Renewable and Sustainable Energy Reviews 57 (2016) 226–237

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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

A topic modeling based bibliometric exploration of hydropower research

Hanchen Jiang, Maoshan Qiang*, Peng Lin

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

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form
10 September 2015
Accepted 17 December 2015
Available online 4 January 2016

Keywords:
Hydropower
Renewable energy
Bibliometrics
Topic modeling
Research trends

ABSTRACT

Scientific research articles can provide rich insights into practitioners' viewpoints around contentious policy making. Although much attention has been paid to hydropower development in the literature, few of them gathered systematic data and performed a large-scale review of scientific articles. In this study, we employed a topic modeling based bibliometric analysis to quantitatively evaluate global scientific literature of hydropower, with a time frame from 1994 to 2013. We analyzed 1726 scholarly articles highly related to hydropower, to discover the research development, current trends and intellectual structure of hydropower literature. Common bibliometric indicators show that hydropower research publications sustain a rapid growth rate, English is the dominant language, and the hotspots of hydropower research can be concluded as "fish", "species", "climate", "emission", "lake", "sediment", "Turkey", etc. We established a 29-topic model to describe the intellectual structure of the 1726 articles, and employed cluster analysis and trend analysis to process the derived topics. We find that post construction issues of hydropower are more attractive for scholars than construction technology itself, and an interdisciplinary trend of hydropower research is emerging. The methodology reported in this study is expected to gain traction as a methodological strategy for energy research reviews and subsequently to promote energy policy making.

Contents

* Corresponding author. Tel.: +86 15210587279; fax: +86 10 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.rser.2015.12.194
1364-0321/

1. Introduction

Worldwide urbanization is stimulating a rapid economic growth together with an increasing demand of energy [1], as energy is the prime agent in the generation of wealth and a significant factor in economic development [2]. However, current trends in energy supply and use are unsustainable economically, environmentally and socially. As traditional fossil fuels still account for a major part of the total energy consumption, greenhouse gas (GHG) emissions caused by usage of these fossil fuels will result in serious consequences including global warming, acid rain, air pollution and many extreme weather disasters [3]. In order to address the global challenges of increasing energy demand, energy security, climate change and sustainable development, there is a pressing need to accelerate the development of renewable energy technologies. Currently, investments in renewable energy sources, including biomass, hydro, geothermal, marine, solar and wind, significantly contribute to the realization of environmental and governmental objectives [4], because renewable energy sources produce zero or almost zero emissions of both air pollutant and GHG [5], and have higher energy security in the face of uncertain markets for traditional fossil fuels [4]. According to the World Energy Outlook 2013 [6], there is a rapid increase in the use of renewable energy, in particular in the power sector. The share of renewable energy in total electricity generation has risen to 20% in 2011, from 15% in 2006, and will rises to 31% in 2035 under the New Policies Scenario [6].

Of the various renewable energy sources, hydropower is the largest one, generating about 3500 TW h (terawatt-hour) electricity in 2011, accounting for 16.3% of world's electricity, more than nuclear power (12.8%), far more than wind, solar, geothermal and other renewable sources combined (3.6%), but still much less than fossil fuel plants (67.2%) [7]. Under the New Policies Scenario, in order to achieve the predicted share of renewable energy in total energy consumption in 2035, major efforts need to be devoted to continuously developing hydropower in the coming decades. Hydropower is fully mature in technology, because of the early development of hydropower projects in many parts of the world. Compared to other renewable energy sources and traditional fossil fuels, hydropower has several advantages, including a high level of reliability, high efficiency, low operating and maintenance costs, flexibility and large storage capacity [2,8,9]. In addition, some large hydropower projects, such as the Three Gorges Project (TGP) in China and the Itaipu Project in Brazil, are usually integrated with other multiple functions including flood control, navigation, and water supply [9,10]. The huge undeveloped potential is another incentive of continuous practice together with research in hydropower sector. The worldwide technical potential for hydropower is usually estimated at around 14,576 TW h per year [11], or about 35% of a theoretical potential derived from the total annual runoff of precipitation [12]. Although there is a long history of human developing hydropower, the percentage of installed technical potential is still low (25% of the total technical potential). From a regional perspective, the percentage of undeveloped technical potential is highest in Africa (92.0%), followed by Asia (80%), Oceania (80%) and Latin America (74%). Even in the most industrialized parts of the world, the undeveloped potential is still significant, at 61% in North America and 47% in Europe [7]. Considering the comprehensive benefits and huge developing potential of hydropower, many countries, such as China [13], Turkey [11], India [14] and Brazil [15], give priority to hydropower development.

However, many serious problems, including environmental issues, socio-economic issues, public acceptance and financing, constraint the development of hydropower projects. Taking the world's largest hydropower project, the TGP, as an example, these four problems are significant during both the construction and operation periods of the project. Environmental impacts of the TGP include an increase in geological risks such as earthquake and landslides [16], water body pollution [17], disruption of the riparian ecosystem [18] and unstable reshaping of the whole Yangtze River system [19]. The socio-economic impacts mainly resulted from the 1.2 million involuntary immigrants [20]. Their survival and developmental issues are of significance to the project success and the social stability. The public acceptance problem of the TGP arose in 1989, when one-tenth of the delegates to the National People's Congress (NPC) signed a petition demanding that the development plan of the TGP should be prudently revised to eliminate negative impacts as much as possible [21]. And in 1992, at the formal vote to approve the project, 1767 delegates voted in favor, 177 delegates opposed the project and 644 abstained [22]. From the beginning of the construction in 1994, the debate of the TGP in China has never stopped [23]. In addition, currently, with the development of social media websites, more and more people discuss public issues through the internet [24], making the public relation problem of a large hydropower project more complex. The financing of the TGP is relatively smooth. The total investment of the TGP is about 200 billion China Yuan, among which 40% is from the TGP Construction Fund, and 20% is from the generation revenue of Gezhouba hydropower station and the TGP itself during construction. The remaining 40% of capital is from the financial market. However, the point is that the TGP Construction Fund was raise from all citizens in China by adding a levy of 0.007–0.015 China Yuan per kilowatt-hour on residential electricity tariffs from 1992 to 2010 [25], leaving a defect which may trigger public dissatisfaction with the project. As a matter of fact, most of these problems with the TGP are commonly existing in hydropower projects all over the world. Unique problems are also found in different hydropower projects, due to the differences in natural and social conditions.

The importance and complexity of the hydropower development issues has attracted scholars across different disciplines, including energy science, hydrology, civil engineering, environmental science, ecology, engineering economics, social science, etc. Although much attention has been paid to hydropower development in these years, few studies attempted to perform a large-scale review of academic articles related to hydropower. An important function of scientific research is that it plays a critical role in identifying many contemporary public policy problems. This function satisfies the increasing requirement of science in the so-called rational decision-making [26]. Science and policy are intimately intertwined. Attempts of making more rational policy necessitate constructing intellectual frameworks of relevant scientific research. Hence, in order to help deploy a rational hydropower development strategy, we performed a topic modeling based bibliometric exploration of peer-reviewed literature reflecting the global status and trend of hydropower research with 1726 target articles published from 1994 to 2013.

2. Methods

Bibliometrics is a set of methods to quantitatively analyze academic literature and describe patterns with a given scientific as well as public concerned topic or field [1,27,28]. With regard to energy issues, particularly, renewable and sustainable energy issues, bibliometric method has been widely used in many review work in recent years. Santos and Sequeira [29] used bibliometric method to examine sodium borohydride ($NaBH_4$) publications to reveal the current perspectives and future outlook of $NaBH_4$ as an efficient energy carrier. Yaoyang and Boeing [30] performed a bibliometric evaluation of research output to map research activities and tendencies of the global biofuel field. Montoya et al. [31] focused on national research status quo and trends, and analyzed the energy research output in Spain based on a bibliometric method. Chen and Ho [32] reported a bibliometric analysis of highly cited articles in biomass research. These studies demonstrated the effectiveness of bibliometric methods in energy review work. Traditional bibliometric studies usually take meta data, such as publication year, author(s), language, etc., as indicators to perform statistic studies. Regarding content analysis, keywords, including author keywords, title keywords or keywords plus (keywords generated by the databases), of articles are also employed as research objects [33–36]. However, in this study, we added "topic" as another fundamental object to describe the intellectual structure of our concerned issues through topic modeling approach. We employed the abstracts of the collected articles as the primary data for topic modeling, giving more comprehensive content analysis results than keyword analysis.

