$

Topic labeling	Weight	Top words
Text classification	0.056	Classification, feature, classify, select, perform, algorithm, machine, base, accuracy, category
Document clustering	0.055	Document, term, word, similar, measure, base, semantic, extract, keyword, sentence
Semantics and ontology	0.047	Knowledge, semantic, concept, ontology, extract, domain, base, relationship, relation, discovery
System development	0.043	System, service, process, tool, application, develop, support, user, design, document
Biomedical corpus annotation	0.041	Biomed, literature, annotation, article, biology, curation, extract, database, abstract, interact

Table 5 Topic–word results for Top 5 topics from LIS dataset at $K = 40$

Topic labeling	Weight	Top words
Bibliometrics	0.0346	Journal, citation, science, article, public, publish, author, scientific, cite, field
Information management	0.0332	Manage, project, implement, system, process, plan, change, success, organ, practice
Cheminformatics (protein structure)	0.0307	Bind, protein, structure, ligand, dock, molecular, model, interact, inhibitor, energy
Group interview	0.0305	Group, participle, data, interview, user, seek, behaviour, interact, behavior, different
Information theory	0.0302	Estimation, distribution, function, algorithm, error, signal, model, general, sample, optimal

As shown in Table 5, the salient topics of LIS include bibliometrics, information management and information theory. In addition, research domains that are highly interactive with LIS are information system, cheminformatics, health communication, bioinformatics, and business informatics ("Appendix 2").

Evaluation of the quality of citation impact-sensitive topic trends

In this sub-section, we reported on the result of quality of citation impact-sensitive topic trends. To this end, we measure the inference performance of g-DMR for grasping the citation-sensitive topic trends in LIS as well as Text Mining by comparing the inferred trends with the actual research trends. We also compared the inference performance of g-DMR with DMR's. However, for the DMR model to perform topic modeling in consideration of year and citation count, year of publication has to be discretized into year unit and the number of citations has to be discretized as well. This is a major limitation of DMR with regards to citation impact-sensitive topic trend analysis. For the reported results, we discretized year and citation count into three, six, and nine bins, respectively. For example, three partitions consist of the bottom 1/3, the top 1/3, and the rest. For DMR, these discretized years and the number of citations are combined by a Cartesian product and finally used as metadata.

As will be described in the subsequent section, higher K values show better perplexity, but the results are more difficult to interpret. Therefore, choosing an appropriate K value is very critical to the overall interpretation of results. In this evaluation, the K value was selected as $K=30$ for Text Mining and $K=40$ for LIS, based on the inflection point where perplexity begins to decrease as K increases.

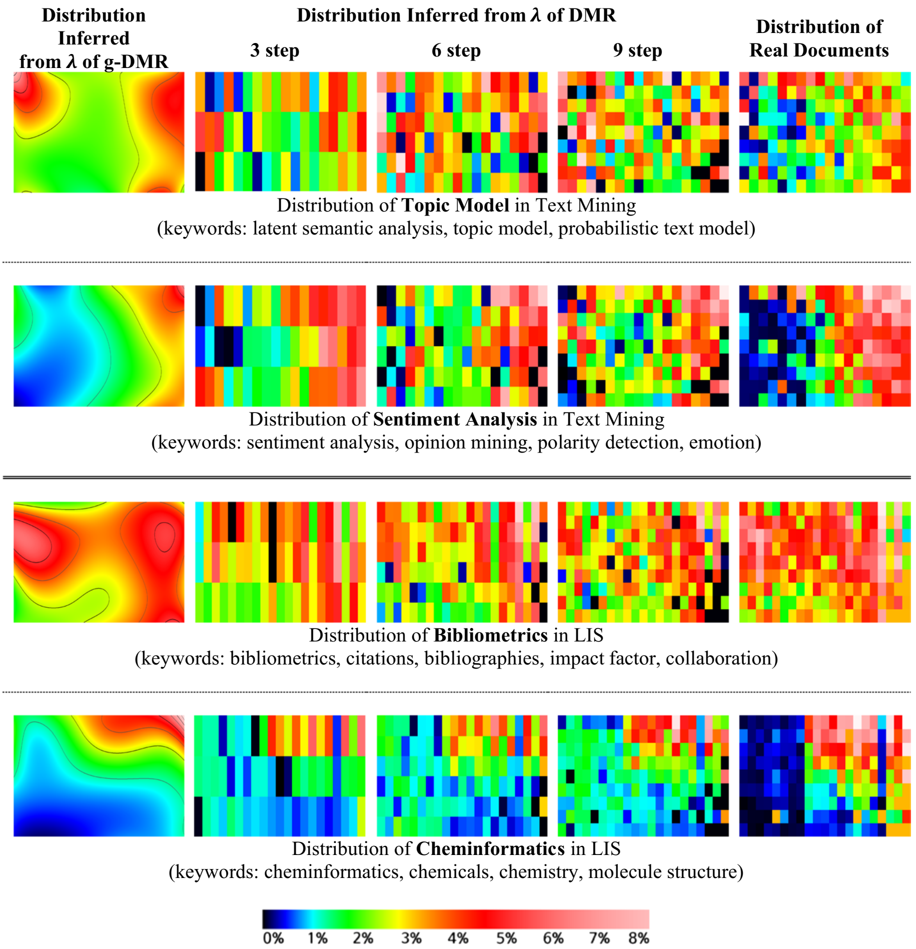
We chose four topics that are not too broad or general since a broad or general topic tends to appear from all areas, and it makes estimated topic distribution undistinguishable among topics. On the other hand, topics that are narrow or specific in a certain area have a complicated distribution, which is suitable for judging the quality of the estimated topic distribution. Thus, we chose "Topic Models" and "Sentiment Analysis" from Text Mining and "Bibliometrics" and "Cheminformatics" from LIS to evaluate the performance of citation impact-sensitive topic trend analysis. As shown in the Fig. 2, these topics are not common and are concentrated at a certain time, making them suitable for measuring the quality of an estimated topic distribution. To obtain real distributions of each topic, we selected relevant documents using abstracts and keywords. For DMR model, we chose the equivalent topics to ones which were selected from g-DMR. The major filtering keywords of each topic, which are displayed in Fig. 2, were used to get the real distribution.

Since each distribution is not a probability distribution, we cannot compare these distributions using statistical distances such as Kullback–Leibler divergence. Instead, we define the sum of squared errors between the log likelihoods of two distributions as a distance function:

$$\mathrm{dist}(P,Q)=\sqrt{\int(\log P(x)-\log Q(x))^2dx},$$

where P and Q are distributions with all values greater than 0 and $\mathrm{dist}(P,Q)=0$ if only $P=Q$.

