Ecological Indicators 60 (2016) 693–701

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

Finding academic concerns of the Three Gorges Project based on a topic modeling approach

Jiang HanChen, Qiang MaoShan*, Lin Peng

State Key Laboratory of Hydroscience and Engineering, Tsinghua University, Haidian, Beijing 100084, China

ARTICLE INFO

Article history:
Received 15 February 2015
Received in revised form 31 July 2015
Accepted 5 August 2015
Available online 27 August 2015

Keywords:
Hydropower project
Three Gorges Project
Topic modeling
Scientific documents
Latent Dirichlet Allocation
Bibliometric indicators

ABSTRACT

The Three Gorges Project (TGP) has gone into the overall completion acceptance stage in 2014. As the world's largest hydropower project, the TGP has attracted worldwide attention over the past few decades. Previous studies mainly focused on a single aspect, such as engineering technologies, social impacts and environmental impacts, of the TGP. However, a large-scale review gathering systematic data to find academic concerns about the TGP is missing. Topic model is a text mining approach for discovering latent topics in a collection of documents. In this article, an emerging topic modeling approach, Latent Dirichlet Allocation (LDA), was introduced to uncover the intellectual structure of the academic literature focusing on the TGP. A collection of 8280 Chinese research articles highly related to the TGP was established with a time frame ranging from 2001 to 2013, and an 18-topic model was used to describe the intellectual structure. Two novel bibliometric indicators, including topic proportion and topic trend, were constructed to describe the academic concerns of the TGP. Topic proportion analysis shows that post-construction issues, including the social and environmental impacts brought by the TGP, have attracted more attention than the construction issues. "Ecology", "Reservoir Operation", "Land Administration", and "Water Pollution", have become the dominant research topics regarding the TGP during these years. Meanwhile, "Construction Technology" and "Design", have gradually lost scholars' interest. The results show that the approach reported in this study can provide sound and credible conclusions of the major academic concerns for a hydropower project. The topic modeling approach is expected to be widely applied as a methodological strategy in future hydropower and other infrastructure project assessment.

1. Introduction

A characteristic development in today's world is the volume and ubiquitous availability of structured as well as unstructured data and their consideration as a resource for mining valuable knowledge (Kulkarni et al., 2014). The tremendous growth of unstructured data sources, especially textual data sources, has inspired the development of new approaches to understand latent intelligence in a variety of events or disciplines (George et al., 2014; Wood et al., 2013). Among these approaches, topic modeling is a powerful text mining method that is able to uncover the latent intellectual structure in textual data, and subsequently provides, researchers and practitioners, with significant support in the broad discipline of decision making.

Information is the basis of decision making. Sufficient information support can effectively reduce risks in decision making (Leitzinger and Stiglitz, 1984). The booming volume of textual data provides a potential to gather more information in a decision making process. Text mining technologies, including topic modeling, have been widely used in decision support tasks. For example, Chen and Tseng (2011) reported a text mining approach to process product reviews from customers and to help business managers make customer-related decisions. Song et al. (2014) proposed a social-media based political decision support system. The system used several text mining techniques, including topic modeling, mention-direction-based network analysis and term co-occurrence retrieval, to process social media textual contents and to help detect and trace the advent of and changes in social issues. In addition to these online user-generated-contents, academic literature is also an important information source for decision support tasks. Nichols (2014) developed a topic modeling approach to measuring interdisciplinary at science foundation. The approach can help science foundation administrators to better understand the content and context of funding portfolio and subsequently promote future

* Corresponding author. Tel.: +86 010 62782027; fax: +86 010 62782027.
E-mail addresses: jhc13@mails.tsinghua.edu.cn (H. Jiang), qiangms@mail.tsinghua.edu.cn (M. Qiang), celinpe@tsinghua.edu.cn (P. Lin).

http://dx.doi.org/10.1016/j.ecolind.2015.08.007
1470-160X/© 2015 Elsevier Ltd. All rights reserved.

science funding plan. With regard to specific discipline, Moro et al. (2015) processed 219 articles focusing on business intelligence in the banking industry with topic modeling. The study found main application trends of technologies in the bank industry and help make reasonable decisions on future research and development projects.

China has got a comprehensive decision making problem in 2014. That is the overall completion acceptance of the Three Gorges Project (TGP). The acceptance is conducted by the Chinese government with the acceptance group leader being the vice premier of the State Council of China. Great attention reflects the complexity and importance of the huge project assessment. For China, the TGP has both symbolic and substantive significance. The TGP is the largest (Chen et al., 2001), most expensive (Wang, 2002) and most powerful (Lu, 1994) hydropower project ever built in the world. Before the construction began in 1994, the project had been discussed and planned for 70 years. The dam is located at the upper reaches of Yangtze River, nearing the city of Yichang, which is the major transportation hub of central China. The large concrete gravity dam raises the normal pool level of the Three Gorges Reservoir to 175 m, creating a 660-kilometer-long and one-kilometer-wide lake along the Yangtze River. And since the first impoundment and the first turbine operation in 2003, the TGP has been providing its major functions including flood control (Stone, 2011), clean power generation (Barros et al., 2011) and waterway navigability improvement (Sutton, 2004). Despite great benefits, major concerns have also been voiced over the negative environmental and social consequences of the TGP (Stone, 2011). Environmental impacts include an increase in geological risks such as earthquakes and landslides (Yang and Lu, 2013), water body pollution (Fu et al., 2010), disruption of the riparian ecosystem (Wu et al., 2003), and unstable reshaping of the whole Yangtze River system (Yang et al., 2011). The social impacts mainly result from the 1.2 million involuntary immigrants (Tullos, 2009). Their survival and developmental issues are of significance to the project success and the stability of the regional society. As a consequence, comprehensive evaluation of the above issues of the TGP will help Chinese government in making more reasonable and democratic decisions in the overall completion acceptance of the TGP.