Topic modeling uses statistical algorithms to extract semantic information from a collection of texts and has become an emerging quantitative method to assessing substantial textual data. Hofmann [37] proposed the first topic model, probabilistic Latent Semantic Indexing (pLSI), which models each word in a document as a sample form a mixture model, where the mixture components are multinomial random variables that can be viewed as representation topics. Blei et al. [38] developed an improved three-layer Bayesian model, Latent Dirichlet Allocation (LDA), taking Dirichlet distribution as the prior distribution and reducing the parameter number to only one. Based on the original LDA model, various extensions, such as Correlated Topic Models (CTM) and Hierarchical Dirichlet Process (HDP), have been proposed in recent years to reduce the computing time and the required memory [39,40]. LDA and its extensions have been widely employed in scientometric research to discover semantic structures and latent topics in a discipline or measure the relations of multiple disciplines [41–43].

Recently, topic modeling and other text mining approaches are becoming more approachable as the availability of accessible software enable researchers to take advantage of these methods. Commercial software packages include SAS Text Miner and SPSS Clementine. Open-access options include some R, Python and Java packages. In this study, the modeling process is based on the R package, Topicmodels, offered by Hornik and Grün [44]. Topicmodels package requires a text mining front-end addition, such as the R package, tm, offered by Meyer et al. [45]. And we also make use of MS excel to assist in processing statistical data and plotting figures.

Data used in this study are derived from the Science Citation Index Expanded (SCI-Expanded) and Social Sciences Citation Index (SSCI) databases, provided by the Institute for Scientific Information (ISI), a part of the Thompson Reuters Corporation. The 2013 edition of the Journal Citation Reports (JCR) lists 8539 peer-reviewed journals in SCI-Expanded and 3080 peer-reviewed journals in SSCI. SCI-Expanded and SSCI are strictly selected academic databases. Every journal indexed by SCI-Expanded or SSCI has met the high standards of an objective evaluation process that eliminates clutter and excess and delivers data that is accurate, meaningful and timely. Hence, these two databases have long been recognized as the most authoritative scientific and technical literature indexing tool providing data on the most important areas in science and technology research [46]. In addition, SCI-Expanded and SSCI databases provide convenient keyword search interface, and support bulk downloading of articles' meta data, including title, keywords, abstract, country and publication date, which are necessary for the topic modeling based bibliometric exploration. Using SCI-Expanded and SSCI databases to retrieve publications is a commonly used data collection strategy in bibliometrc or scientometric research [47–49]. Therefore, we follow this strategy in this study.

The search time frame is set from 1994 to 2013. There are two reasons for setting the time frame. First, SCI-Expanded and SSCI databases do not include abstracts before 1991, hence we can only get abstracts of the articles published after 1991. Second, extending the time frame too much may result in a non-uniform sample, where the way researchers are using languages has undergone paradigmatic changes [50]. Therefore, a proper time frame of 20 years, within which article authors have functioned as a community, is selected in this study. With the above settings, we can find articles related to "hydropower" from 1994 to 2013 based on the online SCI-Expanded and SSCI databases, and compile a bibliography as well as a topic analysis of global academic concerns about hydropower. Technical details of investigation methods used in this study are given in the Appendix A.

3. Results

To ensure the comprehensiveness of the search results, we selected "hydropower", "hydro power", "hydro energy", "hydro-energy", "hydroelectric", "hydro electric", "hydroelectricity" and "hydro electricity" as keywords to search articles published between 1994 and 2013 in the online SCI-Expanded and SSCI databases. We obtained 5093 articles which contained any of these keywords in their title, abstract or author-keywords. Article information including author(s), title, source (journal title), language, author-keywords, abstract, publication year, etc. were also exported from the databases. However, after a preliminary inspection of these articles, we found that some articles, which contained a search keyword only in the abstract, had limited relations with hydropower. In order to get articles which were more relevant to hydropower, we refined the obtained data and picked out 1773 articles which contained at least one search keyword in the title or author-keywords. Notice that, some articles did not have an abstract record in the databases, therefore, we searched these articles in other databases or Internet to complete the data collection. Unfortunately, there were still some articles, which were technical notes, news, or viewpoints, without an available abstract. Finally, we obtained 1726 articles with proper abstracts and these abstracts were our primary data for topic modeling analysis. Other information, such as publication year and language were also taken into consideration in the analysis process.

3.1. Overall analysis

3.1.1. Publication trends

In 1994, there were 28 articles indexed by SCI-Expanded or SSCI and highly related to hydropower, and by 2013 the annual total increased to 238. Therefore, when we use articles in SCI-Expanded and SSCI journals as a measure, the hydropower research field has grown over eight-fold since 1994 in terms of the

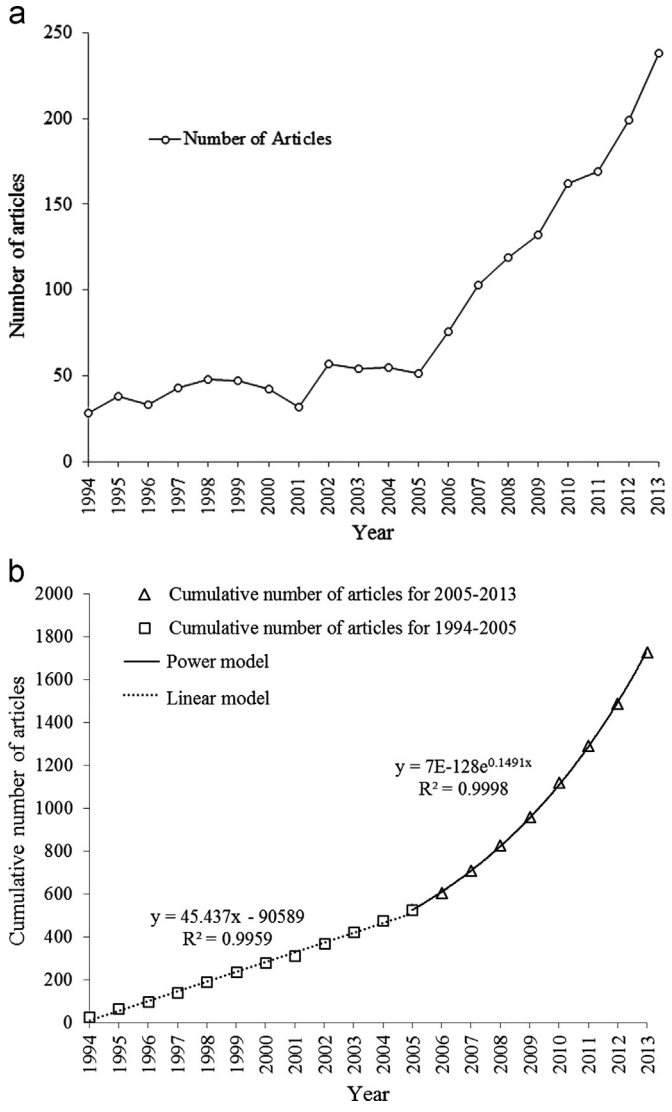

Fig. 1. The trend of (a) number of articles and (b) cumulative number of articles from 1994 to 2013.

number of articles published annually. As shown in Fig. 1(a), we can find an evident increase occurs from 1994 to 2013 due to the increasing concerns on hydropower. The difference in trends of two sub periods, 1994–2005 and 2005–2013, is also significant. As a consequence, we used linear and power models to describe the relationship between the annual cumulative number of articles and the publication year for the two sub periods respectively. The trend of cumulative number of articles is shown in Fig. 1(b). The linear curve fitting result is $y=45.437x-90{,}589$ and the power curve fitting result is $y=7\cdot 10^{-128}\cdot e^{0.1491x}$, where y stands for the cumulative number of articles and x stands for the publication year. Both curves fit the observed data points well with high correlation coefficients ($R^2=0.9959$ for 1994-2005, and $R^2=0.9998$ for 2005–2013). The power model indicates that a rapid growth rate of hydropower research articles since 2005.

3.1.2. Languages

There were 14 languages used for the 1726 articles. The languages in which the articles were published were dominated by English (1521, 88.1%), followed distantly by German (81, 4.7%), French (57, 3.3%), Portuguese (29, 1.7%), Spanish (21, 1.2%), Polish (6, 0.3%), Croatian (4, 0.2%), Estonian (1, 0.06%), Turkish (1, 0.06%), Chinese (1, 0.06%), Russian (1, 0.06%), Slovene (1, 0.06%), Japanese (1, 0.06%) and Czech (1, 0.06%). Results are consistent with previous studies [51–53], which reported that English is the dominant language in many research fields. It could be expected that a high percentage of English would be used, as most journals listed in SCI-Expanded and SSCI are published in English. Note that, although there were 205 articles not written in English, these articles still provided an English abstract. Hence the topic modeling process was not influenced by the usage of multiple languages.