Table 6 shows distribution distances of g-DMR and DMR from its real distribution for the chosen topics and the smallest one is shown in bold. Since in the distribution of real



Distribution of **Topic Model** in Text Mining
(keywords: latent semantic analysis, topic model, probabilistic text model)

Distribution of **Sentiment Analysis** in Text Mining
(keywords: sentiment analysis, opinion mining, polarity detection, emotion)

Distribution of **Bibliometrics** in LIS
(keywords: bibliometrics, citations, bibliographies, impact factor, collaboration)

Distribution of **Cheminformatics** in LIS
(keywords: cheminformatics, chemicals, chemistry, molecule structure)

Fig. 2 Result comparison of text mining and LIS between g-DMR and DMR

Table 6 Distribution distances from real one of g-DMR and DMR

Topics		g-DMR	DMR (3 step)	DMR (6 step)	DMR (9 step)	Baseline
Dataset I (Text Mining)	Topic model	**0.843**	1.088	1.487	1.786	0.901
	Sentiment analysis	**1.352**	1.374	1.748	2.012	1.591
Dataset II (LIS)	Bibliometrics	0.461	1.560	1.037	1.576	**0.399**
	Cheminformatics	**1.165**	1.584	1.611	1.854	1.790

documents, $P(x)$ can be 0 where dist (P, Q) cannot be defined, the distribution of real documents was smoothed by adding a small value 1e-2 for all range. We set a uniform distribution with a certain likelihood regardless of year or the number of citations as a baseline. We could assume that the case of performing the constant TDF, that is, LDA model, will show similar results to the baseline.

As shown in Table 6, the inference performance of DMR varies depending on the number of steps. This implies the inadequateness of DMR for citation impact-sensitive topic trend analysis in that steps we chose were determined empirically.

Topic distributions can be calculated following Eq. (2) by the $\boldsymbol{\lambda}$ parameters generated by the g-DMR model. These topic distributions capture the location where documents of a specific topic are concentrated.

The distribution of documents is discrete by nature and cannot be compared with a continuous distribution directly. Thus, we measure the proportion of related documents in bins divided by the grid to compare inferred with real distributions. The X axis, publication year, was divided by year; the Y axis, the normalized number of citation, was divided into 3, 6, or 9 bins, respectively. For example, the data collection of Text Mining was divided into 162 bins and the data collection of LIS was divided into 189 bins.

Figure 2 shows that the distributions inferred from $\boldsymbol{\lambda}$ of g-DMR approximate the real quite well. The red part of the distribution represents the area where many documents were published, and these areas are well-matched between the left and the right. Also, in all cases, the distance of g-DMR was smaller than DMR.

On the other hand, the DMR model in both 3-step, 6-step and 9-step showed higher distances than the baseline and did not learn the real distribution well. Even the model of the finer grid showed the larger distance. In DMR models, the finer is the grid, the more parameters are to learn, but the fewer documents are included in each grid. Thus, the performance seems to be worse in the finer model.

Minor differences between g-DMR and real distribution can be caused by the following factors: (1) a lack of documents for some time points and (2) a gap between the author's intention and the topic estimated from word distribution. If we minimize these factors effectively, the distribution inferred from λ can be used to observe trends of each topic.

Evaluation of topic quality by g-DMR

Generally, the results of the topic models are evaluated based on the likelihood of the held-out documents or perplexity. A higher likelihood or lower perplexity means that the model predicts more accurately for an unseen document. Thus, likelihood and perplexity can be used to compare overall performance of models. However, as Chang et al. asserted, lower perplexity does not always mean better results for human interpretation and sometimes gives results which is difficult to interpret (Chang et al. 2009). Therefore, it is necessary to evaluate topic model using other measure together rather than using only perplexity. Thus, we also applied a topic coherence evaluation based on Wikipedia (Newman et al. 2010) in addition to perplexity evaluation.

Figure 3 shows the log perplexities of g-DMR and DMR from the number of topics from $K = 10$ to 100 for each dataset. At every point, g-DMR shows better perplexity than DMR. This means that g-DMR can reach the higher fitness to given data than DMR does as g-DMR can incorporate topic distributions beyond monotonicity. Also as mentioned above, we conducted following topic coherence evaluation using external data.

Higher topic coherence is a signal of a good topic to interpret (Newman et al. 2010). If a topic model simply produces low perplexity but its result is difficult for humans to interpret, it would not a suitable model for analysis. Thus, we should consider both perplexity and topic coherence when comparing models' performance. Average topic coherence scores for the top 5 words in each topic from the g-DMR and DMR are shown in Table 7. The higher score among DMR and g-DMR is shown in bold. Since Gibbs sampling is a probabilistic process,

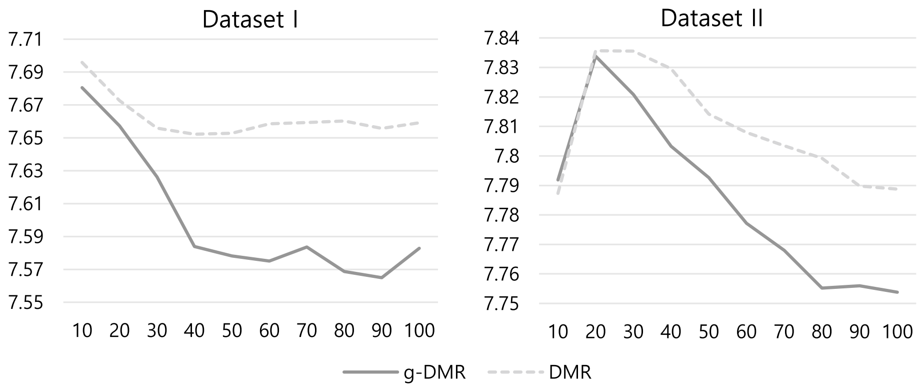


Fig. 3 Log perplexity of g-DMR and DMR from $K = 10$ to 100 (lower is better)

Table 7 Average topic coherence of g-DMR and DMR (higher is better)

	Dataset I (text mining)		Dataset II (LIS)	
	g-DMR	DMR	g-DMR	DMR
10	**1.291**	1.249	**1.152**	1.139
20	1.223	**1.244**	1.157	**1.182**
30	**1.213**	1.206	**1.176**	1.161
40	**1.209**	1.200	1.175	**1.183**
50	**1.225**	1.202	1.180	**1.195**
60	1.209	**1.218**	1.187	**1.201**
70	**1.213**	1.212	**1.198**	1.198
80	1.216	**1.218**	1.194	**1.204**
90	1.210	**1.224**	1.203	**1.208**
100	1.225	**1.231**	1.199	**1.210**

different results can be produced at run time, so one model may luckily outperform the other. In order to minimize this kind of error, the Welch's t-test was performed to 8 samples except 2 extremes out of 10 obtained in each of two models. The result of the statistical test is provided in "Appendix 3". Overall, there was no statistically significant difference between g-DMR and DMR. This means that g-DMR is equivalent to DMR from the perspective of generating coherent topics.

Taking all the above into consideration, it can be concluded that g-DMR generates the same quality of topics as DMR but can reflect the given data in the topic model. In addition, another strength of the g-DMR model is that it can estimate the distribution of topics according to the year of publication or number of citations. This is evaluated in the next section.

Citation impact-sensitive topic trend

As explained above, the experimental results with Dataset I and Dataset II can be used to analyze the topic trends in Text Mining and LIS. The trends of each topic were drawn as a topographical map using λ parameters. As we can see from the topographical map where the heights and valleys are shown, the map helps identify when the topic was studied intensively and how many times it was cited.