Scientific literature is a valuable and rich source of knowledge. Researchers publish a large number of articles which reflect the state of the art in their respective disciplines. Previous studies have used thematic reviews, manual content analysis and citation–cocitation analysis to perform literature review work. However, these approaches are considered to be time-consuming, subjective and difficult to process vast amounts of data (Kulkarni et al., 2014). In order to improve the efficiency of processing scientific documents, recently, text mining approaches, such as topic modeling, have been introduced into scientometric and bibliometric studies (Nichols, 2014; Yau et al., 2014) to help find the intellectual structure of a discipline. These text mining approaches provide a potential of investigating the underlying intellectual structure of the academic concerns toward a huge and complex hydropower project like the TGP. And with regard to the TGP, researchers, especially Chinese researchers, have indeed published numerous articles discussing scientific, technical and social problems related to the project during these decades. The goal of this article is to introduce topic modeling for revealing the intellectual structure of the academic research of the TGP. The topic modeling results provide researchers and practitioners with a brief and accurate summary of the academic concerns of the TGP, and subsequently promote the rationality of decision makings in the overall completion acceptance of the project. In general, this article intends to answer the following scientific questions. What are the major topics that the TGP research articles focus on? How much is the proportion of each topic in all literature? And do these identified research topics show significant trends over the project life cycle?

2. Latent Dirichlet Allocation

Topic models are regarded as statistical or probabilistic models for uncovering the underlying intellectual structure of a collection of documents based on the following assumptions: (1) words are exchangeable in a document; (2) a topic is modeled as a multinomial distribution on words from a vocabulary; and (3) a document is composed of words from some different topics (Blei et al., 2003). Topic models aim at building the generative process of a document from a probabilistic perspective. Considering that a given document contains T topics over a vocabulary of V terms, the probability that a word w instantiates term v in the document can be calculated as follows:

$$P(w=v)=\sum_{j=1}^{T}P(w=v|z=j)P(z=j) \tag{1}$$

where z is a latent variable indicating the topic from which the word w was drawn, $P(w=v|z=j)$ is the probability that w instantiates terms v in the latent topic $z=j$ and $P(z=j)$ is the probability of the jth topic appearing in the document. If a word w has a high value of $P(w|z)$, it would be an important or representative word in topic z. And if a topic z has a high value of $P(z)$, it would be a dominant topic in the given document. Based on regarding documents as mixtures of probabilistic topics, the problem of revealing the set of topics that are used in a collection of documents can be formulated as fitting a probability model. Supposing that there are D documents containing T topics using V unique terms, the main objectives of topic model inference are: to find (1) the term distribution $P(w|z=j)=\phi_j$ for each topic j and (2) the topic distribution $P(z|d=m)=\theta_m$ for each document m. The estimated parameter sets $\Phi=\{\phi_j\}_{j=1}^{T}$ and $\Theta=\{\theta_m\}_{m=1}^{D}$ are the basis for latent semantic representation of words and documents. In order to estimate Φ and Θ for a given collection of documents, Hofmann (2001) proposed to maximize $P(w|\phi,\theta)$ directly by using the Expectation-Maximization (EM) algorithm to find maximum likelihood estimates of ϕ and θ. However, this EM algorithm may cause overfitting and is slow to converge, encouraging new models that make assumption about the source of ϕ and θ.

Latent Dirichlet Allocation (LDA) is one such model, introducing prior probability distributions (Dirichlet distribution) both on ϕ and θ. In order to give a clear review of LDA, additionally, let $\mathrm{Dir}_T(\alpha)$ denote a Dirichlet distribution over T topics with a parameter α, and $\mathrm{Dir}_V(\beta)$ denote a Dirichlet distribution over V unique terms with a parameter β. The generative process for a collection of documents is as follows:

For each topic $j\in[1, T]$, sample multinomial distribution $\phi_j\sim\mathrm{Dir}_V(\beta)$.
For each document $m\in[1, D]$, sample multinomial distribution $\theta_m\sim\mathrm{Dir}_T(\alpha)$, and sample document length $N_m\sim\mathrm{Poisson}(\xi)$.
For each word $n\in[1, N_m]$ in document m, sample topic index $z_{m,n}\sim\mathrm{Multinomial}(\theta_m)$, and sample term for word $w_{m,n}\sim\mathrm{Multinomial}(\phi_{z_{m,n}})$.

According to the generative process, the complete-data likelihood of a document can be specified using a joint distribution of all known and hidden variables, given the hyperparameters (α and β), as follows:

$$p(w_m,z_m,\theta_m,\Phi|\alpha,\beta)=\prod_{n=1}^{N_m}p(w_{m,n}|\phi_{z_{m,n}})p(z_{m,n}|\theta_m)p(\theta_m|\alpha)p(\Phi|\beta) \tag{2}$$

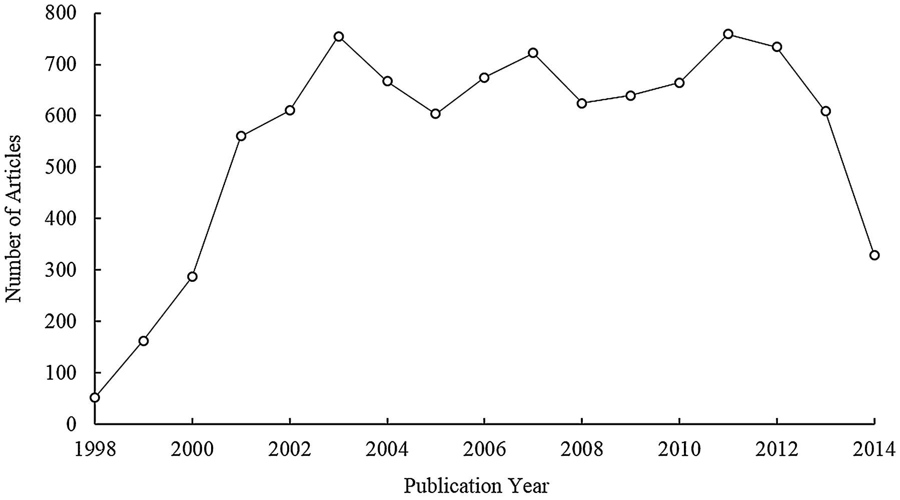


Fig. 1. Annual numbers of TGP research articles during 1998–2014.

Table 1
Professional dictionaries used in the segmentation process.

