Soft Computing (2018) 22:7875–7892
https://doi.org/10.1007/s00500-018-3511-4

FOCUS

A bibliometric analysis of text mining in medical research

Tianyong Hao[1] · Xieling Chen[2] · Guozheng Li[3] · Jun Yan[4]

Published online: 6 September 2018

Abstract
Text mining has become an increasingly significant role in processing medical information. The research of text mining enhanced medical has attracted much attention in view from the substantial expansion of literature. This study aims to systematically review the existing academic research outputs of the field from Web of Science and PubMed by using techniques such as geographic visualization, collaboration degree, social network analysis, and topic modeling analysis. Specifically, publication statistical characteristics, geographical distribution, collaboration relations, and research topic are quantitatively analyzed. This study contributes to the text mining enhanced medical research field in a number of ways. First, it provides the latest research status for researchers who are interested in the field through literature analysis. Second, it helps scholars become more aware of the research subfields through hot topic identification. Third, it provides insights to researchers engaging in the field and motivates attention on the relevant research.

Keywords Text mining · Medical · Bibliometric analysis · Topic modeling

1 Introduction

Text mining is the discovery of new, previously unknown information, by the automatic extraction of information from different text resources by computer (Hearst 2003). Text mining methods can be regarded as an extension of data mining to text data (Romero and Ventura 2007), and data mining techniques are also widely applied for image domain processing, e.g., clustering (Zhang et al. 2017), classification (Tan et al. 2018; Tan and Gao 2017), discriminant analysis (Li et al. 2017), and information retrieval (Luo et al. 2017). An important aim of text mining is to shift through large volumes of text for the extraction of patterns and models to be incorporated in intelligent applications (Apte et al. 1998). Usually, text mining is widely applied to the process of structuring the input text, generating patterns within the structured data, as well as evaluating and interpreting the output.[1] In addition, it allows researchers to identify out needed information more efficiently, uncover relations hidden in the sheer volume of available information, and generally shift the burden of information overload from the researchers to the computer by adopting algorithmic, statistical and data management methods to the vast amount of knowledge existing in unstructured texts. On the other hand, medicine is a large and complex domain with abundant synonymy and semantically similar and related concepts (Batet et al. 2011). Most clinical information resources such as electronic medical records and medical knowledge contain considerable amount of information, much of which comes in free text (Meystre and Haug 2005). Therefore, text mining has the great potential in improving health care and advancing medicine through the processing of large amount of medical text data.

Text mining in medical research field has drawn more and more attention from the academia. Especially in recent years,

Communicated by B. B. Gupta.

✉ Xieling Chen
shaylyn_chen@163.com

Tianyong Hao
haoty@126.com

Guozheng Li
gzli@ndctcm.cn

Jun Yan
jun.yan@yiducloud.cn

[1] School of Computer Science, South China Normal University, Guangzhou, China

[2] College of Economics, Jinan University, Guangzhou, China

[3] National Data Center for Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, China

[4] AI Lab, Yidu Cloud (Beijing) Technology Co. Ltd., Beijing, China

[1] https://en.wikipedia.org/wiki/Text_mining.

researchers begin to explore how text mining techniques can be applied in processing medical information. Some examples are as follows. With the basis of dual-process theory and the knowledge adoption model, Jin et al. (2016) introduced a healthcare information adoption model for the exploration of patients' healthcare information- seeking behavior in online communities. Savova et al. (2010) developed and evaluated an open-source natural language processing system to extract information from electronic medical record clinical free text. Lucini et al. (2017) processed data from early emergency department patient records with the application of text mining methods. In addition, a remarkable growth of interest in problems of systems optimization enables the wide application of optimization techniques (e.g., Wang et al. 2018; He et al. 2016, 2017; Lin et al. 2017). Saraswathi and Tamilarasi (2016) proposed an ant colony optimization-based feature selection method for opinion mining classification. Other research interests of medical information processing with text mining techniques include obesity event mining (Chou et al. 2014), sexual event mining (Knight et al. 2012), smoking event mining (Hoek et al. 2014). Consequently, there is an increasing number of academic publications in this interdisciplinary research field.