Figure 4 shows the distributions of 30 topics from the Text Mining field. In each map, the X axis represents the year of publication, and the Y axis represents the relative citation weight. The blue and dark colors indicate a low proportion of documents, and the red and bright colors indicate that the proportion of the documents is high. For example, a map of two general topics shows red or white colors in its entire range. This means that these three topics appear in almost every document of the corpus. As another example, the map of Social Network Analysis is green overall, but reddish after 2010s. One can interpret from this that in the field of Text Mining, Social Networks Analysis were introduced in the early 2000s, but did not cause an impact. However, the topic has been drawing attention recently.

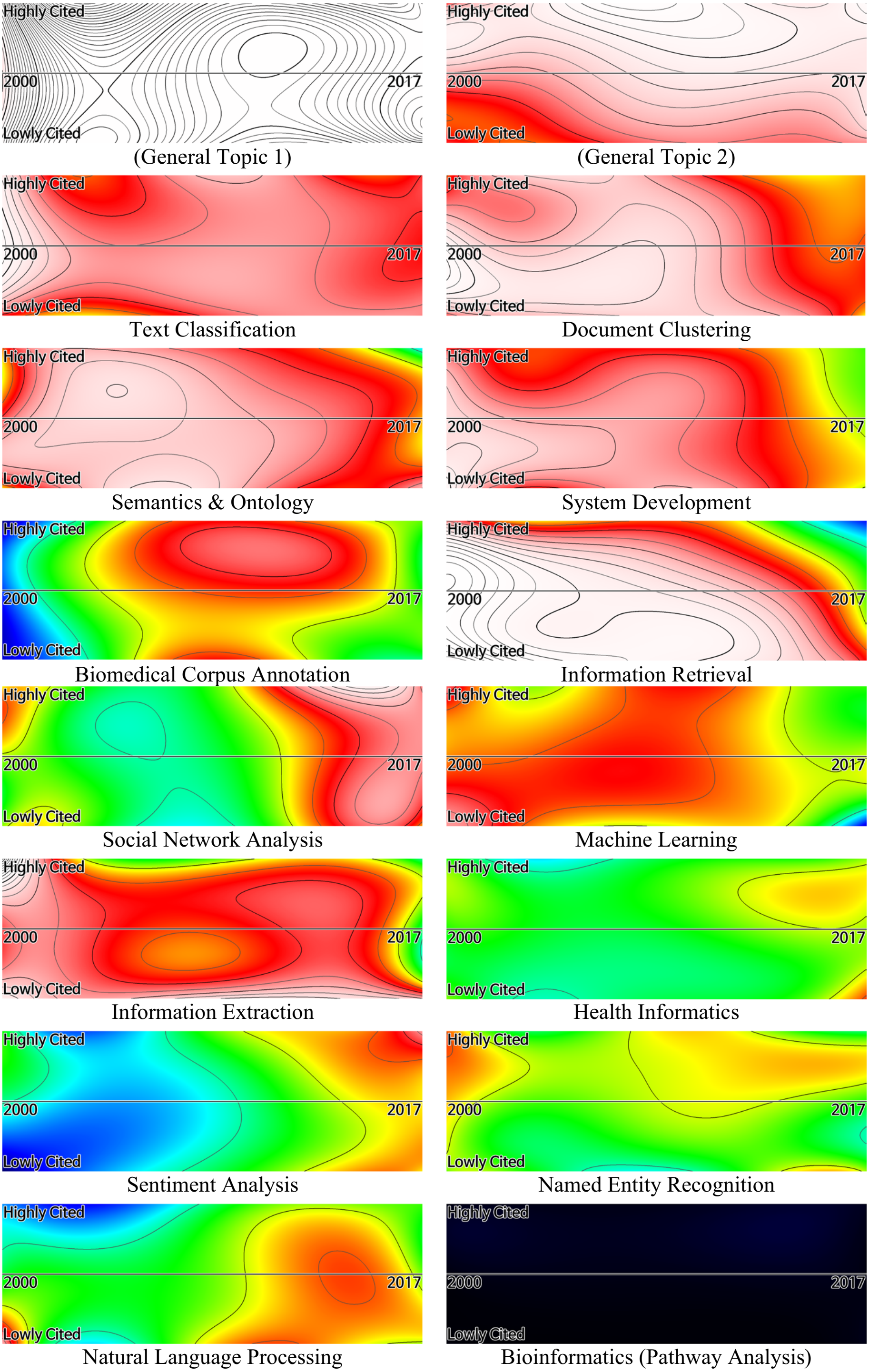
We also categorized the topics into four groups based on citations and recent trends in Table 8. The “Cold Recently” group includes the topics that had high weights in the 2000s but have had low weights recently. Conversely, the “Hot Recently” group has the topics that had low weights in the 2000s but have obtained high weights recently. The topics in the “Lowly Cited” group have a lower number of citations than expected, and the topics in the “Highly Cited” group have a higher number of recent citations than expected.

There are some topics whose weights are too small to observe research trends. For these topics, it is difficult to analyze trends because the value of the overall distribution falls within the error bounds. Therefore, we omitted them from the table.

The combination of Hot or Cold Recently and Highly or Lowly Cited yields four categories in total. The topics in the category of Hot and Lowly Cited have been widely studied by many scholars in recent years, but relatively few papers have cited them. This category includes Health Informatics, Natural Language Processing, Recommendation System, etc. On the other hand, the topics in the Cold and Highly Cited category have a relatively large number of citations, but the number of papers published on them recently is small. These include Biomedical Corpus Annotation, Bioinformatics (Genomic Database & Interaction Database), etc. The topics in the Cold and Lowly Cited category, such as System Development and Data Mining, have been paid lesser attention to by scholars than before. Finally, the topics of the Hot and Highly Cited category, such as Social Network Analysis, Sentiment Analysis, Bibliometrics, etc., are the fields in which the number of citations is relatively high and there have recently been increasing interests in research topics in these areas.

Figure 5 shows the distributions of 40 topics from the LIS field. It provides a much broader spectrum of the field than what the trends of Text Mining show. Although the compositions of the figure and table are the same as those of Fig. 4 and Table 8, the overall topography of these results is more complicated because of the larger size of the corpus. It contains three general topics, which are related to methods of research, not subjects, such as Survey Research and Group Interview. We omitted those general topics and minor topics whose distribution weights are too small.

In the LIS field, there are some Cold and Lowly Cited topics, which have been widely studied in the past and received relatively little attention and a low number of citations, as shown in Table 9. These include Information Management and Collection Management. The topics in the Hot and Lowly Cited category vary from Information Literacy to Information Ethics and are currently being studied extensively in LIS. On the other hand, in the Cold and Highly Cited category, there appear interdisciplinary topic areas such as Information Theories and Business Informatics. Lastly, in the Hot and Highly Cited category, there are also many interdisciplinary topics similar to Cold and Highly Cited group, such as Bibliometrics, Social Media and Cheminformatics. This shows the increasing trends in



(General Topic 1)

(General Topic 2)

Text Classification

Document Clustering

Semantics & Ontology

System Development

Biomedical Corpus Annotation

Information Retrieval

Social Network Analysis

Machine Learning

Information Extraction

Health Informatics

Sentiment Analysis

Named Entity Recognition

Natural Language Processing

Bioinformatics (Pathway Analysis)