3.1.3. Globally salient terms

Statistics of terms with high frequency of occurrence in the collection of abstracts can provide a roughly panorama of a research field. Some preprocessing steps were performed before the statistic work and subsequent topic modeling. Common stop words, referring to frequent but trivial words, such as “the”, “of”, “and”, etc., were excluded, because these words carried little information. Numbers, punctuations and terms with no more than three characters were also removed. Terms sharing a common stem were consolidated [50,54]. For instance, “study”, “studies” and “studying” were all stemmed as “studi-”. After preprocessing, the top 20 terms of high occurrence frequency are listed in Table 1, using 10 year intervals to guarantee a reasonable time span. Excluding the search keywords (“hydropower”, “hydroelectric” and “hydroelectricity”), “water” (1660), “use” (1643), “model” (1354), “river” (1212), “energi-” (1198), “system” (1095), “dam” (1008), “reservoir” (1066), “develop-” (969) and “plant” (951) remain as the top 10 frequently used terms in the abstracts of hydropower research articles we collected.

However, frequent terms can just provide limited information, as most terms in Table 1 are either trivial in hydropower sector, such as “water”, “energi-”, “river”, “dam”, etc., or trivial in general scientific articles, such as “model”, “use”, “studi-”, “paper”, etc. Therefore, we performed a transformation on the corpus using Term Frequency-Inverse Document Frequencies (TF-IDFs transformation), which penalizes frequent terms that occur in many documents [55,56]. We calculated the TF-IDF values of all terms and sorted these terms by their TF-IDF values. Then, we manually examined these terms from the one with the lowest TF-IDF value and defined a threshold as 0.05. We removed terms which had a TF-IDF value no more than the threshold, and recounted the occurrence number of each remaining term. Table 2 lists the top

Table 1
Top 20 terms of high occurrence frequency from 1994 to 2013.

Stemmed terms	Occurrence number		
	Total	1994–2003	2004–2013
Hydropow-	1892	259	1633
Water	1660	410	1250
Use	1643	379	1264
Model	1354	241	1113
Power	1334	305	1029
River	1212	346	866
Energi-	1198	205	993
System	1095	273	822
Hydroelectr-	1039	317	722
Dam	1008	268	740
Reservoir	1006	227	779
Develop	969	210	759
Plant	951	279	672
Result	944	340	604
Generat-	862	226	636
Oper-	862	360	502
Flow	809	221	588
Studi-	809	231	578
Fish	778	222	556
Paper	757	218	539

Table 2
Top 20 terms (after filtering by TF-IDF) of high occurrence frequency from 1994 to 2013.

Stemmed terms	Occurrence number		
	Total	1994–2003	2004–2013
Fish	778	222	556
Optim-	509	107	402
Control	393	98	295
Turbin-	358	88	270
Cost	352	73	279
Basin	330	54	276
Hydro	325	74	251
Speci-	313	106	207
Renew	311	55	256
Climat-	309	36	273
Flood	306	106	200
Emiss-	299	75	224
Unit	250	54	196
Releas-	231	71	160
Program	224	65	159
Lake	221	96	125
Hydraul-	220	51	169
Turkey	215	38	177
Dynam-	212	56	156
Sediment	209	50	159

20 frequent terms with a TF-IDF value of at least 0.05. Apparently, terms listed in Table 2 are more specific terminology of hydropower research issues. There are several engineering or technology related terms, like “turbin-” (358) and “unit” (250), with a high occurrence number, suggesting the significance of power generation devices research. And the major part of terms listed in Table 2 are related to environmental, ecologic or sustainable issues, such as “fish” (778), “speci-” (313), “climat-” (309), “emiss-” (299), “lake” (221) and “sediment” (209). “Turkey” (215) is the only country name appearing in Table 2, indicating that Turkey has been focusing on developing hydropower during these decades.

3.2. Topic modeling analysis

We employed LDA model to reveal the latent intellectual topics in the literature corpus. To fit the model, we should determine the only parameter, which is the number of topics, of the LDA model. Hence, based on previous studies [57], we computed the posterior likelihoods of a set of models with different numbers of topics to find a maximum. Fig. 2 presents the log-likelihoods of models with different numbers of topics. *Y*-axis of Fig. 2 represents the posterior probabilities of the data occurred in these models with different numbers of topics. The result suggests that the data are best accounted for by a model incorporating 29 topics. We used manual checks to make sure about the validity and robustness of the model. Table 3 displays the top 20 high-frequency words for all topics in the 29-topic model. (The order of topics in Table 3 is not meaningful.)

3.2.1. Topic interpretations

In light of our prior knowledge about hydropower, topics listed in Table 3 are readily recognizable as they are related to major issues in hydropower research field. We provide interpretations of some representative topics as follows.

Topic 3 contains words like “emiss-”, “gas”, “carbon”, “greenhous-”, “flux”, “methan-”, “releas-” and “reservoirs”, and thus pertains to the phenomenon that artificial reservoirs do release GHGs. This is an interesting topic in hydropower research. Intuitively, hydropower is usually known as a clear and renewable energy source producing zero or almost zero GHG emissions. However, scientists find that the decomposition of flooded organic matter in artificial reservoirs produces long-term emissions of carbon dioxide, methane and other GHGs. This topic mainly aims to discuss the influences of reservoirs releasing GHGs and to measure the emissions. Reservoir GHG emissions are significantly observed in tropical area. As a result, the term “tropic-” also ranks highly on the term list of topic 3.

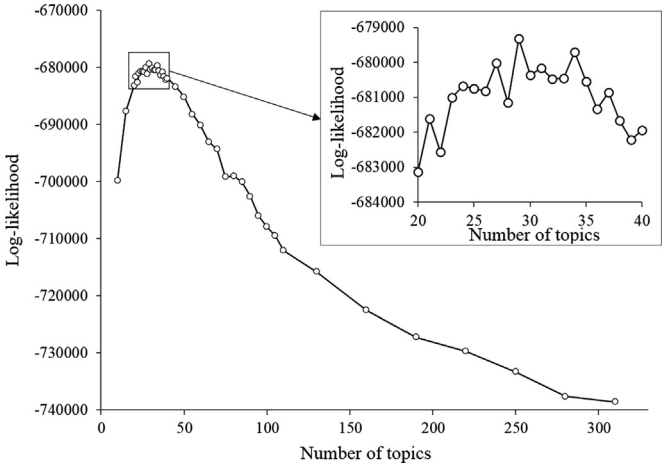

Fig. 2. Posterior log-likelihood of models with different numbers of topic.

Topic 5 contains words like “fish”, “salmon”, “migrat-”, “passag-”, “pass”, “surviv-”, “bypass”, “spawn” and “smolt”, and thus apparently refers to the influences of hydropower facilities on migratory fishes. Life history of these kinds of fishes is the basic knowledge in this research field, therefore, terms like “spawn”, “smolt (young salmon at the stage when it migrates from fresh water to the sea)”, “juvenil-”, “adult” and “moral” are frequently employed in topic 5. Additionally, many of these studies have taken the Chinook salmon living in the Pacific Ocean and the Columbia River as an example. Hence, terms like “Chinook” and “Columbia” are also listed in Table 3.

Topic 21 also addresses fish issues with the highest frequent term of “fish”. However, different from topic 5, topic 21 focuses on the fish ecology and fisheries in hydropower reservoirs. Terms like “weight”, “length”, “net”, “growth”, “catch”, “densiti-” and “oxygen” are all fish ecology or fishery terminology. An important issue in this topic is the heavy metal accumulation in fishes living in reservoirs. Hence, a typical kind of heavy metal, mercury (“mercuri-”), ranks highly on the term list of topic 21.

Topic 13 and topic 20 discuss socio-economic issues related to hydropower. Topic 13 contains words like “rural”, “villag-”, “health”, “survey”, “peopl-”, “communiti-”, “resettl-”, “live” and “household”, and refers to the influences of hydropower projects on the health and well-being of local communities. Topic 20 focuses more on policy and decision making level, hence, employs words like “social”, “projects”, “polici-”, “govern”, “decis-”, “benefit”, “public”, “interest”, “conflict” and “decisionmak-”.

Topic 28 contains words like “hydro”, “renew-”, “wind”, “energy”, “solar” and “thermal”, and thus apparently discusses the role of hydropower in the increasing renewable penetrations. It also includes words like “technolog-”, “storag-”, “pump” and “smallscal-”, indicating that novel hydropower technologies, such as pumped storage power station and small hydropower station, become more and more important in energy development strategy.

Other topics in Table 3 are not discussed in detail due to space constraints, however, they also suggest explicit research themes related to hydropower. These established topics conform to our prior knowledge of hydropower indicated in the introduction section, where hydropower is regarded as a complex energy

Table 3
Top 20 high-frequency terms for each topic in the 29-topic model.