Dictionary name	Domain	Number of words
Construction bureau	Civil engineering	145
Environmental protection	Environmental engineering	2643
Information of Yichang city	Place name	8543
Hydraulic engineering dictionary	Hydraulic engineering	28,389

So the probability that a word $w_{m,n}$ instantiates a particular term v, given that the LDA parameters (θ and Φ), is obtained by marginalizing the latent variable $z_{m,n}$ and omitting the hyperparameters as follows:

$$p(w_{m,n} = v|\theta_m, \Phi) = \sum_{j=1}^{T} p(w_{m,n} = v|\phi_j)p(z_{m,n} = j|\theta_m) \tag{3}$$

Note that for different m and n, the token observations $w_{m,n}$ are independent events, therefore, the likelihood of the corpus $W = \{w_m\}_{m=1}^{D}$ can be calculated as follows:

$$p(W|\Theta, \Phi) = \prod_{m=1}^{D} p(w_m|\theta_m, \Phi) = \prod_{m=1}^{D}\prod_{n=1}^{N_m} p(w_{m,n}|\theta_m, \Phi) \tag{4}$$

The task of estimating parameters Φ and Θ in LDA can be accomplished using statistical techniques such as variation EM algorithm (Blei et al., 2003) and Gibbs sampling (Griffiths and Steyvers, 2004). The latter technique solves the estimation problem by using a Monte Carlo procedure, resulting in a simple and practicable implementation, requires few computing resources and has been used in this study.

Recently, topic modeling and other text mining approaches are becoming more approachable as the availability of accessible software enables researches to take advantage of these methods. Commercial software packages include SAS Text Miner and SPSS Clementine. Open source options include some relevant R, Python, and Java packages. In this study, the application is based on the R package, Topicmodels, supporting LDA modeling and Gibbs sampling, offered by Hornik and Grün (2011). Topicmodels package requires a text mining front-end addition, such as another R package, tm, offered by Meyer et al. (2008). And this article also make use of MS excel to assist in processing statistical data and plotting figures.

3. Research method

In this section, LDA is employed to reveal the intellectual structure of TGP research articles. Technical details of the analysis process appear as follows.

3.1. Data collection

The data collection process focused on searching academic articles in the major journals. We used “Three Gorges” as a keyword to search articles in a Chinese journal database (China Online Journal database) and an international Journal database (Science Citation Index Expanded and Social Science Citation Index), respectively. We found that regarding TGP, Chinese articles were much more abundant than articles in other languages (mainly English). As performing topic modeling with multiple languages is difficult and translating thousands of articles to unify language is also impractical, in this study, the more abundant data source, Chinese articles, is chosen. The employed academic database in this study is the China Online Journals (COJ) database. One reason for choosing COJ as the data source is that it has a rich data volume. The COJ is provided by WanFang Data (www.wanfangdata.com.cn) and includes over 7000 full-text Chinese academic journals from the year 1998 up to now. Another reason is that COJ supports batch export of article information, including the title, abstract, keywords, and publication date, providing a great convenience for the data collection process. We accessed the WanFang Data in November 2014 for performing the data collection task. The search criteria included two parameters: keywords and publication date. (1) Keywords: as Chinese people discuss the TGP or a part of the TGP with several terms, a series of words, including “Three Gorges Project”, “Three Gorges Dam”, “Three Gorges Reservoir”, “Three Gorge Hydropower Station”, “Three Gorges Power Station”, “Three Gorges Reservoir Area”, “Three Gorges Hydro-junction” and “Three Gorges Hydro-project”, were selected as keywords for search. If an article recorded by COJ contains any of the above words in its title or author keywords, it was included in this study. (2) Publication date: based on the “Keywords” search criterion, 9455 articles focusing on the TGP were obtained from COJ since 1998. The annual numbers of articles published from 1998 to 2014 are shown in Fig. 1. The number of articles published in 1998, 1999, 2000 and 2014 were significantly less than those in other years. As this study aimed to perform a research trend analysis during consecutive years, in order to reduce any influence caused by non-uniformity of the annual data, the time frame was defined as a 13-year period from 2001 to 2013. This time frame, including the major construction period (before 2006) and the operation period (after 2003), will help find different academic concerns of the TGP. In total, the search process yielded 8625 articles.

Using abstract texts as input data for topic modeling is a common strategy in previous studies which aim at finding topics in scientific articles (Griffiths and Steyvers, 2004; Kim and Yoon, 2014; Kulkarni et al., 2014; Zheng et al., 2006). Following this strategy, abstracts of the collected articles were used as the input data for topic modeling in this study. Hence, articles without a Chinese abstract were filtered out and the final data set contained 8280 article abstracts.

3.2. Preprocessing

As Chinese texts do not segment words by spaces, the collected abstracts needed to be preprocessed with segmentation method before modeling. The segmentation was executed by an existing tool: the Institute of Computing Technology Chinese Lexical Analysis System (ICTCLAS) (Zhang et al., 2003). In view of the fact that the TGP research articles employ substantial terminologies of hydraulic engineering, civil engineering, and environmental engineering, and special place names in the local area, four Chinese professional dictionaries downloaded from Sougou Lab (http://www.sogou.com/labs/dl/w.html/) were imported into the segmentation system to improve the accuracy. Information on the four dictionaries is shown in Table 1. After segmentation, common stop words, indicating some frequent but trivial terms such as “a”, “of”, “the”, “and”, etc., were excluded, because they carry little information.

In order to build an input with a proper data structure that LDA could process, a document-term matrix (DTM) was established to present the data set based on Inverse Document Frequencies (TF-IDFs transformation), which reduces the weights of frequent terms appearing in many documents and promotes the weights of rare terms (Robertson, 2004). Terms with a single Chinese character or obtaining a TF-IDF value less than 0.1 were removed to reduce the input and finally got an 8280 × 62,996 DTM (8280 documents and 62,996 terms), denoted as **M**.

3.3. Parameter regulation

The LDA model is conditioned on three parameters: the Dirichlet hyperparameter α and β and the number of topics T. In order to fit the model, this study followed the strategy proposed by Griffiths and Steyvers (2004), who suggested to fix $\alpha = 50/T$ and $\beta = 0.1$ and explore the consequences of varying T. For such a Bayesian probabilistic model, the common practice to make a choice in a set of models is to compute the posterior probability of that set of models given the observed data. In this case, the observed data is the DTM **M**, and the models are specified by different values of T.

Previous studies have shown that using only statistical measures to estimate the parameter of a LDA model may be less semantically meaningful (Chang et al., 2009; Grimmer and Stewart, 2013; Levy and Franklin, 2013). Using manual interpretation and inspection to fit the model can maximize topic interpretability, but it will be nontrivial. Hence, a trade-off was used, which referred to a two-round estimation procedure, to determine an acceptable parameter T in terms of both probability and interpretability. The first round estimation was merely from the probabilistic perspective. Estimates of $P(\mathbf{M}|T)$ with different T values from 2 to 50 were calculated to find a local maximum. The value range of T was relatively small, as this study aimed to find mesoscale topics in the data set. In all cases, the log-likelihood values are shown in Fig. 2. The results suggested that the data were best accounted for by a model incorporating 21 topics.