In the analysis of existing publications, bibliometric analysis is an effective and widely applied strategy. The term bibliometrics is interpreted as "the application of mathematical and statistical methods to books and other media of communication" in 1969 by Glanzel (2003). Used initially in the field of library and information science, bibliometrics has now been widely applied to other areas and has demonstrated significant effectiveness from long-term practice. With the coming of the era of big data, bibliometrics has been a quantitative and qualitative analysis tool of distribution, research hotspots, and tendency for a given research field (Chen et al. 2017a, b; Li and Zhao 2015), as well as a widely accepted tool for identifying future research directions to guide younger researchers (Fu et al. 2010). Benefits of bibliometric analysis are remarkable, e.g., information organization in a specific field (Merigó et al. 2015), scientific developments evaluation in knowledge of a specific subject (Bouyssou and Marchant 2011), research performance comparison across different countries and institutes, and emerging research hotspots identification (Mazloumian 2012). In particular, it has also been applied in interdisciplinary research fields, e.g., natural language processing in mobile computing (Chen et al. 2018a), the natural resource accounting (Zhong et al. 2016), and the fuzzy theory research field (Yu et al. 2018).

To the best of our knowledge, there is no bibliometric analysis of the research field of text mining in medical yet. Therefore, this study conducts a bibliometric analysis on scientific publications retrieved from Web of Science and PubMed during the year 2008–2017 for the exploration of the status and development of the field. The main objectives of this study include: (1) publication statistical characteristics identification, (2) publication geographical distribution exploration, (3) collaboration degrees acquisition, (4) scientific collaboration relation visualization, and (5) current research hotspots and their evolution discovery.

The remainder of this paper is organized as follows: In Sect. 2, we introduce methods and materials. The analyzing results of overall characteristics, collaboration analysis, and topic modeling analysis are presented in Sect. 3. Section 4 is the set of more relevant discussion. This study finishes with conclusions in Sect. 5.

2 Materials and methods

2.1 Materials

Web of Science (WoS) and PubMed are the most commonly used databases in the academia. WoS is the most authoritative citation database with publications of high quality, while PubMed is the largest data source on life sciences and biomedical topics. In our study, to make full use of their complementary advantages, we use all the relevant publications from these two databases.

First of all, a list of keywords (Table 8 in "Appendix") related to "text mining" was determined by relevant domain experts in the field. In WoS Core Collection database, Topic Subject was used as retrieval field. "Science Citation Index Expanded (SCI-E)" and "Social Sciences Citation Index (SSCI)" were set to be the citation indexes to ensure publication quality. 2284 publications between 2008 and 2017 with "Article" and "Proceedings paper" as article types, and WoS category containing terms "Health", "Medicine", "Medical", "Clinical", and "Nursing" were identified. Furthermore, after manually removing 62 irrelevant publications with "image" or "imaging" containing in title, 2222 publications were finally identified out.

As for PubMed database, Title/Abstract was used as search column. 6331 publications between 2008 and 2017 were retrieved, where 3346 were in "Journal Article" type with "humans" as species and "MEDLINE" as journal category. Similarly, after removing 165 irrelevant publications with "image" or "imaging" containing in title, 3283 publications were identified out. 1967 publications were finally obtained for analysis after removing 1316 publications that were already contained in WoS through manual review according to publication title, author, publication year, and publication source.

The raw data of the totally 4189 publications from WoS and PubMed were downloaded as both plain text and XML format. Key elements including title, publication source, published year, abstract, author address, author keywords, and Keywords Plus/PubMed MeSH were extracted. Manual

information supplement was conducted. Finally, according to the author address information, the corresponding institutes and countries were identified. The statistical characteristics of the publications are shown in Table 1.