Topic	High-frequent terms
Topic 1	optim-, program, algorithm, schedul-, stochast-, dynam-, nonlinear, hydro, inflow, constraint, unit, rule, formul-, decis-, maxim, comput-, releas-, fuzzi-, polici-, linear
Topic 2	speci-, divers, abund-, species, zone, habitat, veget-, loss, rich, composit-, forest, occurr-, submerg-, highest, tree, riparian, diet, popul-, correl- least
Topic 3	emiss-, gas, carbon, greenhous-, flux, methan-, releas-, emissions, reservoirs, per, ghg-, tropic-, dissolv-, net, cycl-, gase-, much, diffus-, life, atmospher-
Topic 4	hydrogen, Brazil, option, industri-, mitig-, intern, opportun-, resources, capacity, CDM, replac-, barrier, grid, excess, address, clean, short, impacts, cost, attract
Topic 5	fish, salmon, migrat-, passag-, pass, surviv-, bypass, juvenil-, spawn, smolt, Chinook, releas-, mortal, adult, facil-, Columbia, turbin-, movement, success, fall
Topic 6	collect, captur-, vector, Brazil, trap, malaria, mosquito, presenc-, preval-, speci-, fauna, surround, yellow, drift, earli-, preserv-, stations, transmiss-, anophel-, frequenc-
Topic 7	temperatur-, scale, air, paramet-, index, refer, heat, interact, field, across, thermal, aspect, transfer, larger, coeffici-, exploit, experi-, modifi-, temperature, distanc-
Topic 8	eel, treatment, trend, reach, behavior-, group, silver, recoveri-, attribut-, start, modifi-, origin, sea, European, past, section, spillway, sensit-, anguilla-, hep-
Topic 9	climat-, basin, forecast, scenario, predict, runoff, uncertainti-, degre-, inflow, precipit-, volum-, change, streamflow, rainfal-, storag-, models, histor-, glacier, respect, catchment
Topic 10	discharg-, scheme, peak, intak-, built, ice, canal, veloc-, channel, flows, attempt, cours-, much, weir, almost, profil-, schemes, used, meter, relationship
Topic 11	control, flood, risk, mode, strategi-, probabl-, activ-, defin-, event, reliabl-, return, class, optimum, optimis-, operation, rout, fault, periods, hgu-, residu-
Topic 12	monitor, network, point, detail, map, analysis, statist, spatial, neural, phase, wide, identif-, transform, novel, analyt-, linear, deriv-, error, hazard, technique
Topic 13	land, rural, villag-, health, survey, landscap-, peopl-, communiti-, access, resettl-, live, place, electrif-, remot-, household, poor, agricultur-, lack, suffici-, incom-
Topic 14	structur-, build, dynam-, procedur-, unit, mainten-, engin-, long, technic, offer, operation, softwar-, presented, transmiss-, link, concept, exampl-, view, visual, learn
Topic 15	turbin-, head, shp-, speed, output, turbines, runner, equip, american-, turbine, blade, paramet-, engineers, convert, civil, detail, wheel, compon-, convent, devic-
Topic 16	regul-, habitat, regim-, stream, popul-, trout, communiti-, reach, fluctuat-, channel, growth, biomass, section, biolog-, group, sites, rapid, domin-, dri-, age
Topic 17	qualiti-, standard, author, status, around, field, challeng-, now, report, Norway, green, old, south, long, intern, final, north, still, america-, technic
Topic 18	irrig-, basin, loss, winter, central, run, strategi-, anoth-, histor-, summer, illustr-, put, multipurpos-, releas-, alloc-, management, mountain, core, scienc-, uses
Topic 19	load, unit, machin-, rotor, compon-, connect, frequenc-, forc-, element, mechan-, voltag-, magnet, vibrat-, induct, ump, equival-, radial, bear, finit-, dynam-
Topic 20	social, projects, polici-, articl-, govern, decis-, benefit, public, interest, issu-, protect, conflict, feder-, institut-, way, adapt, face, framework, polit-, decisionmak-
Topic 21	fish, weight, fisheri-, length, net, mercuri-, correl-, growth, catch, stock, densiti-, million, highest, oxygen, relationship, per, seem, elev-, end, tributari-
Topic 22	specif-, safeti-, materi-, discussed, stage, steel, life, taken, give, bear, yield, extern, reason, essenti-, regard, point, relev-, experi-, weak, criteria
Topic 23	hydraul-, pressur-, numer-, comput-, experiment, tank, transient, gate, veloc-, equat-, surg-, mathemat-, close, predict, penstock, rise, pipe, draft, tube, fluid
Topic 24	sediment, lake, concentr-, soil, flood, organ, depth, reservoirs, transport, MeHg, surfac-, particl-, matter, contamin-, eros-, column, nutrient, metal, phyto-plankton, dri-
Topic 25	ecolog-, ecosystem, China, cascad-, servic-, Mekong, framework, conserv-, benefit, upper, restor-, watersh-, gorg-, user, forest, rivers, adequ-, spatial, aquat-, largescal-
Topic 26	cost, market, price, methodolog-, invest, financi-, compani-, plants, facil-, profit, privat-, sector, runofriv-, trade, gain, coordin-, costs, enabl-, overview, efficiency
Topic 27	Turkey, countri-, world, renew-, hydropower, plants, consumpt-, grow, role, potential, countries, meet, fuel, sector, fossil, play, nuclear, primari-, biomass, pollut-
Topic 28	hydro, renew-, wind, technolog-, storag-, pump, energy, solar, scienc-, convent, thermal, hybrid, sources, best, independ-, smallscal-, storage, considered, island, alreadi-
Topic 29	rock, slope, tunnel, stress, mass, stabil-, geolog-, engin-, excav-, deform, failur-, zone, crack, fractur-, mechan-, joint, numer-, surfac-, bank, strength

source involving many scientific, technical, environmental, ecologic and socio-economic issues. As a summary, we manually sorted these topics into four categories, including construction technology, operation technology, environmental issues and socio-economic issues (including energy strategy research), as shown in Table 4. In fact, operation technology, environmental issues and socio-economic issues can all be regarded as post construction issues of hydropower. Therefore, we can infer that post construction issues are dominant in hydropower research.

3.2.2. Topic proportions

The 29-topic model assigned topic proportions to each abstract we collected. For example, Fig. 3 presents the title and abstract of a randomly selected article and the diagnostic proportions of all the 29 topics for the article (abstract). As shown in the right part of Fig. 3, topic 16 and topic 21 are the two strongest diagnostic topics, which have significantly higher proportions than other topics do. Referring to the title and abstract shown in the left part of Fig. 3, it can be known that the article indeed discusses the influences of hydropower projects on fish habitats (topic 16) and fish ecology (topic 21). Hence it demonstrates that the assignment of topic proportions agrees with the topic contents reflected by the high-frequency terms listed in Table 3

Integrating topic proportions for all the abstracts in the corpus, we obtained a topic distribution for the whole corpus, as shown in Fig. 4. Based on Fig. 4, the five highest-frequency research topics of hydropower are topic 1 (4.9%), topic 5 (4.2%), topic 27 (4.0%), topic 20 (3.9%) and topic 29 (3.8%), while the five lowest-frequency research topics are topic 13 (3.1%), topic 22 (3.1%), topic 18 (3.0%), topic 6 (3.0%) and topic 8 (3.0%). We can infer the detailed contents of these topics from Table 3.

Table 4
Categorization of the 29 hydropower research topics.

Category	Topic
Construction technology	Topic 14, Topic 15, Topic 17, Topic 19, Topic 22, Topic 23, Topic 29
Operation technology	Topic 1, Topic 10, Topic 11, Topic 12
Environmental issues	Topic 2, Topic 3, Topic 5, Topic 6, Topic 7, Topic 8, Topic 9, Topic 16, Topic 18, Topic 21, Topic 24, Topic 25
Socio-Economic issues	Topic 4, Topic 13, Topic 20, Topic 26, Topic 27, Topic 28

The proportions of the four categories in Table 4 were also calculated. Environmental issues account for the largest proportion (40.9%), followed by construction technology (23.5%), socio-economic issues (21.1%) and operation technology (14.4%). The results also support the notion that post construction issues attract more attention than construction issues in hydropower research. Two reasons are suggested for the results. First, hydropower construction is a proven technology which was widely discussed in the earlier (before 1990 s) literature, hence only a few articles we collected in this study address the construction technology of hydropower. Second, regarding the global trend

Differences in habitat utilization among native, native stocked, and non-native stocked brown trout (Salmo trutta) in a hydroelectric reservoir

Native and native-stocked brown trout (Salmo trutta) in Lake Tesse, a regulated hydroelectric reservoir (southern Norway), were spatially segregated according to size: small individuals occurred mainly in the epibenthic habitat and larger individuals mainly in the pelagic habitat. In contrast, all size groups of non-native stocked brown trout were mostly restricted to the epibenthic habitat. Age-specific lengths were generally larger for non-native than for native stocked trout, which were larger than native fish, However, growth rate between age 3 and 4 was significantly lower for non-native stocked fish than for native and native stocked fish. Differences in body length were mainly due to strain but also to some extent to habitat.....