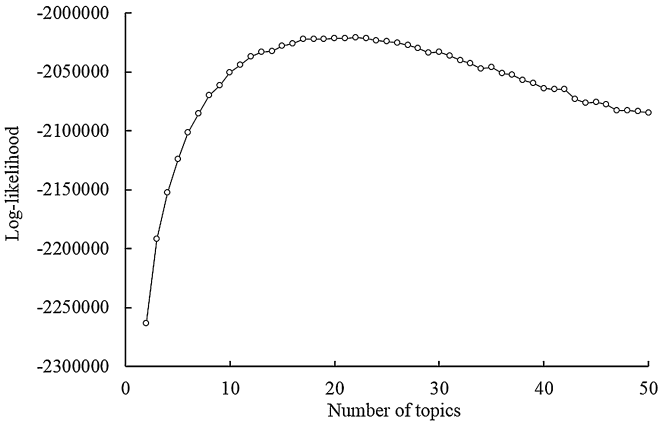


Fig. 2. First round parameter estimation results.

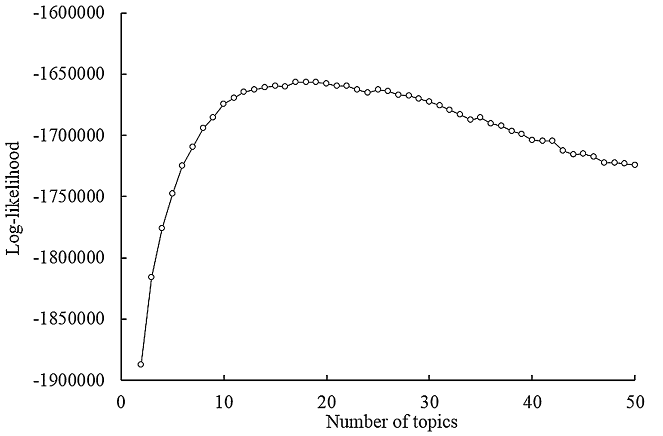


Fig. 3. Second round parameter estimation results.

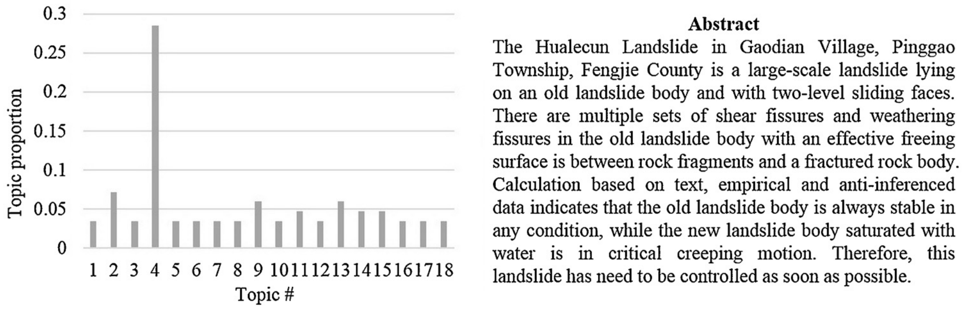
The second round estimation added an interpretability perspective. The term list of each topic generated by the 21-topic model was established and interpreted into a substantive issue related to the TGP. It was found that there were three noisy topics which referred to unspecific issues of the TGP. In other words, these three topics were composed of high-frequency terms including “find”, “result”, “investigate”, “relation”, “work”, “case”, etc. These terms are widely used in common scientific documents and contribute little professional information of any specific issue regarding the TGP. Hence, the stop word list was updated by adding these 275 recognized trivial words. The preprocessing step was repeated to establish a new DTM. The new DTM was denoted as $\mathbf{M}'$, with 8280 documents and 60,472 terms. Similarly, estimates of $P(\mathbf{M}'|T)$ with different T values from 2 to 50 were calculated and a maximal log-likelihood value was reached at $T = 18$, as shown in Fig. 3. After manual inspection, the 18-topic model was accepted.

3.4. Topic label assignment and indicator establishment

Table 2 lists the 18 topics and the corresponding five high-frequent and representative terms for each topic. Based on our prior knowledge of the TGP, a label (topic name) was assigned to each topic by manual examinations. These labels are presented in Table 2 under the topic number.

Table 2
TGP research topics with five high-frequency and representative terms.

Topic 1 Immigration	Topic 2 Reservoir Operation	Topic 3 Construction Technology	Topic 4 Geological Disaster	Topic 5 Comprehensive Benefit	Topic 6 Water Pollution
Development	Reservoir	Construction	Geology	Benefits	Reservoir
Economy	Impoundment	Engineering	Disaster	Power generation	Water body
Society	Influence	Quality	Landslide	Navigation	Water quality
Immigrant	Water level	Technology	Deformation	Flood control	Impoundment
Industry	Variation	Design	Prevention	Yangtze river	Pollution
Topic 7 Design	Topic 8 Land Administration	Topic 9 Construction Management	Topic 10 Ecology	Topic 11 People's Livelihood	Topic 12 Achievement
Power station	Type	Management	Plant	Investigation	Engineering construction
Design	Space	Construction	Grow	Health	Nation
Ship lock	Land	Reinforce	Density	Resident	Project
Power unit	Area	Improve	Species	Prevention	Task
Device	Region	Mechanism	Variation	Daily life	Completion
Topic 13 Modeling	Topic 14 Monitoring System	Topic 15 Resource Conservation	Topic 16 Regional Research	Topic 17 Navigation	Topic 18 Industrial Research
Model	System	Protection	Chongqing city	Ship	Production
Calculate	Monitoring	Resources	Region	Navigation	Model
Forecast	Data	Activity	Wanzhou city	Yangtze river	Proper
Simulation	Implementation	Nature	City	Improvement	Plant
Parameter	Function	Abundant	Central	Stable	Manual



Abstract
The Hualecun Landslide in Gaodian Village, Pinggao Township, Fengjie County is a large-scale landslide lying on an old landslide body and with two-level sliding faces. There are multiple sets of shear fissures and weathering fissures in the old landslide body with an effective freeing surface is between rock fragments and a fractured rock body. Calculation based on text, empirical and anti-inferenced data indicates that the old landslide body is always stable in any condition, while the new landslide body saturated with water is in critical creeping motion. Therefore, this landslide has need to be controlled as soon as possible.