Fig. 4 Trends of 30 topics in the text mining field

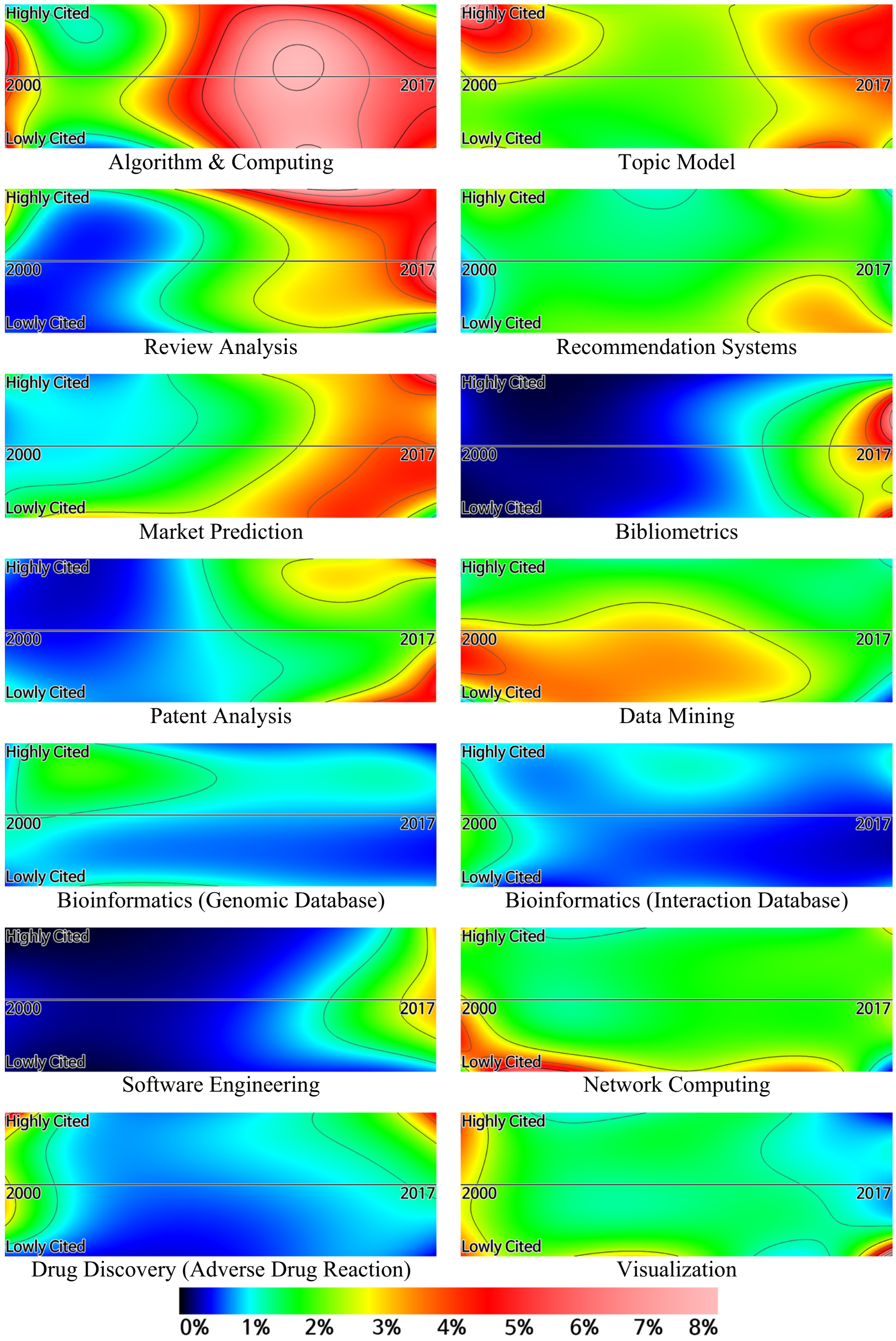
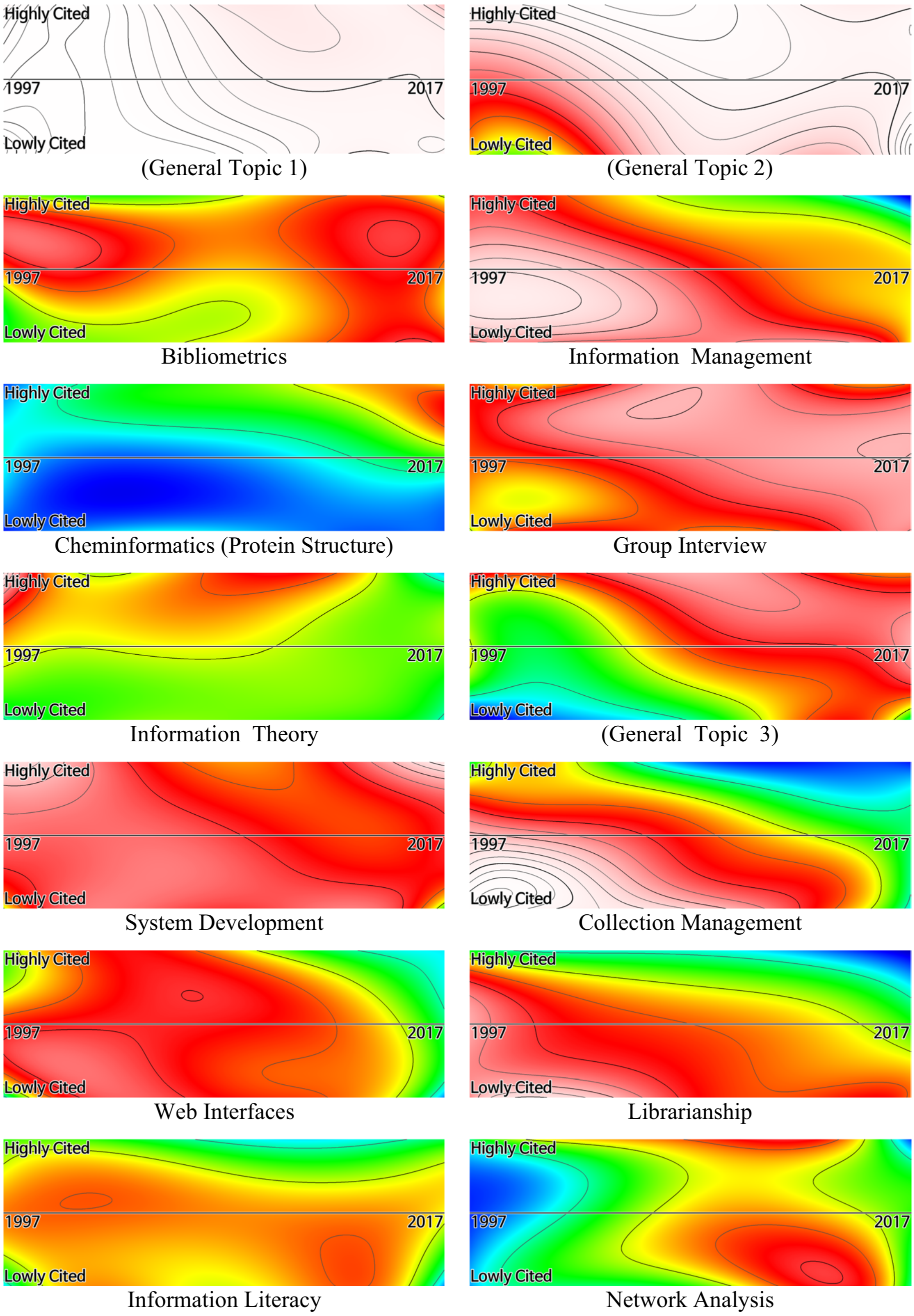