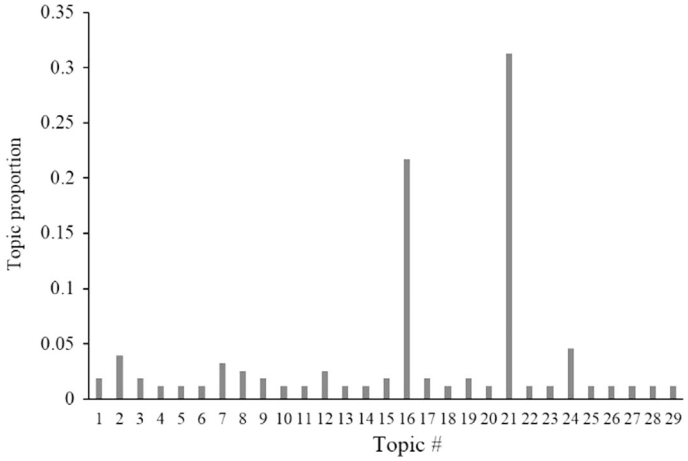

Fig. 3. Title and abstract of a randomly selected article and diagnostic topic proportions for the article.

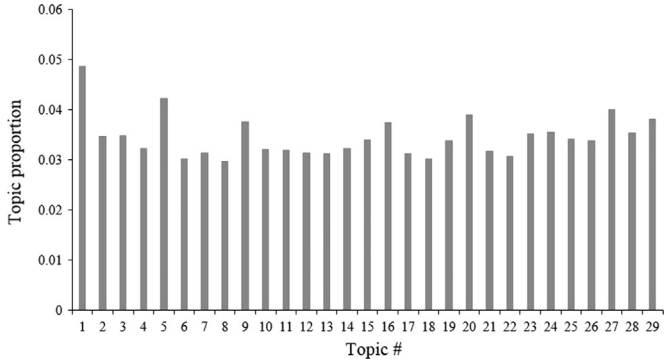

Fig. 4. Topic proportions for the whole corpus.

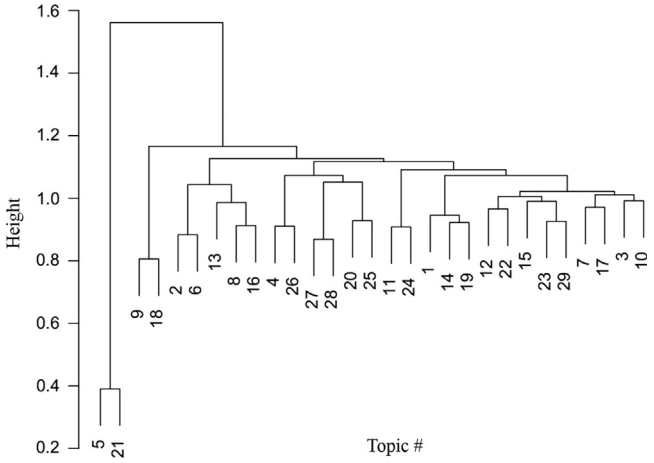

Fig. 5. Dendrogram of the term-level similarity clustering result.

of sustainable development in recent years, the environmental and socio-economic impacts as well as the operation optimization of numerous existing hydropower projects naturally have become the major focuses of hydropower research.

3.3. Topic cluster analysis

We used the hierarchical cluster analysis with Ward's method [26,34] to perform the cluster analysis of the 29 topics listed in Table 3. There are two ways of measuring topic similarity: topics may contain some of the same terms, or topics may appear in some of the same documents. We called topic similarities measured by the two ways as term-level similarity and document-level similarity respectively.

3.3.1. Term-level similarity clustering

The dendrogram of the clustering result based on term-level similarity is shown in Fig. 5. The integer numbers in Fig. 5 are topic numbers corresponding to those in Table 3. And the vertical axis, height, represents the disparity of topics. Lower location of connecting line means topics are more similar. As can be seen from Fig. 5, topic 5 and topic 21 have high term-level similarity and are far distant from other topics. From the topic interpretations above, topic 5 and topic 21 both address fish issues, using many analogous terms. However, fish issues are little correlated with other hydropower issues. Hence, the cluster of topics 5 and topic 21 is separate from the main stem of the dendrogram. Other topics have less term-level similarity than topic 5 and topic 21 do, and are mapped in the middle of the dendrogram. For instance, topic 9 and topic 18 both discuss basin hydrology issues, but they only share one high-frequency term ("basin") in Table 3. Topic 9 focuses on the climate or micro-climate change of a basin where a hydropower project is located. Meanwhile, topic 18 focuses on the water resource management in a basin. Cluster of topic 27 and topic 28 is another example. The two topics both discuss the role of hydropower in various renewable energy sources. Topic 27 is from a national perspective, focusing on Turkey's hydropower development strategy. And topic 28 is from a technology innovation perspective, focusing on novel hydropower technology and the linkage between hydropower and other energy sources. In summary, the dendrogram shown in Fig. 5 clearly presents the term usage similarity structure of hydropower research topics.

3.3.2. Document-level similarity clustering

The dendrogram of the clustering result based on document-level similarity is shown in Fig. 6. Different from term-level similarity clustering, document-level similarity clustering aims to describe the interaction structure of hydropower research topics. As shown in Fig. 6, topic 5 and topic 8 have a high document-level similarity, which means articles with a high topic proportion of topic 5 often have a high topic proportion of topic 8 simultaneously. This measure of topic similarity has the same meaning of interdisciplinary analysis [42]. If two topics frequently appear in the same articles, there is a big potential to foster a novel interdisciplinary research field. Hence, Fig. 6 presents the emerging trends of hydropower research in the past two decades.

Combining Figs. 6 and 5, it can be found that some clusters are consistent in terms of both term-level and document-level similarities, such as clusters of topic 9 and 18, topic 20 and 25, and

topic 27 and 28. There are also topic pairs which have high term-level similarity but low document-level similarity, such as topic 3 and topic 10, or which have high document-level similarity but low term-level similarity, such as topic 1 and topic 26. The consistency and difference between term-level similarity and document-level similarity also reveal the intellectual structure of hydropower research in the past two decades.

3.4. Topic trend analysis

It is believed that research shows strong trends, with topics rising and falling regularly in popularity. Identifying emerging research topics within a rapidly growing and changing field may provide valuable insights into how and why concerns evolve. Likewise, identifying fading research topics can also help understand the intellectual structure of a research field. Therefore, we normalized the proportions of the 29 topics year by year, and obtained the annual trends of the 29 hydropower research topics as shown in Fig. 7. We used Mann–Kendall test [58], a nonparametric trend test, to examine whether increasing or decreasing trends were existing in the 29 topics. Test results show that seven topics, including topic 9, topic 13, topic 15, topic 19, topic 23, topic 25 and topic 26, present a statistically significant increasing trend, and three topics, including, topic 4, topic 16 and topic 24, present a statistically significant decreasing trend, both at the two-sided $P=0.05$ level.

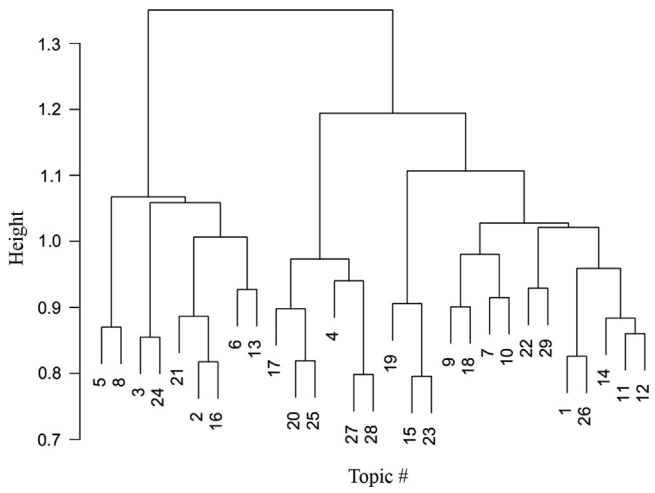

Fig. 6. Dendrogram of the docment-level similarity clustering result.

We provide brief explanations of these emerging or fading topics. (1) Emerging topics: topic 19 discusses the operation characteristics, especially the load characteristic, of hydropower generator units; topic 23 focuses on the hydraulic problems of hydropower projects; topic 25 addresses the impacts of hydropower development on ecosystem, and pays special attention to China and Mekong River; topic 26 discusses the market, pricing and financing issues of hydropower; topic 9, topic 13 and topic 15 have been explained in the previous parts of this article. (2) Fading topics: topic 4 discusses the combination of hydropower and other renewable energy sources, such as hydrogen, to achieve a Clean Development Mechanism (CDM); topic 24 discusses the pollution in reservoirs, and specially focuses on the methylmercury (MeHg); topic 16 has been explained in the previous parts of this article.