Fig. 4. A collected abstract and diagnostic topic proportions for the Abstract.

The fitted model assigned topic proportions to each abstract in the whole corpus. For example, Fig. 4 presents the text of a collected abstract and the diagnostic proportions of all the 18 topics in the text. It can be found that topic 4, labeled as "Geological Disaster", is the most dominant topic for the abstract. Referring to the text, it can be known that the abstract indeed discusses the geological disaster (landslide) in the Three Gorges Reservoir area. For each topic, its proportions assigned to all abstracts were integrated, and subsequently the overall topic proportion for the whole corpus was obtained. In addition, as the collected textual data contained the publication date of each article, annual proportions from 2001 to 2013 of each topic were also calculated, generating a topic proportion time series. Mann–Kendall test (Mann, 1945), which is a nonparametric trend test, was used to determine whether a trend is increasing or decreasing. Based on these results, two bibliometric indicators were established for further explanations of these identified topics. One indicator is the overall topic proportion, which reflects the distribution of the TGP related academic concerns over the whole Chinese academic circle. The other indicator is the topic trend, which presents the time variations of topic prevalences. It should be noted that, the topic label assignment is, to some extent, subjective, and the established indicators also need further explanations. Thus, detailed interpretations of key topics and further analysis of topic proportions and trends are provided in Sections 4 and 5.

4. Results

Results are given to answer the scientific questions proposed in Section 1. First, some key research topics are shown in detail. Then the two established bibliometric indicators, including overall topic proportions and topic trends, are further described.

4.1. Descriptions of key topics

Some of these topics in Table 2 apparently refer to existing hydraulic engineering subjects. For instance, topic 4 contains words such as "geology", "disaster", "landslide", "deformation", "prevention", "earthquake", and "failure" and thus clearly pertains to the geological disaster issues of the TGP. These terms are of great subject specialty, and the fitted model gathered them into one topic with high probabilities. A representative abstract, to which the fitted model has assigned high proportion (33.4%) of topic 4, reads as follows: "*Huangtupo landslide is one of the largest landslide in the Three Gorges Reservoir area, where the Badong field laboratory is built by Three Gorges Research Center for Geo-hazards, Ministry of Education because it is typical and significant. The geological structure characteristics and formation mechanism of Huangtupo landslide can be clarified by exposing slope structure through the text tunnel excavation . . .*" (Jian and Yang, 2013). Definitely, the article of this abstract discusses the geological disaster issues related to the TGP.

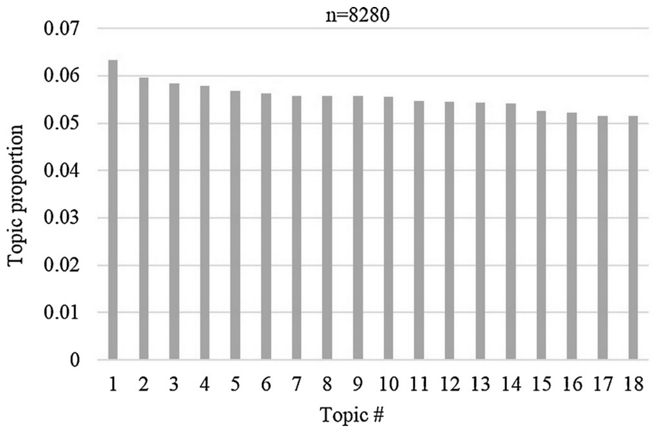


Fig. 5. Overall topic proportions for the whole corpus.

As a matter of fact, most of the topics in Table 2 focus on existing subjects and their summary phrases are also comprehensible.

However, some topic interpretations in Table 2 need further explanations, as terms in these topics are scattered amongst different subjects. For example, topic 5 contains words like “benefits”, “power generation”, “navigation”, “flood control”, “Yangtze River”, and “hydro-junction”. These terms do not belong to a concentrated subject. Hence, some abstracts, to which the model has given high proportions of topic 5, were examined. One abstract reads as follows: “*The Three Gorges Project on the Yangtze River is the largest water power station in the world; it is now attracting the world wide attention. Possessing comprehensive utilization benefits mainly for flood control, power generation and navigation improvement, TGP will be a vital important and backbone project in harnessing and developing of the Yangtze River*...” (Wang, 2009). Based on analyzing high-frequency terms and representative abstracts of topic 5, it can be known that the topic indicates the multiple functions or benefits of the TGP. Hence, topic 5 was named as “Comprehensive Benefit”. The same interpreting procedure was applied to topic 11 (People’s Livelihood), topic 12 (Achievement) and topic 16 (Regional Research).

4.2. *TGP research topic proportions*

Fig. 5 shows the calculated overall topic proportions for the whole corpus. The topic numbers in Table 2 are sorted by their proportions for the whole corpus. That is, based on the results of the fitted model, the five highest-frequency topics are “Immigration (6.3%)”, “Reservoir Operation (6.0%)”, “Construction Technology (5.8%)”, “Geological Disaster (5.8%)”, and “Comprehensive Benefits (5.7%)”, while the five lowest-frequency research topics are “Monitoring System (5.4%)”, “Resource Conservation (5.3%)”, “Regional Research (5.2%)”, “Navigation (5.1%)” and “Industrial Research (5.1%)”.

In order to present a more general classification result, these topics were categorized into three main categories named “Engineering Technology”, “Social Impact” and “Environmental Impact”, respectively. The classification is shown in Table 3. The result reveals that the “Engineering Technology” and “Social Impact” are more frequently discussed in the corpus, with the same proportion of 38.9%, and the “Environmental Impact” is less frequently discussed with a proportion of 22.2%.

4.2.1. *TGP research topic trends*

Fig. 6 presents the annual trends of the 18 TGP research topics during 2001–2013 respectively. Mann–Kendall test (Mann, 1945), a nonparametric trend test, was used to examine whether increasing or decreasing trends existed in the 18 topics. The result is consistent with the idea that research shows strong trends with topics rising and falling regularly in prevalence (Griffiths and Steyvers, 2004). Four of the topics, including “Ecology”, “Reservoir Operation”, “Land Administration” and “Water Pollution”, show a statistically significant increasing trend, and two of the topics, including “Construction Technology” and “Design”, show a statistically significant decreasing trend, both at the $P = 0.0005$ level. As can be seen from Fig. 6, some topics, such as “Navigation”, “Industrial Research” and “Resource Conservation”, show a stable trend during 2001–2013. Meanwhile, some topics, such as “Geological” and “Immigration”, show a trend with relatively sharp fluctuations. And “People’s Livelihood” topic shows an increasing trend before 2007 and a decreasing trend after 2007. These research trends reflect varying research interests related to the TGP in the Chinese academic circle.