Fig. 4 (continued)

Table 8 Brief Categorization of Topics from Text Mining

	Cold recently	Hot recently
Lowly Cited	System development, data mining	Health informatics, natural language processing, recommendation system, market prediction, patent analysis
Highly Cited	Biomedical corpus annotation, bioinformatics, machine learning	Social network analysis, sentiment analysis, topic model, review analysis, bibliometrics



(General Topic 1)

(General Topic 2)

Bibliometrics

Information Management

Cheminformatics (Protein Structure)

Group Interview

Information Theory

(General Topic 3)

System Development

Collection Management

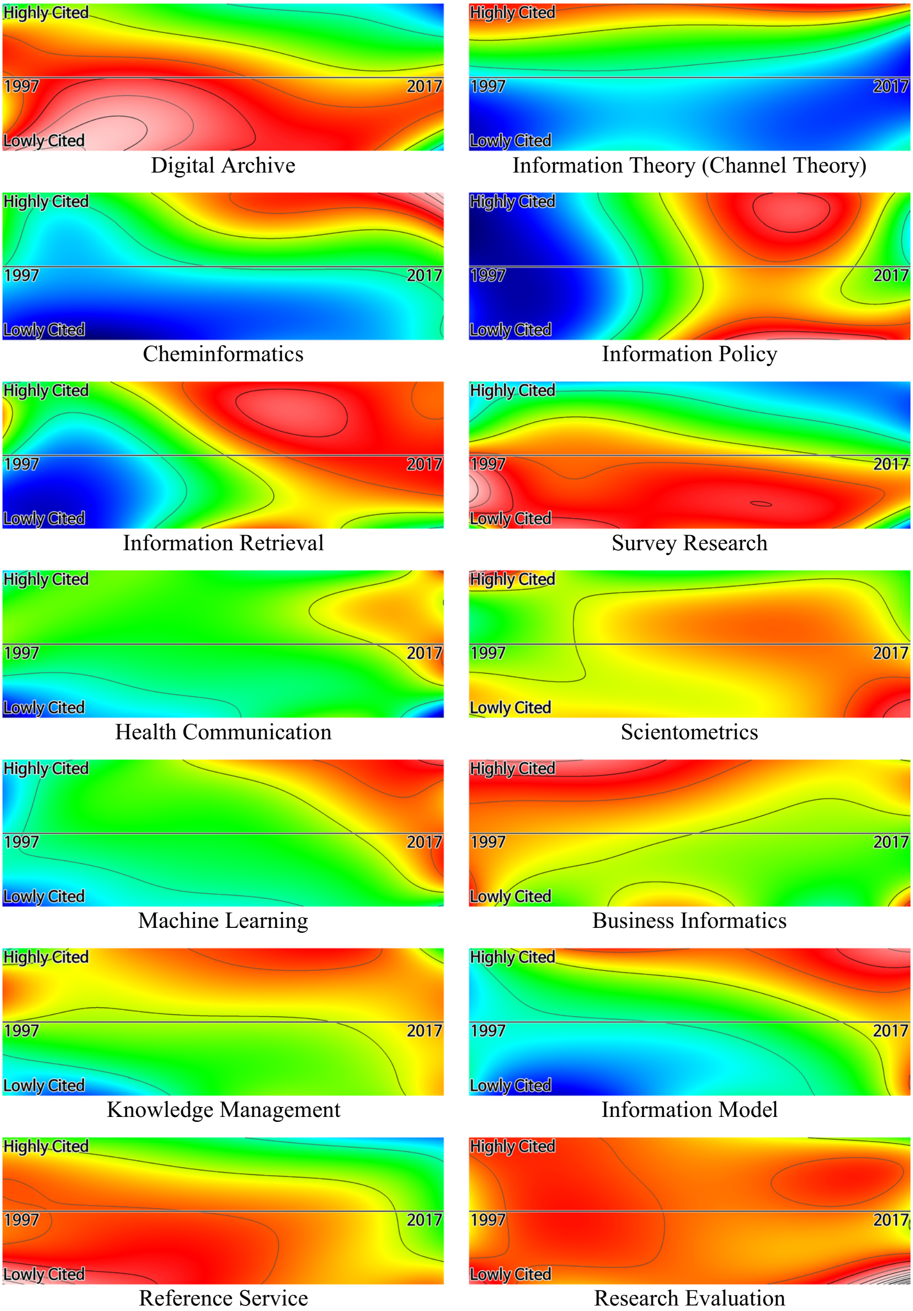
Web Interfaces

Librarianship

Information Literacy

Network Analysis

Fig. 5 Trends of 40 topics in the LIS field



Digital Archive

Information Theory (Channel Theory)