Other topics, which do not reveal a significant increasing or decreasing trend, also vary in the characteristics of their trends. As can be seen from Fig. 7, some topics, such as topic 14 and topic 18, show a stable trend during 1994 to 2013. Meanwhile, some topics, such as topic 1, topic 3 and topic 29, show a trend with relatively sharp fluctuations. And topic 27 shows an increasing trend before

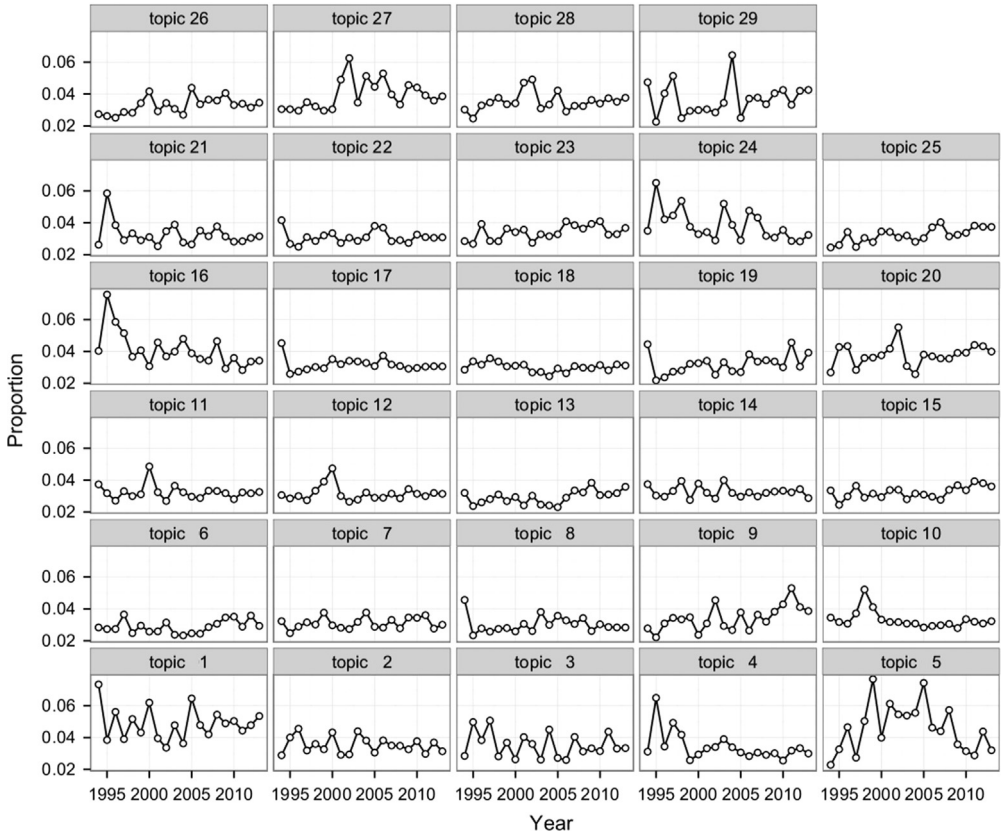

Fig. 7. Hydropower research topic trends during 1994–2013.

2002 and a decreasing trend after 2002. Detailed explanations of these trends and characteristics need a further investigation of the model and a more in-depth understanding of hydropower issues.

4. Discussion

In terms of total capacity and generation, hydropower is the dominant source of renewable energy currently. Guney [59] expected that the world annual renewable energy generation will reach 5.8 trillion kW h by 2020, where hydropower will share 4.4 trillion kWh from total or 76%. The International Energy Agency (IEA) [7] also reported that hydropower is a competitive energy source already today and will remain so for a long time, but its further deployment still faces many challenges including environmental, socio-economic, public acceptance and financial issues. We see this study as strong support for this positioning of hydropower from the academic perspective. Firstly, the rapid growth rate of hydropower research articles is in correspondence with the vigorous development of hydropower in recent decades. Secondly, the topic distribution reflects that the major academic concerns of current hydropower development are ex-post construction issues rather than construction technologies. In view of the fact that scientific publications are often in advance of industrial practice, these ex-post construction issues will be more and more critical in further deployment of hydropower. Thirdly, the interdisciplinary trend of hydropower research also shows the complexity in hydropower development. Interactions of technology, society and governance will become vital constraints of hydropower development in the future.

Previous studies [60,61] have suggested that social scientists and decision makers can benefit from large-scale text data with proper text mining methods. Bibliometric analysis is such a kind of method which can arrange copious scientific literature data and promote decision making. In addition, scientific literature data provide more condensed, informative and objective information than other data sources, such as news, reports and social media comments. However, traditional bibliometric analysis methods only focus on meta data, such as publication year, author(s), institution(s), source, language, keywords, citation index, etc., and can just offer some statistic indicators for a research field. For example, Han et al. [1] performed a similar bibliometric analysis for sustainable hydropower development research using bibliometric indicators only. Although they provided a systematic overview of 434 hydropower sustainability related articles, their study gave more contributions to the bibliometrics or scientometrics rather than to the hydropower industry. Hydropower is quite a comprehensive field involving many professional issues, any of which should be regarded as a separate topic. Therefore, topic modeling, which is able to reveal the latent intellectual structure of a collection of documents, has been combined with traditional bibliometric analysis in this study to provide a more comprehensive and informative overview of the collected hydropower research articles. For practitioners and decision makers, this kind of summary provided by topic modeling just caters to their requirements in seizing the major academic concerns of hydropower. Moreover, these academic concerns do reflect practical issues in the whole life cycle of hydropower development, because the data and material of these SCI-Expanded or SSCI articles are always obtained by scientific and rigorous methods, such as long term observation, field investigation and in-situ testing.

In addition to promoting practice and policy making, this study makes contributions to understand the factors that shape hydropower research priorities. As indicated in the introduction, scientific research plays an important role in identifying many contemporary problems. The knowledge of how research priorities emerge is critically important to the understanding of the role that science plays in society [26]. This study, with its approach to identify substantial topics, topic proportions and trends, and emerging interdisciplinary fields, provides necessary background information for further exploration of hydropower research questions, such as: (1) How do science, technology, society and governance interact in the whole life cycle of a hydropower project? (2) How do political trends (e.g., interest in southern hemisphere countries, tropical forests or endangered species), enterprise funding and research relate to each other in hydropower development over a long period? (3) How do interdisciplinary topics emerge with the accelerated process of the worldwide hydropower development? The analysis method and data we used in this study allow us to explore these questions in future work.

There are some limitations in this study. One is that citation data have not be employed in the analysis, although they are indeed valuable to describe relations between scientific articles. Thus, further investigation are required to take citation data into consideration, with an in-depth understanding of the citing rationale. The other limitation is from the usage of topic modeling, as any new form of research methodology faces complications in execution. The number of topics was selected by a statistical measure of model fit in this study. The measure showed that a 29-topic model was the most likely model fitting the data set. However, mechanical reliance on statistical measures may lead to the selection of a less meaningful topic model [62]. Hence, we used some manual checks to assess the validity and robustness of the results. First, the topics revealed by the model are confirmed by qualitative assessment, based on our prior knowledge in hydropower. For each topic, we checked the semantic coherence of its high-frequency terms and examined the abstracts which had a high proportion of this topic. Second, we ran models with different numbers of topics and compared these models with the 29-topic model. In fact, models with 25-35 topics had little difference in revealing the intellectual structure of the data set. However, in order to provide a repeatable process, using statistical measure to find a mathematically optimal model would be a good choice. Third, we ran multiple iterations of the 29-topic model, starting from different randomly selected seeds. The results, including high-frequency terms and topic proportions, were quite consistent across these iterations. It should be admitted that the validation of a topic model can only be performed by manual inspection which may be subjective.

Notwithstanding its limitations, this topic modeling based bibliometric exploration has successfully mapped an intellectual structure of the research topics of hydropower. In further study, we propose to introduce this method into reviews of other energy research to carry out a comparative study of hydropower and other energy sources. Energy is a complex and interdisciplinary research field, which involves many relevant or distinct topics. Scientists, journalists and even common people are concerned with energy issues and produce many text data, such as research articles, news reports and blogs, to express their opinions. These text data are of value for energy research. Recent 'Big Data' trend has inspired computer scientists to develop approaches to understand latent intelligence in collections of text data. However, while computer scientists have produced powerful tools, including topic modeling, for automated content analyses of big text data, they lack application scenarios and related theoretical directions to extract knowledge from the data. Meanwhile, other scientists, including energy researchers, have a huge demand of analyzing massive text data for tackling real-world problems, but lack the methodological capacity to explore them beyond micro-levels of analysis [63]. This study gives an instruction to researchers who are not familiar with topic modeling, but want to gain insights from recent 'Big Data' trend. We expect the methodology in this

study to gain traction as a methodological strategy for energy research reviews and subsequently promote energy policy making, especially the policies that impact research priorities.