Table 3
Three-category classification of the 18 topics.

Category	Proportion	Topic
Engineering Technology	38.9%	Topic 2 Reservoir Operation Topic 3 Construction Technology Topic 7 Design Topic 9 Construction Management Topic 13 Modeling Topic 14 Monitoring System Topic 17 Navigation
Social Impact	38.9%	Topic 1 Immigration Topic 5 Comprehensive Benefit Topic 8 Land Administration Topic 11 People’s Livelihood Topic 12 Achievement Topic 16 Regional Research Topic 18 industrial Research
Environmental Impact	22.2%	Topic 4 Geological Disaster Topic 6 Water Pollution Topic 10 Ecology Topic 15 Resource Conservation

5. Discussion

Previous studies (Coglianese, 2004; Shulman et al., 2003) have suggested that social scientists and decision makers can benefit from large-scale textual data with proper text mining methods which help promote public participation and democratic processes. Using topic modeling to improve decision makings in public issues has also been widely reported (Shulman et al., 2003; Roberts et al., 2014). The commonly used data are online public comments, which are usually unsophisticated (Cuéllar, 2005), subjective and less likely to have substantive impact on rules (Steelman, 1999). However, the data source used in this study is the scientific literature, which is more condensed, informative and objective than online public comments, causing that the analysis results in this article are indeed valuable for comprehensive understanding of the TGP.

5.1. *Implications of topic modeling results*

5.1.1. *Most urgent topic of the TGP*

In order to demonstrate that the topic modeling results are in accord with the actual development of the TGP, further discussion of some dominant topics is presented. Involuntary migration is the biggest social impact brought by the TGP. The Three Gorges Reservoir, with a normal water storage level at 175 m, completely or partially flooded 13 cities and towns, 365 townships, and 1711 villages in 20 countries in Hubei and Chongqing Provinces (Hwang et al., 2011), causing that 25.9 thousand hectares of farmland was lost and at least 1.2 million residents (59% urban and 41% rural) were relocated (Duan and Steil, 2003; Gleick, 2009). The cost of the involuntary migration is 86 billion Chinese Yuan, accounting

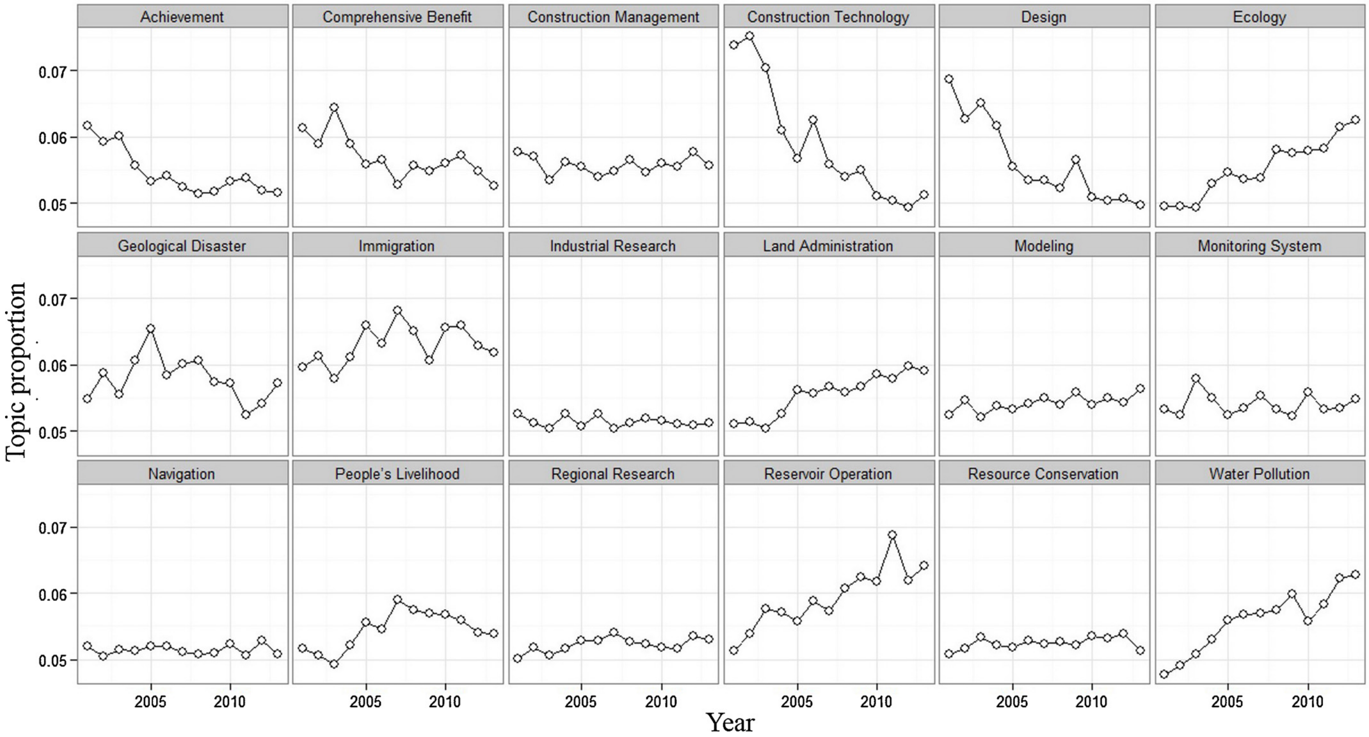


Fig. 6. TGP research topic trends during 2001–2013.

for 41% of the total cost of the project. Both the project company and the government have been suffering from the involuntary migration. The immigrants also face challenges from development, economy and society issues. Correspondingly, the "Immigration" research topic has the biggest proportion over the 18 topics. The extracted high-frequency terms, including "development", "economy" and "society", can reflect the major issues of immigrants. The trend analysis shows that the prevalence of the "Immigration" topic has maintained at a high level, which is in accord with the long period (from 1999 to 2009) of the migration process and the continuous reverse migration events (Wang and Shi, 2004) in recent years. These analysis results indicate that the project company and the government should continue paying attention to the migration issue of the TGP. In particular, they should help the immigrants adapt to life in new environments, learn new skills and archive self-development, rather than just give them money, which may result in reverse migration.