2.2 Collaboration degree analysis

The collaboration degree is a measure of scientific research's connective relation to the level of authors, institutes, and countries (Zhang et al. 2016). The calculations of author's collaboration degree, institute's collaboration degree, and country's collaboration degree are expressed in Eq. (1) in order (Wei et al. 2013).

$$C_{Ai} = \frac{\sum_{j=1}^{N} \alpha_j}{N}, \quad C_{Ii} = \frac{\sum_{j=1}^{N} \beta_j}{N}, \quad C_{Ci} = \frac{\sum_{j=1}^{N} \gamma_j}{N} \tag{1}$$

In the equation, C_{Ai}, C_{Ii}, and C_{Ci} represent the author, institute, and country's collaboration degree of the i year. α_j, β_j, and γ_j indicate the number of authors, institutes and countries for each publication. N donates the annual total number of publications in the research field.

2.3 Social network analysis (SNA)

Complex social systems are usually formed from the interaction of social actors with each other at multiple physical or social interfaces and across layers. A complex social system can be expressed through a social network with "actors" as nodes and "interactions" as link lines. Social network is thus a collection of social actors and their interaction relations. The relations between nodes represent similarities, interactions, social relations, and flows (Borgatti et al. 2009). It is very interesting for scholars and managers to investigate how complex social systems change and evolve to emerge dynamic patterns. By studying the social network, dynamic patterns of interactions emergence and their evolution with time can then be explored. The social network analysis (SNA) has formed a quantitative analysis ground on the development of the mathematical method and the graph theory and thus provides a quantitative assessment on relations between social actors.

In this paper, we apply SNA to explore the collaboration relations for specific countries/regions, institutes, and authors in the research field. Collaboration relations between them can be visualized by SNA by counting the number of times they (e.g., two countries/regions) appear in the same publication together. In the network, each country/region, institute, or author is presented as a node with the node size representing its proportion of publications. The node color denotes the continent or country. The thickness of each line indicates collaboration strength between two countries/regions, institutes, or authors. One could explore the collaboration relations for specific countries/regions, institutes, or authors by clicking the nodes.

2.4 Latent Dirichlet allocation (LDA)

As an emerging quantitative method to assessing substantial textual data, topic modeling extracts semantic information from a collection of texts with the use of statistical algo-

Table 1 Statistical characteristics of the publications

Characteristics	Statistics
Total number of publication from WoS/PubMed	2222/3283
Overlap number of publication from WoS and PubMed	1316
Total unique number of publication	4189
Number of unique countries/regions	88
Number of unique first authors/last authors/total authors	3538/3241/13,717
Number of unique first institutes/total institutes	1789/3208
Top 10 terms in author keywords, Keywords Plus and PubMed MeSH	Human (2.82%); health (2.04%); language (1.85%); information (1.49%); analysis (1.33%); processing (1.31%); natural (1.21%); data (1.12%); mining (1.09%); system (0.94%)
Top 10 terms in titles	Analysis (1.35%); health (1.35%); clinical (1.08%); using (1.06%); text (0.95%); information (0.94%); discourse (0.86%); study (0.80%); language (0.77%); patient (0.75%)
Top 10 terms in abstracts	Information (0.80%); data (0.74%); study (0.67%); method (0.66%); analysis (0.65%); patient (0.61%); health (0.60%); result (0.59%); system (0.58%); using (0.54%)

rithms. The first topic model, probabilistic latent semantic indexing (pLSI), was proposed by Hofmann (1999). It models the probability of each co-occurrence as a mixture of conditionally independent multinomial distributions, where the mixture components can be viewed as representation topics. An improved three-layer Bayesian model, latent Dirichlet allocation (LDA), was developed by Blei et al. (2003), which takes Dirichlet distribution as the prior distribution and reduces the parameter number to only one. In LDA, documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words, and topics are assumed to be uncorrelated. In order to reduce the computing time and the required memory (Blei and Lafferty 2007; Teh et al. 2005), some various extensions such as Correlated Topic Models and Hierarchical Dirichlet Process have been proposed based on the original LDA model in recent years. LDA and its extensions have been widely applied in scientometric research for discovering semantic structures and latent topics in a discipline or measuring the relations of multiple disciplines (Lu and Wolfram 2012; Nichols 2014; Yau et al. 2014). LDA defines the following terms:

(1) A *word* is an item from a vocabulary indexed by $\{1, \ldots, V\}$;
(2) A *document* is a sequence of N words denoted by $d = (w_1, \ldots, w_N)$;
(3) A *corpus* is a collection of M documents denoted by $D = \{d_1, \ldots, d_M\}$.