Cheminformatics

Information Policy

Information Retrieval

Survey Research

Health Communication

Scientometrics

Machine Learning

Business Informatics

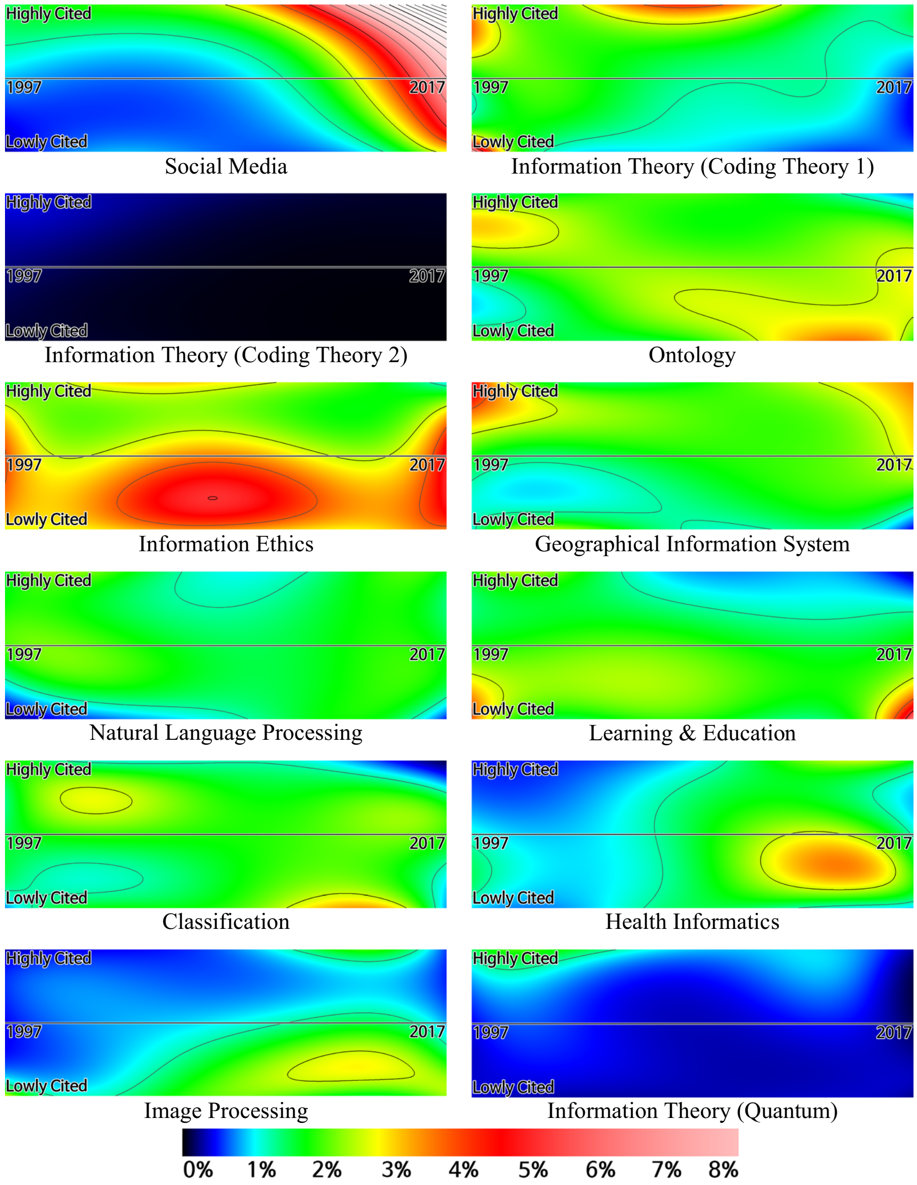
Knowledge Management

Information Model

Reference Service

Research Evaluation

Fig. 5 (continued)


Social Media

Information Theory (Coding Theory 1)

Information Theory (Coding Theory 2)

Ontology

Information Ethics

Geographical Information System

Natural Language Processing

Learning & Education

Classification

Health Informatics

Image Processing

Information Theory (Quantum)

Fig. 5 (continued)

collaboration between LIS and other disciplines, and these topics gain more attraction than traditional LIS topics.

We further examine the recent hot topic "Scientometrics". A total of 50 articles, which have more than 50% of the topic Scientometrics, were published in 2016. 37 of them were

Table 9 Brief categorization of topics from LIS

	Cold recently	Hot recently
Lowly cited	Information management, collection management, librarianship	Information literacy, network analysis, research evaluation, information ethics, health informatics
Highly cited	Information theory, business informatics	Bibliometrics, cheminformatics, information retrieval, social media

published in the Scientometrics journal. On the subject of research, most of them dealt mainly with collaboration networks (Zhang et al. 2016; Bouabid et al. 2016; Zhao and Zhao 2016). They covered various different collaboration types of countries, disciplines, and authors, and those papers were relatively highly cited. Another major subject was trend analysis, which included discipline-specific case studies (Cavacini 2016; Milanez et al. 2016; Stein et al. 2016). There were studies using VOSViewer, a visualization software (Moed 2016), and few researches proposed a visualization technique itself (Liu and Mei 2016). There were a few studies on patent analysis (Fukugawa 2016; Dou and Kister 2016; Kang and Sohn 2016). On average, they received fewer citations than collaboration research did.

Conclusion and future work

We proposed the g-DMR topic model that is a universal method for multi-dimensional featured topic models by incorporating multiple arbitrary features into topic model. In the present paper, we applied the proposed model for analyzing citation impact-sensitive research trends. To evaluate the performance of g-DMR, we downloaded 10,907 articles on text mining and 54,703 articles on LIS from Scopus. We compared the inferred topic trends in consideration of citation count with the actual topic trends for LIS as well as text mining. As reported in the Sect. 4.2, DMR is not suitable for citation impact-sensitive topic trends analysis because we have to arbitrarily partition the datasets by both citation counts and year. In contrast to DMR, g-DMR overcomes this problem by proposing Legendre TDF. The comparison results between the inferred and the actual topic trends show that g-DMR enables to predict the topic trends accurately whereas DMR with the several partitioning options fails to predict the topic trends. Moreover, the results also show that g-DMR enables to analyze citation impact-sensitive topic trends. Due to the nature of g-DMR, the proposed approach is highly applicable for other domains. As stated in the introduction section, g-DMR is publicly available. Thus, if similar studies in other domains are to be conducted in the future, it would shed a new light to understand the research trends in the applied domain by harmoniously incorporating citation counts into topic trend model.

In addition to the TDF presented in this paper, a different type of TDF can be proposed to handle several continuous metadata and categorical metadata variables together, and it is also possible to model two or more metadata categories of the documents that are independent of each other. g-DMR provides more general generative models than DMR does and can perform topic inference with more complicated metadata. It can be also used for social text or other analysis where more than one metadata variable is attached. For example, if there is a set of social texts scored for multiple sentiment aspects, we can find word distributions related to each sentiment through g-DMR analysis. Furthermore, it is expected that it is also possible to predict the best-matching distribution of the given word set and the sentiment score.

However, there are a couple of limitations of g-DMR. First, the overall performance of the model depends heavily on parameters *I*, *J* (the order of Legendre Polynomials) and *K* (the number of topics). Thus, the appropriate value of the parameters should be obtained by an experiment. Another limitation is that g-DMR is flat. In other words, g-DMR cannot grasp the hierarchy of or correlation between topics. Since the topics in scholarly documents are not independent but have hierarchies or relations with each other, g-DMR may miss information about this.

Our model can be extended in several potentially useful ways. First, a non-parametric inference, which is used in HDP (Teh et al. 2005) or hLDA (Griffiths et al. 2004), can be applied to g-DMR. This solves the problem of selecting appropriate values for parameters by fitting the number of topics, *K*, from the given data.

In addition, integrating a hierarchical model into g-DMR is another useful extension that addresses the problem of topic independence. In this way, research trends may be viewed considering different levels of hierarchy as well as time and citation impact, and we are retaining this as an idea for future work.

Acknowledgements This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2018S1A3A2075114).

Appendix 1

See Table 10.