5. Conclusions and managerial implications

Hydropower will orient the development strategy of renewable energy due to its huge development potential, economic and social benefits and proven technology. However, we cannot omit the negative impacts brought by hydropower projects in terms of environment, ecology and socio-economy. Scientific literature related to hydropower is an abundant and reliable data pool, from which we can understand the major academic concerns about hydropower and hence deploy a proper development strategy of hydropower. Based on the 1726 articles collected from the SCI-Expanded and SSCI databases, a topic modeling based bibliometric exploration regarding the global research trend of hydropower is conducted. Results of this exploration present a comprehensive overview and an intellectual structure of hydropower research, especially, hydropower research topics, from 1994 to 2013.

For the collected articles, linear and exponential relations between yearly cumulative number of articles and publication year have been obtained for the period from 1994 to 2005 and 2005 to 2013, revealing that annual article publications sustain a constant growth rate. English is the dominate language, accounting for 88.1% of the whole collection of articles. The 29-topic model has been successfully applied to discover the latent thematic patterns in the corpus. And the derived 29 topics are clustered based on term-level similarity and document-level similarity, to find latent relations and emerging interdisciplinary fields of these topics. Increasing and decreasing topics are also recognized through statistic test. Based on these topic analysis results, we find that ex-post construction issues of hydropower development are more attractive for scholars than energy technology itself, and an interdisciplinary trend of hydropower research is emerging from the interaction of natural science, social science and engineering technology related to hydropower.

This study directly contributes to our understanding of what academic concerns of hydropower are in the past two decades. Besides, the findings of this study have implications for future energy policy. Firstly, the rapid growth rate of publications indicates that there is a huge demand for hydropower related research. And along with the accelerated process of hydropower development, government should provide more funding for this research field. Secondly, the derived topic distribution suggests that for industrialized countries with fully or almost fully developed hydropower potential, ex-post issues, upgrading or redevelopment of existing plants should be emphasized to alleviate negative impacts and deliver additional benefits. Thirdly, as most of the growth in hydroelectricity generation will come from large projects in emerging economies and developing countries, these countries should learn lessons from the experience of hydropower development in industrialized countries. And this study just provides a systematic overview for this purposes.

Acknowledgments

This research work was supported by National Natural Science Foundation of China (General Program nos. 51479100, 51179086, 11272178, and 51379104), and grants (Nos. 2013-KY-5 and 2015-KY-5) from State Key Laboratory of Hydroscience and Engineering, China.

Appendix A. Technical details of investigation methods

TF-IDF transformation

TF-IDF stands for term frequency-inverse document frequency. TF-IDF weight is often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [56].

The TF-IDF weight is composed by two terms. The first is Term frequency (TF), which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalization. Given that there is a corpus containing D documents using V unique terms, the TF value of the ith term in the jth document is calculated as follows:

$$TF_{ij} = C_{ij}/N_j \tag{A.1}$$

where TF_{ij} is the TF value of the ith term in the jth documents, C_{ij} is the number of times the ith term appears in the jth document, and N_j is the total number of terms in the jth document.

The second is Inverse Document Frequency (IDF), which means how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

$$IDF_i = \log_2(D/D_i) \tag{A.2}$$

where IDF_i is the IDF value of the ith term, D is the total number of documents, and D_i is the number of documents with the ith term in it. And the TF value of the ith term in the whole corpus is calculated as follows:

$$TF_i = \sum_{j=1}^{D} TF_{ij} * IDF_i \tag{A.3}$$

LDA model

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 [38]. 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:

$$P(w=v) = \sum_{j=1}^{T} P(w=v \mid z=j)P(z=j) \tag{A.4}$$

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. Suppose 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 sematic representation of words and documents. In order to estimate Φ and Θ for a given collection of documents. Hofmann proposed to maximize $P(w|\phi, \theta)$ directly by using the Expectation-Maximization (EM) algorithm to find maximum likelihood estimates of ϕ and θ [64]. However, this EM algorithm may cause overfitting and be slow to converge, encouraging new models that make assumption about the source of ϕ and θ.

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:

$$p(w_m,z_m,\theta_m,\Phi\mid\alpha,\beta)=\prod_{n=1}^{N_m}p(w_{m,n}|\phi_{z_{m,n}})p(z_{m,n}|\theta_m)p(\theta_m|\alpha)p(\Phi|\beta) \tag{A.5}$$

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:

$$p(w_{m,n}=v|\theta_m,\Phi)=\sum_{j=1}^{T}p(w_{m,n}=v|\phi_j)p(z_{m,n}=j|\theta_m) \tag{A.6}$$

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:

$$p(W|\Theta,\Phi)=\prod_{m=1}^{D}p(w_m|\theta_m,\Phi)=\prod_{m=1}^{D}\prod_{n=1}^{N_m}p(w_{m,n}|\theta_m,\Phi) \tag{A.7}$$

The task of estimating parameters Φ and Θ in LDA can be accomplished using statistical techniques such as variation EM algorithm [38] and Gibbs sampling [57]. 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.

Topic clustering

The clustering method in this study is the hierarchical cluster analysis with Ward's method [65]. Ward's method has been widely used since 1963, and is appropriate for quantitative variables. Mathematical details of Ward's method can be found in Ward's original paper [65]. To this study, the more important issue is the similarity metric of topic. Detailed explanations of term-level similarity and document-level similarity are shown as follows:

Both term-level similarity and document-level similarity are based on cosine similarity. Given two vectors of attributes, A and B, the cosine similarity, $\cos(A, B)$, is represented using a dot product and magnitude as follows:

$$similarity=\cos(A,B)=\frac{A\cdot B}{\|A\|\|B\|}=\frac{\sum A_i\times B_i}{\sqrt{\sum(A_i)^2}\times\sqrt{\sum(B_i)^2}} \tag{A.8}$$

There are two types of vectors related to a topic in topic modeling. First, a topic is a term distribution on a vocabulary, which is also the definition of topic in topic modeling. Hence, given a ordered vocabulary containing V unique terms, the topic k can be represented by a vector $VT=(\phi_{k,1}, \phi_{k,2},\ldots, \phi_{k,V})$. $\phi_{k,i}$ means the probability that term i appears in topic k. And the term-level similarity between two topics, k and l, can be calculated as follows:

$$\cos_{term}(k,l)=\frac{\sum_{i=1}^{V}\phi_{k,i}\times\phi_{l,i}}{\sqrt{\sum_{i=1}^{V}(\phi_{k,i})^2}\times\sqrt{\sum_{i=1}^{V}(\phi_{l,i})^2}} \tag{A.9}$$

Second, a topic is assigned to all documents in the corpus. Hence, a topic can also be regarded as a document distribution on a corpus. Given an ordered document list containing D documents, the assignment of the topic k to all documents can be denoted by a vector $VD=(\theta_{k,1}, \theta_{k,2},\ldots, \theta_{k,D})$. $\theta_{k,i}$ means the proportion of topic k in document i. And the document-level similarity between two topics, k and l, can be calculated as follows:

$$\cos_{document}(k,l)=\frac{\sum_{i=1}^{D}\theta_{k,i}\times\theta_{l,i}}{\sqrt{\sum_{i=1}^{D}(\theta_{k,i})^2}\times\sqrt{\sum_{i=1}^{D}(\theta_{l,i})^2}} \tag{A.10}$$

Mann–Kendall test

The Mann–Kendall trend test [58] is based on the correlation between the ranks of a time series and their time order. For a time series $X=\{x_1, x_2,\ldots, x_n\}$, the test statistic is given by:

$$S=\sum_{i<j}a_{ij} \tag{A.11}$$

where:

$$a_{ij}=\mathrm{sign}(x_j-x_i)=\begin{cases}1 & x_i<x_j\\ 0 & x_i=x_j\\ -1 & x_i>x_j\end{cases} \tag{A.12}$$

Under the assumption that the data are independent and identically distributed, the mean and variance of the S statistic in Eq. (A.11) above are given by

$$E(S)=0 \tag{A.13}$$

$$V(S)=n(n-1)(2n+5)/18 \tag{A.14}$$

where n is the number of observations. The distribution of S tends to normality as the number of observations becomes large. The significance of trends can be tested by comparing the standardized variable u in Eq. (A.15) with the standard normal variate at the desired significance level α, where the subtraction or addition of unity in Eq. (A.15) is a continuity correction.

$$u=\begin{cases}(S-1)/\sqrt{V(S)} & S>0\\ 0 & S=0\\ (S+1)/\sqrt{V(S)} & S<0\end{cases} \tag{A.15}$$