5.1.2. University–industry–research cooperation mode of the TGP

The university–industry–research cooperation mode of the TGP is typical in China (Chen, 1999). The topic modeling results reveal some characteristics of the mode. The huge number of TGP research articles reflects that Chinese researchers have been supported by abundant industrial and national funding for the TGP. The research topic trends reveal that the TGP has faced different key problems during its whole life cycle. For instance, "Reservoir Operation" and "Construction Technology" are second and third dominant topics respectively. However, they show entirely different trends with each other. "Reservoir Operation" presents a significant increasing trend, while "Construction Technology" presents a significant decreasing trend. The reason is obvious. The construction of the major part of the TGP has been gradually completed since 2003, causing a decreasing demand of construction technology research. However, the operation of the reservoir just started in 2003, leading to more research in related domains. In addition, recent extreme climate events in the Yangtze Valley (Dai et al., 2010) also result in an increasing demand of research on the "Reservoir Operation" topic. The decreasing topics include "Construction Technology" and "Design", which are construction engineering issues. The increasing topics include "Ecology", "Reservoir Operation", "Land Administration", and "Water Pollution", which are post-construction issues. These trends reveal that the focus of the academic concerns, as well as the key challenges of the TGP, have moved from engineering technologies to post-construction issues, suggesting that the project company and the government should put more resources into these post-construction issues. Currently, China's hydropower development is still in its peak period (Chang et al., 2010; Zhao et al., 2012). Many large hydropower projects, such as Xiluodu Project, Baihetan Project and Motuo Project, being of the same order of magnitude as the TGP, are under or will be under construction. The recognized research trends of the TGP will provide valuable experience for these projects making their research and development plan.

5.1.3. Support for building assessment indicator frameworks

The topic modeling approach also supports the establishment of indicator frameworks for assessment tasks. The main feature of indicators is their ability to summarize, focus and condense the enormous complexity of the dynamic environment to a manageable amount of meaningful information (Singh et al., 2009). Taking advantage from existing literature is a common used strategy for item identification in building an indicator framework. However, manual inspection is tedious and impossible to cover the large volume of literature. Thus, the topic modeling approach in this study can efficiently propose important items which have been recognized by previous studies for building an indicator framework. And the proposed bibliometric indicators, including the topic proportion and topic trend, can describe the prevalence and temporal variation of a topic, adding weights to the indicator framework. Hence, the proposed topic modeling approach and bibliometric indicators can be extended to a useful and innovative tool for improving assessment practices in future.

5.2. Limitations and future work

However, any new form of research faces complications and limitations in execution. First, although a two-round and semi-automated analysis process was employed, the proposed method is still somewhat subjective and non-trivial. Updating the stop word list in the parameter regulation process, and more important, the interpreting and naming the latent topics both rely on some manual inspection. Second, the application reported in this article only used Chinese textual data, which may cause a loss of information from the international perspective. The methodology in this article can perform any monolingual task. But multilingual tasks require more complicated preprocessing, such as automatic translation, which is not easily accessible at present. Third, when a project did not attract sufficient attention of researchers, there is no guarantee that the LDA will always uncover all important and emerging topics for the project, simply because of the lack of available data. Nevertheless, if the project is not that important to draw the attention of researchers, there is no need to evaluate such a project in that much detail by the proposed method. And if it is the case that the project is important, but the academic community has not been aware of the importance of the project, it is necessary to wait for the long gestation period of academic research in terms of problem formulation to commissioning. Hence, this is a broader limitation of any study like ours.

In future research, multiple data sources, such as social media comments, online news, official reports and project internal documents, will be used to extract different or shifted topics and foci, which can be compared with the results in this article. In addition, as the TGP is a world-famous hydro project, international (non-Chinese) researchers have also published a certain number of research articles discussing problems related to the TGP. The intellectual structure of these international articles can provide a more comprehensive perspective on understanding the TGP related issues. And future research will also investigate other large hydro projects in other countries with different natural conditions, public needs and political policies.

6. Conclusions

In this study, a detailed analysis method is presented to gain insight into the content of scientific articles related to the Three Gorges Project. The topics uncovered by the proposed topic modeling approach present meaningful aspects of the intellectual structure of the project related research. The analysis results, including the contents, proportions and trends of the diagnostic topics, conclude the research profile of the TGP. The identified 18 research topics can be regarded as the major academic concerns of the TGP in the Chinese academic circle. Topic proportion analysis shows that the post-construction issues, including the social and environmental impacts brought by the TGP, have attracted more attention than the construction issues. Specifically, topics, such as “Ecology”, “Reservoir Operation”, “Land Administration”, and “Water Pollution”, which show a significant increasing trend, reveal the dominant thematic hotspots of the TGP research during recent years. These results provide valuable information supports for further decision making and research tasks of the TGP. This study demonstrates that topic modeling is a useful and innovative tool for discovering semantic topics in a large corpus related to a world-famous infrastructure project. As a statistical model, the numerical results generated by topic modeling can be easily processed with corresponding semantics to establish novel indicators for different purposes. Hence, we expect topic modeling to gain traction as a methodological strategy for hydropower and other infrastructure project assessment in future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 51479100, 51179086, 11272178, and 51379104; and State Key Laboratory of Hydroscience and Engineering of China under Grant 2015-KY-5 and 2013-KY-5.

References

Barros, N., Cole, J.J., Tranvik, L.J., Prairie, Y.T., Bastviken, D., Huszar, V.L., del Giorgio, P., Roland, F., 2011. Carbon emission from hydroelectric reservoirs linked to reservoir age and latitude. Nat. Geosci. 4 (9), 593–596.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M., 2009. Reading tea leaves: how humans interpret topic models. In: Proceedings of the Neural Information Processing Systems 2009, Vancouver, BC, Canada, December 7–10, 2009, pp. 288–296.

Chang, X., Liu, X., Zhou, W., 2010. Hydropower in China at present and its further development. Energy 35 (11), 4400–4406.

Chen, C.C., Tseng, Y.D., 2011. Quality evaluation of product reviews using an information quality framework. Decis. Supp. Syst. 50 (4), 755–768.

Chen, S.H., 1999. Mode of basic research combined with practice. Sci. Found. China 5, 313–314 (in Chinese).