Table 10 Topic–word results from text mining dataset at $K=30$

Topic labeling	Weight	top words
(General Topic 1)	0.188	Analysis, technique, process, application, different, paper, present, develop, important, work
(General Topic 2)	0.074	Model, learn, label, domain, feature, task, perform, train, framework, different
Text Classification	0.056	Classification, feature, classify, select, perform, algorithm, machine, base, accuracy, category
Document Clustering	0.055	Document, term, word, similar, measure, base, semantic, extract, keyword, sentence
Semantics and Ontology	0.047	Knowledge, semantic, concept, ontology, extract, domain, base, relationship, relation, discovery
System Development	0.043	System, service, process, tool, application, develop, support, user, design, document
Biomedical corpus annotation	0.041	Biomed, literature, annotation, article, biology, curation, extract, database, abstract, interact
Information retrieval	0.036	Web, search, image, query, page, retrieve, user, content, engine, video
Social network analysis	0.034	Social, media, network, user, twitter, tweet, community, content, online, post
Machine learning	0.031	Cluster, algorithm, document, matrix, set, similar, paper, base, space, factor
Information extraction	0.031	Extract, event, relation, system, sentence, semantic, annotation, corpus, structure, detect
Health informatics	0.030	Clinic, medic, patient, report, record, health, disease, system, electron, care
Sentiment analysis	0.030	Sentiment, analysis, opinion, emotion, polar, posit, neg, express, word, review
Named entity recognition	0.025	Entity, name, system, task, recognition, extract, mention, perform, evaluate, corpus
Natural language processing	0.023	Language, word, nature, Arabic, process, English, translate, question, Chinese, answer
Bioinformatics (pathway analysis)	0.023	Gene, disease, association, cancer, network, express, pathway, cell, identify, analysis
Algorithm and computing	0.023	Algorithm, time, compute, efficient, detect, match, large, parallel, string, perform
Topic model	0.021	Topic, model, latent, stream, document, LDA, time, detect, temporal, Dirichlet
Review analysis	0.021	Review, product, custom, online, consumer, user, opinion, service, rate, market
Recommendation systems	0.017	User, recommend, detect, person, system, email, mobile, spam, filter, base
Market prediction	0.017	News, predict, market, financial, article, stock, analysis, company, price, business
Bibliometrics	0.017	Public, journal, citation, science, literature, article, scientific, author, discovery, country
Patent analysis	0.016	Patent, technology, analysis, student, develop, learn, innovation, education, identify, collaborate
Data mining	0.016	Pattern, rule, associated, frequent, algorithm, discover, set, sequence, paper, tree
Bioinformatics (genomic database)	0.015	Gene, biology, genomic, sequence, function, phenotype, database, bioinformation, tool, annotation

Table 10 (continued)

Topic labeling	Weight	top words
Bioinformatics (interaction database)	0.014	Protein, interact, database, network, function, predict, structure, chemic, mutate, sequence
Software engineering	0.014	Software, report, system, code, bug, develop, project, source, predict, incident
Network computing	0.014	Network, book, compute, model, system, neural, human, table, learn, age
Drug discovery (adverse drug reaction)	0.014	Drug, review, systematic, effect, adverse, identify, clinic, medic, trial, search
Visualization	0.013	Visual, map, geographic, location, spatial, region, explore, analysis, cloud, user

Appendix 2

See Table 11.

Table 11 Topic–word results from LIS dataset at $K = 40$

Topic labeling	Weight	Top words
(General topic 1)	0.0721	Change, technology, community, work, article, way, social, concept, culture, discuss
(General topic 2)	0.0614	Approach, design, model, methodology, value, practice, find, system, framework, origin
Bibliometrics	0.0346	Journal, citation, science, article, public, publish, author, scientific, cite, field
Information management	0.0332	Manage, project, implement, system, process, plan, change, success, organ, practice
Cheminformatics (protein structure)	0.0307	Bind, protein, structure, ligand, dock, molecular, model, interact, inhibitor, energy
Group interview	0.0305	Group, participle, data, interview, user, seek, behaviour, interact, behavior, different
Information theory	0.0302	Estimation, distribution, function, algorithm, error, signal, model, general, sample, optimal
(General topic 3)	0.0299	Measure, perform, indicate, impact, differ, data, rank, evaluate, assess, correlation
System development	0.0298	System, software, compute, data, application, technology, design, user, process, implement
Collection management	0.0290	Electron, collect, book, resource, universe, print, journal, access, academic, publish
Web interfaces	0.0274	Web, search, user, site, engine, page, access, website, tool, online
Librarianship	0.0273	Librarian, academic, profession, educ, li, program, survey, science, role, skill
Information literacy	0.0272	Student, literacy, instruct, course, learn, librarian, universe, skill, faculty, program
Network analysis	0.0268	Network, control, system, algorithm, model, node, optimal, dynamic, time, simul
Digital archive	0.0252	Digit, archive, access, record, catalog, open, repository, collect, preserve, institution
Information theory (channel theory)	0.0252	Channel, capacity, receive, achieve, rate, optimal, interfere, relay, region, gaussian
Cheminformatics	0.0251	Compound, chemic, structure, screen, active, drug, molecule, set, similar, molecular
Information policy	0.0249	Public, govern, policy, community, service, access, e-govern, technology, ICT, nation
Information retrieval	0.0244	Retrieve, query, search, document, algorithm, relevant, user, system, approach, perform
Survey research	0.0230	Universe, survey, find, internet, questionnaire, resource, data, access, respond, Africa
Health communication	0.0228	Health, cancer, women, community, effect, behavior, risk, message, literacy, association

Table 11 (continued)

Topic labeling	Weight	Top words
Scientometrics	0.0228	Patent, country, collaborate, intern, technology, nation, universe, scientific, china, institution
Machine learning	0.0223	Model, data, predict, set, classification, cluster, train, feature, algorithm, learn
Business informatics	0.0218	Business, market, firm, product, company, custom, technology, industry, cost, ERP
Knowledge management	0.0215	Knowledge, manage, share, organ, innovation, organize, process, KM, organism, technology
Information model	0.0214	Factor, model, influence, perceive, user, effect, intent, trust, adopt, relationship
Reference service	0.0213	Service, refer, user, mobile, librarian, academic, universe, technology, patron, virtual
Research evaluation	0.0209	Evaluate, review, quality, question, assess, survey, answer, criteria, effect, test
Social media	0.0197	Social, media, network, community, online, politic, user, blog, content, internet
Information theory (Coding theory 1)	0.0192	Code, construct, sequence, bound, general, linear, function, distance, binary, length
Information theory (coding theory 2)	0.0187	Code, decode, source, error, channel, bound, rate, encode, algorithm, perform
Ontology	0.0186	Data, metadata, semantic, ontology, model, standard, resource, link, object, web
Information ethics	0.0176	Ethic, law, privacy, secure, legal, protect, right, copyright, issue, policy
Geographical information system	0.0174	Spatial, data, model, map, geographic, GIS, area, urban, pattern, location
Natural language processing	0.0147	Language, text, word, extract, term, document, automat, translate, system, annotation
Learning and education	0.0143	Learn, education, school, student, teacher, technology, game, teach, environ, children
Classification	0.0134	Classification, subject, analysis, map, structure, tag, visual, knowledge, document, cluster
Health informatics	0.0127	Health, medic, patient, clinic, care, hospital, nurse, medicine, resource, librarian
Image processing	0.0123	Image, detect, base, video, algorithm, feature, visual, recognition, scheme, segment
Information theory (Quantum)	0.0085	Quantum, channel, secure, entropy, bound, general, scheme, capacity, key, community

Appendix 3

See Table 12.