LDA assumes the following generation process:

(1) The term distribution β containing the probability of a word occurring in a given topic is determined by $\beta \sim \text{Dirichlet}(\delta)$;
(2) The proportions θ of the topic distribution for a document d are determined by $\theta \sim \text{Dirichlet}(\alpha)$;
(3) For each word w_i in the document d, a topic is chosen by the distribution $z_i \sim \text{Multinomial}(\theta)$, and a word is chosen from a multinomial probability distribution conditioned on the topic z_i : $p(w_i|z_i, \beta)$.

The log-likelihood for one document $d \in D$ in variational expectation–maximization (VEM) estimation is given by Eq. (2).

$$\begin{aligned} l(\alpha, \beta) &= \log(p(d|\alpha, \beta)) \\ &= \log \int \left\{ \sum_z \left[\prod_{i=1}^{N} p(w_i|z_i, \beta) p(z_i|\theta) \right] \right\} p(\theta|\alpha) \mathrm{d}\theta \end{aligned} \tag{2}$$

Gibbs sampling is a Markov chain Monte Carlo method (Finkel et al. 2005) aiming at constructing a Markov chain converging to the target probability distribution in the high-dimensional model and then extracting the sample distribution closest to the target probability distribution. The log-likelihood for Gibbs sampling is as Eq. (3).

$$\begin{aligned} \log(p(d|z)) = k \ \log & \left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V} \right) \\ & + \sum_{K=1}^{k} \left\{ \left[\sum_{j=1}^{V} \log \left(\Gamma(n_K^{(j)} + \delta) \right) \right] \right. \\ & \left. - \log \left(\Gamma(n_K^{(\cdot)} + V\delta) \right) \right\} \end{aligned} \tag{3}$$

The topic modeling analysis in this study follows the following steps:

(1) Weights 0.4, 0.4 and 0.2 determined in our former experiment (Chen et al. 2018b) are assigned to segmented author keywords, Keywords Plus and PubMed MeSH, publication title, and abstract, respectively.
(2) Since frequent terms usually provide just limited information, as most terms in Table 1 are either trivial in text mining in medical sector, such as "health", "mining", "clinical", and "patient", or trivial in general scientific publications, such as "analysis", "using", "study", and "method". Thus, we perform a transformation on the corpus using Term Frequency-Inverse Document Frequencies (TF-IDF) to penalize frequent terms occurring in many publications (Salton et al. 1975; Robertson 2004). We calculate the TF-IDF values of all terms and sort them according to the values. A threshold is determined as 0.1 empirically by manually examining these ranked terms. Terms with a TF-IDF value no more than the threshold are removed.
(3) Through sampling, 17 different topic numbers are set to $c(2 : 10, 15, 20, 30, 40, 50, 80, 150, 250)$. For each topic number, tenfold cross-validation is used to evaluate model performance. Perplexity criteria are used to select optimal topic number (Blei et al. 2003). α for Gibbs sampling is initialized as the mean value of α values for model fitting using VEM with the optimal topic number.
(4) We then adopt Gibbs sampling and VEM method to estimate the LDA model with the optimal topic number and an initialized α.
(5) By matching the topics detected by VEM and Gibbs sampling based on Hellinger distance as Eq. (4), the best matches with the smallest distance can be identified. In Eq. (4), P and Q denote two probability measures.

$$H^2(P, Q) = \frac{1}{2} \int \left(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q} \right)^2 \tag{4}$$