Chen, Z., Li, J., Shen, H., Wang, Z., 2001. Yangtze River of China: historical analysis of discharge variability and sediment flux. Geomorphology 41 (2), 77–91.

Coglianese, C., 2004. Information technology and regulatory policy new directions for digital government research. Soc. Sci. Comput. Rev. 22 (1), 85–91.

Cuéllar, M.F., 2005. Rethinking regulatory democracy. Admin. Law R 57, 411–499.

Fu, B.J., Wu, B.F., Lu, Y.H., Xu, Z.H., Cao, J.H., Niu, D., Yang, G.S., Zhou, Y.M., 2010. Three Gorges Project: efforts and challenges for the environment. Prog. Phys. Geogr. 34, 741–754.

Dai, Z.J., Du, J.Z., Chu, A., Li, J.F., Chen, J.Y., Zhang, X.L., 2010. Groundwater discharge to the Changjiang River, China, during the drought season of 2006: effects of the extreme drought and the impoundment of the Three Gorges Dam. Hydrogeol. J. 18 (2), 359–369.

Duan, Y., Steil, S., 2003. China Three Gorges Project resettlement: policy, planning and implementation. J. Refugee Stud. 16 (4), 422–443.

George, G., Haas, M.R., Pentland, A., 2014. Big data and management. Acad. Manage. J. 57 (2), 321–326.

Gleick, P.H.,2009. Three Gorges dam project, Yangtze River, China. In: The World's Waters 2008–2009. Island Press, Washington, DC, pp. 139–150.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. 101 (Suppl. 1), 5228–5235.

Grimmer, J., Stewart, B.M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 21, 267–297.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42 (1-2), 177–196.

Hornik, K., Grün, B., 2011. Topicmodels: an R package for fitting topic models. J. Stat. Softw. 40 (13), 1–30.

Hwang, S.S., Cao, Y., Xi, J., 2011. The short-term impact of involuntary migration in China's Three Gorges: a prospective study. Soc. Indic. Res. 101 (1), 73–92.

Jian, W., Yang, J., 2013. Formation mechanism of No.1 part slide of Huangtupo landslide in the Three Gorges Reservoir Area. J. China Univ. Geosci. (Earth Sci.) 38 (3), 625–631 (in Chinese).

Kim, S., Yoon, J., 2014. Link-topic model for biomedical abbreviation disambiguation. J. Biomed. Inform. 53, 267–380.

Kulkarni, S.S., Apte, U.M., Evangelopoulos, N.E., 2014. The use of latent semantic analysis in operations management research. Decis. Sci. 45 (5), 971–994.

Leitzinger, J.J., Stiglitz, J.E., 1984. Information externalities in oil and gas leasing. Contemp. Econ. Policy 2 (5), 44–57.

Levy, K.E., Franklin, M., 2013. Driving regulation: using topic models to examine political contention in the US trucking industry. Soc. Sci. Comput. Rev. 32, 182–194.

Lu, Y.M., 1994. Three Gorges Project: a progress report. Int. Water Power Dam Constr. 46 (8), 20–23.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica 13, 245–259.

Meyer, D., Hornik, K., Feinerer, I., 2008. Text mining infrastructure in R. J. Stat. Softw. 25 (5), 1–54.

Moro, S., Cortez, P., Rita, P., 2015. Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst. Appl. 42 (3), 1314–1324.

Nichols, L.G., 2014. A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics 100 (3), 741–754.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural topic models for open-ended survey responses. Am. J. Polit. Sci. 58 (4), 1064–1082.

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. J. Document. 60 (5), 503–520.

Shulman, S.W., Schlosberg, D., Zavestoski, S., Courard-Hauri, D., 2003. Electronic rulemaking a public participation research agenda for the social sciences. Soc. Sci. Comput. Rev. 21 (2), 162–178.

Singh, R.K., Murty, H.R., Gupta, S.K., Dikshit, A.K., 2009. An overview of sustainability assessment methodologies. Ecol. Indic. 9 (2), 189–212.

Steelman, T.A., 1999. The public comment process: what do citizens contribute to national forest management? J. Forest. 97 (1), 22–26.

Song, M., Kim, M.C., Jeong, Y.K., 2014. Analyzing the political landscape of 2012 Korean presidential election in Twitter. IEEE Intell. Syst. 29 (2), 18–26.

Stone, R., 2011. The legacy of the Three Gorges dam. Science 333 (6044), 817.

Sutton, A., 2004. The Three Gorges Project on the Yangtze River in China. Geography 89 (2), 111–126.

Tullos, D., 2009. Assessing the influence of environmental impact assessments on science and policy: an analysis of the Three Gorges Project. J. Environ. Manage. 90, S208–S223.

Wang, J., 2002. Three Gorges Project: the largest water conservancy project in the world. Public Admin. Dev. 22 (5), 369–375.

Wang, M.F., Shi, Z., 2004. The reservoir migrants' return migration in the institutional perspective. J. Huazhong Univ. Sci. Technol. Ed. Soc. Sci. 3, 34–38.

Wang, R., 2009. Retrospect of Three Gorges Project demonstration. J. China Three Gorges Univ. (Nat. Sci.) 31 (6), 1–5 (in Chinese).

Wood, S.A., Guerry, A.D., Silver F.M., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. Sci. Rep. 3, 2976.

Wu, J., Huang, J., Han, X., Xie, Z., Gao, X., 2003. Three-Gorges dam – experiment in habitat fragmentation? Science 300, 1239–1240.

Yang, S.L., Milliman, J.D., Li, P., Xu, K., 2011. 50,000 dams later: erosion of the Yangtze River and its delta. Glob. Planet. Change 75 (1), 14–20.

Yang, X., Lu, X.X., 2013. Ten years of the Three Gorges dam: a call for policy overhaul. Environ. Res. Lett. 8 (4), 041006.

Yau, C.K., Porter, A., Newman, N., Suominen, A., 2014. Clustering scientific documents with topic modeling. Scientometrics 100 (3), 767–786.

Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q., 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing – vol. 17, Sapporo, Japan, July 11–12, 2003, pp. 184–187.

Zheng, B., McLean, D.C., Lu, X., 2006. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. BMC Bioinform. 7 (1), 58.

Zhao, X.G., Liu, L., Liu, X.M., Wang, J.Y., Liu, P.K., 2012. A critical-analysis on the development of China hydropower. Renew. Energy 44, 1–6.