References

[1] Han M, Sui X, Huang Z, Wu X, Xia X, Hayat T, Alsaedi A. Bibliometric indicators for sustainable hydropower development. Ecol Indic 2014;47:231–8.
[2] Yüksel I. Hydropower for sustainable water and energy development. Renew Sustain Energy Rev 2010;14:462–9.
[3] Liang S, Wang C, Zhang T. An improved input–output model for energy analysis: a case study of Suzhou. Ecol Econ 2010;69:1805–13.
[4] Johnstone N, Haščič I, Popp D. Renewable energy policies and technological innovation: evidence based on patent counts. Environ Resour Econ 2010;45:133–55.
[5] Demirbaş A. Global renewable energy resources. Energy Sources 2006;28:779–92.
[6] IEA. World Energy Outlook. Paris: International Energy Agency; 2013.
[7] IEA. World Energy Outlook. Paris: International Energy Agency; 2012.
[8] Kaldellis JK. Critical evaluation of the hydropower applications in Greece. Renew Sustain Energy Rev 2008;12:218–34.
[9] Liu J, Zuo J, Sun Z, Zillante G, Chen X. Sustainability in hydropower development – a case study. Renew Sustain Energy Rev 2013;19:230–7.
[10] Ozturk M, Bezir NC, Ozek N. Hydropower–water and renewable energy in Turkey: sources and policy. Renew Sustain Energy Rev 2009;13:605–15.
[11] Balat M. Hydropower systems and hydropower potential in the European Union countries. Energy Sources Part A 2006;28:965–78.
[12] IJHD (International Journal on Hydropower & Dams). World atlas & industry guide. Wallington, Surrey, UK; 2010.
[13] Huang H, Yan Z. Present situation and future prospect of hydropower in China. Renew Sustain Energy Rev 2009;13:1652–6.
[14] Yüksel I. Development of hydropower: a case study in developing countries. Energy Sources Part B 2007;2:113–21.
[15] Sousa Júnior WC, Reid J. Uncertainties in Amazon hydropower development: risk scenarios and environmental issues around the Belo Monte dam. Water Altern 2010;3:249–68.
[16] Yang X, Lu X. Ten years of the Three Gorges Dam: a call for policy overhaul. Environ Res Lett 2013;8:041006.
[17] Fu B, Wu B, Lü Y, Xu Z, Cao J, Niu D, Yang G, Zhou Y. Three Gorges Project: efforts and challenges for the environment. Prog Phys Geogr 2010;34:741–54.
[18] Wu J, Huang J, Han X, Xie Z, Gao X. Three-Gorges dam–experiment in habitat fragmentation? Science 2003;300:1239–40.
[19] Yang S, Milliman J, Li P, Xu K. 50,000 dams later: erosion of the Yangtze River and its delta. Glob Planet Change 2011;75:14–20.
[20] Tullos D. Assessing the influence of environmental impact assessments on science and policy: an analysis of the Three Gorges Project. J Environ Manag 2009;90:S208–23.
[21] Reng X. The banning of Yangtze! Yangtze!. In: Dai Q, Adams P, Thibodeau J, editors. Yangtze! Yangtze!. London: Earthscan; 1994. p. 13–21.
[22] He S, Si K. The comeback of the Three Gorges dam. In: Dai Q, Adams P, Thibodeau J, editors. Yangtze! Yangtze!. London: Earthscan; 1994. p. 22–45.
[23] Lowry W. Potential focusing projects and policy change. Policy Stud J 2006;34:313–35.
[24] Franch F. (Wisdom of the crowds) 2: 2010 UK election prediction with social media. J Inf Technol Politics 2013;10:57–71.
[25] Lu Y. The exploitation and sustainable development of hydropower in China. Water Resour Hydropower Eng 2004;36:1–4 Chinese.
[26] Neff MW, Corley EA. 35 years and 160,000 articles: a bibliometric exploration of the evolution of ecology. Scientometrics 2009;80:657–82.
[27] Ho YS. Bibliometric analysis of biosorption technology in water treatment research from 1991 to 2004. Int J Environ Pollut 2008;34:1–13.
[28] Pritchard A. Statistical bibliography or bibliometrics. J Doc 1969;25:348.
[29] Santos DMF, Sequeira CAC. Sodium borohydride as a fuel for the future. Renew Sustain Energy Rev 2011;15(8):3980–4001.
[30] Yaoyang X, Boeing WJ. Mapping biofuel field: a bibliometric evaluation of research output. Renew Sustain Energy Rev 2013;28:82–91.
[31] Montoya FG, Montoya MG, Gómez J, Manzano-Agugliaro F, Alameda-Hernández E. The research on energy in Spain: a scientometric approach. Renew Sustain Energy Rev 2014;29:173–83.
[32] Chen H, Ho YS. Highly cited articles in biomass research: a bibliometric analysis. Renew Sustain Energy Rev 2015;49:12–20.
[33] Chiu WT, Ho YS. Bibliometric analysis of tsunami research. Scientometrics 2007;73:3–17.
[34] Zong QJ, Shen HZ, Yuan QJ, Hu XW, Hou ZP, Deng SG. Doctoral dissertations of Library and Information Science in China: a co-word analysis. Scientometrics 2013;94:781–99.
[35] Li T, Ho YS, Li CY. Bibliometric analysis on global Parkinson's disease research trends during 1991–2006. Neurosci Lett 2008;441:248–52.
[36] Zhang G, Xie S, Ho YS. A bibliometric analysis of world volatile organic compounds research trends. Scientometrics 2010;83:477–92.
[37] Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval; 1999 Aug 15–19; Berkeley, USA. New York: ACM; 1999.
[38] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
[39] Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat 2007;1:17–35.
[40] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical dirichlet processes. J Am Stat Assoc 2006;101:1566–81.
[42] Lu K, Wolfram D. Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches. J Am Soc Inf Sci Technol 2012;63:1973–86.
[42] Nichols LG. A topic model approach to measuring interdisciplinarity at the National Science Foundation. Scientometrics 2014;100:741–54.
[43] Yau CK, Porter A, Newman N, Suominen A. Clustering scientific documents with topic modeling. Scientometrics 2014;100:767–86.
[44] Hornik K, Grün B. Topicmodels: an R package for fitting topic models. J Stat Softw 2011;40:1–30.
[45] Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. J Stat Softw 2008;25:1–54.
[46] Yao Q, Chen K, Yao L, Lyu PH, Yang TA, Luo F, Chen SQ, He ZY, Liu ZY. Scientometric trends and knowledge maps of global health systems research. Health Res Policy Syst 2014;12:26.
[47] Boyack KW, Klavans R, Börner K. Mapping the backbone of science. Scientometrics 2005;64:351–74.
[48] Takeda Y, Mae S, Kajikawa Y, Matsushima K. Nanobiotechnology as an emerging research domain from nanotechnology: a bibliometric approach. Scientometrics 2009;80:23–38.
[49] Wang H, Liu M, Hong S, Zhuang Y. A historical review and bibliometric analysis of GPS research from 1991–2010. Scientometrics 2013;95:35–44.
[50] Kulkarni SS, Apte UM, Evangelopoulos NE. The use of latent semantic analysis in operations management research. Decis Sci 2014;45:971–94.
[51] Adler MD, Johnson KB. Quantifying the literature of computer-aided instruction in medical education. Acad Med 2000;75:1025–8.
[52] Hsieh WH, Chiu WT, Lee YS, Ho YS. Bibliometric analysis of patent ductus arteriosus treatments. Scientometrics 2004;60:105–215.
[53] Meneghini R, Packer AL. Is there science beyond English? EMBO Rep 2007;8:112–6.
[54] Porter MF. An algorithm for suffix stripping. Program: Electron Libr Inf Syst 1980;14:130–7.
[55] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Commun ACM 1975;18:613–20.
[56] Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 2004;60:503–20.
[57] Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA 2004;101:5228–35.
[58] Mann HB. Nonparametric tests against trendc. Econom: J Econom Soc 1945;13:245–59.
[59] Guney MS. Evaluation and measures to increase performance coefficient of hydrokinetic turbines. Renew Sustain Energy Rev 2011;15:3669–75.
[60] Coglianese C. Information technology and regulatory policy new directions for digital government research. Soc Sci Comput Rev 2004;22:85–91.
[61] Shulman SW, Schlosberg D, Zavestoski S, Courard-Hauri D. Electronic rulemaking a public participation research agenda for the social sciences. Soc Sci Comput Rev 2003;21:162–78.
[62] Levy KEC, Franklin M. Driving regulation: using topic models to examine political contention in the U.S. trucking industry. Soc Sci Comput Rev 2013;32:182–94.
[63] Bail CA. The cultural environment: measuring culture with big data. Theory Soc 2014;43:465–82.
[64] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 2001;42:177–96.
[65] Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–44.