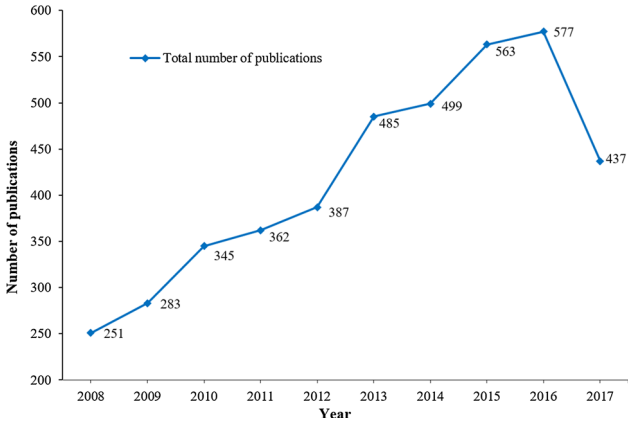

Fig. 1 Publication number distribution by year

Recently, with the development and the availability of accessible software, topic modeling and other text mining approaches are becoming more approachable. Open-access options include some R, Python and Java packages. In this study, the topic modeling process is conducted with an R package called *Topicmodels* offered by Grün and Hornik (2011). The package requires a text mining front-end addition, such as the R package, *tm* (Feinerer et al. 2008).

3 Result

3.1 General publication statistics

3.1.1 Publication with year

The publication number distribution by year is demonstrated in Fig. 1. The publication number keeps increasing by year from 251 (year 2008) to 577 (year 2016), but experiences decline in 2017 to 437. The decline may be caused by time lag of some publications to be included in the databases in 2017. The annual growth rate reaches 7.31% on average, while the rate reaches up to 25.32% from 2012 to 2013, witnessing the research upsurge in 2013.

3.1.2 Productive publication sources

The top 20 productive publication sources in the research field are presented in Table 2. These publication sources together contribute 38.46% of the total publications. All the 20 publication sources are journals except *AMIA Annual Symposium Proceedings* and *Studies in Health Technology and Informatics*, which are two top conferences in medical informatics. The most productive journal is *Journal of Biomedical Informatics* with 297 publications, followed by *Journal of the American Medical Informatics Association* with 212 publications, and *PLOS One* with 146 publications. All the 18 journals on the list have an IF of over 1.00. *Nucleic Acids Research* possesses the highest IF as 11.561, reflecting high quality of its publications. Interestingly, the causality among total publications in one journal and IF is not found in such a field. This may due to the fact that most journals with higher reputation actually cover many research fields, in which text mining in medical research is just one of them.

3.1.3 Geographical distribution

The analysis of world geographical distribution is based on author institute address. All the authors participating in each publication are considered. Also, since an author may be affiliated with more than one institutes, all the countries/regions and institutes of authors are used for the geographical distribution analysis.

The 4189 publications are from 88 countries/regions. Figure 2 illustrates geographical distribution of the publications. The top 4 countries are: the USA (1680 publications), UK (546 publications), Canada (314 publications), and China (285 publications). The publication number of the USA is nearly 3 times than that of the second productive country, indicating its dominant position in the research field. As for the top 20 countries/regions, most are developed countries except China (rank 4th) and Brazil (rank 7th), reflecting their huge enthusiasm in the research field.

Since the publications are mainly distributed in the 5 countries, we further explore the annual publication distributions for these countries, as shown in Fig. 3. The number of publications for the USA is on the whole presenting an upward trend in fluctuation from 82 in 2008 to 241 in 2015, but dwindles since 2015. As for UK, the publication number presents slow growth before 2013, and a slight decline appears from 2013 to 2015. After that, a sharp growth is noticeable in 2016. As for the other three countries, the publication numbers are on the whole presenting upward trends in fluctuation with years going on. In short, the research field has received increasing attention from these countries.

3.1.4 Productive authors and institutes

The top 20 productive authors are listed in Table 3. All of them come from the USA except *Darmoni, Stefan J.* from France, which again demonstrates the USA's high productivity in the research field. The top 3 are all from the USA, including *Denny, Joshua C.* (52 publications), *Xu, Hua* (52 publications), and *Savova, Guergana K.* (38 publications), followed by *Liu, Hongfang* (33 publications) from the USA and *Lu, Zhiyong* (32 publications) from the USA. Most of the 20 authors serve more as last authors than as first authors, and almost all collaborate with other authors in all their publications except *Denny, Joshua C.* and *Chute, Christopher G.*