Table 12 Welch's t-test result between topic coherence of DMR and of g-DMR

		Dataset LIS			Dataset TM		
		avg	std	*p*	avg	std	*p*
$K = 10$	DMR	1.139	0.018	0.151	1.249	0.030	0.009*
	g-DMR	1.152	0.013		1.291	0.018	
20	DMR	1.182	0.035	0.146	1.244	0.022	0.084
	g-DMR	1.157	0.025		1.223	0.020	
30	DMR	1.161	0.030	0.240	1.206	0.016	0.397
	g-DMR	1.176	0.014		1.213	0.015	
40	DMR	1.183	0.024	0.479	1.200	0.015	0.316
	g-DMR	1.175	0.015		1.209	0.017	
50	DMR	1.195	0.011	0.025*	1.202	0.010	0.001*
	g-DMR	1.180	0.012		1.225	0.011	
60	DMR	1.201	0.011	0.054	1.218	0.013	0.230
	g-DMR	1.187	0.014		1.209	0.013	
70	DMR	1.198	0.010	0.989	1.212	0.010	0.738
	g-DMR	1.198	0.014		1.213	0.012	
80	DMR	1.204	0.010	0.140	1.218	0.010	0.842
	g-DMR	1.194	0.014		1.216	0.014	
90	DMR	1.208	0.010	0.389	1.224	0.014	0.066
	g-DMR	1.203	0.010		1.210	0.011	
100	DMR	1.210	0.008	0.018*	1.231	0.011	0.355
	g-DMR	1.199	0.007		1.225	0.011	

The sample size of DMR and g-DMR is 8, and there are only four statistically significant cases (Dataset LIS, $K = 50$ and Dataset TM, $K = 60$, $K = 90$ at $p < 0.05$). As a result, we can say the two models show equivalent performance in topic coherence

References

Andrews, L. C., & Andrews, L. C. (1992). *Special functions of mathematics for engineers*. New York: McGraw-Hill.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning*, (pp. 113–120).

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

Bouabid, H., Paul-Hus, A., & Larivière, V. (2016). Scientific collaboration and high-technology exchanges among BRICS and G-7 countries. *Scientometrics, 106*, 873–899.

Cavacini, A. (2016). Recent trends in Middle Eastern scientific production. *Scientometrics, 109*, 423–432.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Chen, C., Wang, Z., Li, W., & Sun, X. (2018). Modeling scientific influence for research trending topic prediction. In *Thirty-second AAAI conference on artificial intelligence.*

Dietz, L., Bickel, S., & Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on machine learning* (pp. 233–240).

Dou, H., & Kister, J. (2016). Research and development on Moringa Oleifera-Comparison between academic research and patents. *World Patent Information, 47*, 21–33.

Finardi, U., & Buratti, A. (2016). Scientific collaboration framework of BRICS countries: An analysis of international coauthorship. *Scientometrics, 109*, 433–446.

Fukugawa, N. (2016). Knowledge creation and dissemination by Kosetsushi in sectoral innovation systems: insights from patent data. *Scientometrics, 109*, 2303–2327.

Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences, 115*, 3308–3313.

Gerrish, S., & Blei, D. M. (2010). A Language-based Approach to Measuring Scholarly Impact. *ICML, 10*, 375–382.

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems* (pp. 17–24).

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 363–371).

Hawkins, D. T. (2001). Bibliometrics of electronic journals in information science. *Information Research, 7*, 7.

Jabeen, M., Yun, L., Rafiq, M., & Jabeen, M. (2015). Research productivity of library scholars: Bibliometric analysis of growth and trends of LIS publications. *New Library World, 116*, 433–454.

Jo, Y., Hopcroft, J. E., & Lagoze, C. (2011). The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web* (pp. 257–266).

Kang, K., & Sohn, S. Y. (2016). Evaluating the patenting activities of pharmaceutical research organizations based on new technology indices. *Journal of Informetrics, 10*, 74–81.

Kawamae, N., & Higashinaka, R. (2010). Trend detection model. In *Proceedings of the 19th international conference on World wide web* (pp. 1129–1130).

Kim, M., Baek, I., & Song, M. (2018). Topic diffusion analysis of a weighted citation network in biomedical literature. *Journal of the Association for Information Science and Technology, 69*, 329–342.

Li, L.-L., Ding, G., Feng, N., Wang, M.-H., & Ho, Y.-S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics, 80*, 39–58.

Liu, L., & Mei, S. (2016). Visualizing the GVC research: a co-occurrence network based bibliometric analysis. *Scientometrics, 109*, 953–977.

Lv, P. H., Wang, G.-F., Wan, Y., Liu, J., Liu, Q., & Ma, F.-C. (2011). Bibliometric trend analysis on global graphene research. *Scientometrics, 88*, 399–419.

Maisonobe, M., Eckert, D., Grossetti, M., Jégou, L., & Milard, B. (2016). The world network of scientific collaborations between cities: Domestic or international dynamics? *Journal of Informetrics, 10*, 1025–1036.

Mann, G. S., Mimno, D., & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 65–74).

Milanez, D. H., Noyons, E., & Faria, L. I. (2016). A delineating procedure to retrieve relevant publication data in research areas: The case of nanocellulose. *Scientometrics, 107*, 627–643.

Mimno, D., & McCallum, A. (2012). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. arXiv preprint, arXiv:1206.3278.

Moed, H. F. (2016). Iran's scientific dominance and the emergence of South-East Asian countries as scientific collaborators in the Persian Gulf Region. *Scientometrics, 108*, 305–314.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, (pp. 100–108).

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*, 1064–1082.

Sethi, B. B., & Panda, K. C. (2012). Growth and nature of international LIS research: An analysis of two journals. *The International Information & Library Review, 44*, 86–99.

Song, M., Kim, S., & Lee, K. (2017). Ensemble analysis of topical journal ranking in bioinformatics. *Journal of the Association for Information Science and Technology, 68*, 1564–1583.

Song, M., Kim, S., Zhang, G., Ding, Y., & Chambers, T. (2014). Productivity and influence in bioinformatics: A bibliometric analysis using PubMed central. *Journal of the Association for Information Science and Technology, 65*, 352–371.

Stein, M.-K., Galliers, R. D., & Whitley, E. A. (2016). Twenty years of the European information systems academy at ECIS: Emergent trends and research topics. *European Journal of Information Systems, 25*, 1–15.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in neural information processing systems* (pp. 1385–1392).

Timakum, T., Kim, G., & Song, M. (2018). A data-driven analysis of the knowledge structure of library science with full-text journal articles. *Journal of Librarianship and Information Science*. https://doi.org/10.1177/0961000618793977.

Tran, B., Pham, T., Ha, G., Ngo, A., Nguyen, L., Vu, T., et al. (2018). A bibliometric analysis of the global research trend in child maltreatment. *International Journal of Environmental Research and Public Health, 15*, 1456.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. arXiv preprint, arXiv:1206.3298.

Wang, X., & McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 424–433).

Wang, X., Zhai, C., & Roth, D. (2013). Understanding evolution of research themes: a probabilistic generative model for citations. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*, (pp. 1115–1123).

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13*, 100983.

Yan, F., Xu, N., & Qi, Y. (2009). Parallel inference for latent dirichlet allocation on graphics processing units. *Advances in neural information processing systems* (pp. 2134–2142).

Zhang, Y., Chen, K., Zhu, G., Yam, R. C., & Guan, J. (2016). Inter-organizational scientific collaborations and policy effects: An ego-network evolutionary perspective of the Chinese Academy of Sciences. *Scientometrics, 108*, 1383–1415.

Zhao, Y., & Zhao, R. (2016). An evolutionary analysis of collaboration networks in scientometrics. *Scientometrics, 107*, 759–772.

Zhao, Y., Li, D., Han, M., Li, C., & Li, D. (2016). Characteristics of research collaboration in biotechnology in China: Evidence from publications indexed in the SCIE. *Scientometrics, 107*, 1373–1387.

Zou, C. (2018). Analyzing research trends on drug safety using topic modeling. *Expert Opinion on Drug Safety, 17*, 629–636.