Table 2 Top 20 productive publication sources in the research

R	Publication sources	TP	P%	IF
1	Journal of Biomedical Informatics	297	7.09	2.882
2	Journal of the American Medical Informatics Association	212	5.06	4.270
3	PloS one	146	3.49	2.766
4	AMIA Annual Symposium Proceedings	113	2.70	N/A
5	BMC Bioinformatics	105	2.51	2.213
6	Studies in Health Technology and Informatics	92	2.20	N/A
7	BMC Medical Informatics and Decision Making	75	1.79	2.134
8	Journal of Medical Internet Research	73	1.74	4.671
9	International Journal of Medical Informatics	63	1.50	2.957
10	Database-The Journal of Biological Databases and Curation	61	1.46	3.978
11	Qualitative Health Research	53	1.27	2.413
12	Social Science & Medicine	44	1.05	3.007
13	Methods of Information in Medicine	41	0.98	1.531
14	Nucleic Acids Research	40	0.95	11.561
15	Artificial Intelligence in Medicine	39	0.93	2.879
16	Bioinformatics	38	0.91	5.481
17	Health	32	0.76	1.413
18	Behavior Research Methods	30	0.72	3.597
19	Nursing Inquiry	29	0.69	1.159
20	Journal of Medical Systems	28	0.67	2.098

R, rank; TP, total publications; P%, proportion of publication number; IF, impact factor 2017

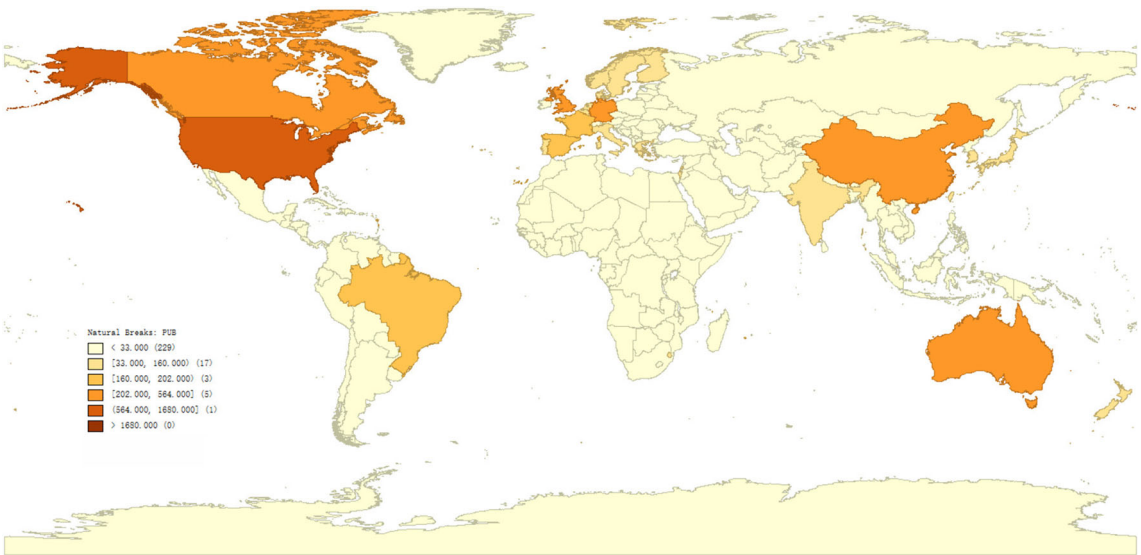

Fig. 2 Geomap of publications by countries

Table 4 depicts the most productive first authors and last authors. All the 9 productive first authors are from the USA. All the 9 productive last authors come from the USA except *Darmoni, Stefan J.* from France. The top 3 first authors are *Pakhomov, Serguei V. S.* (11 publications), *Denny, Joshua C.* (10 publications), and *Meystre, Stephane M.* (9 publications). The top 3 last authors are *Xu, Hua (26 publications)*, *Lu, Zhiyong* (23 publications), and *Denny, Joshua C.* (17 publications). It is worth noting that *Denny, Joshua C.* and *Xu, Hua* appear in both two lists, which to a certain degree demonstrates their influence in the research.

3208 institutes from 88 countries have performed researches in the field. Table 5 shows the most productive institutes. Most of the 19 institutes are from the

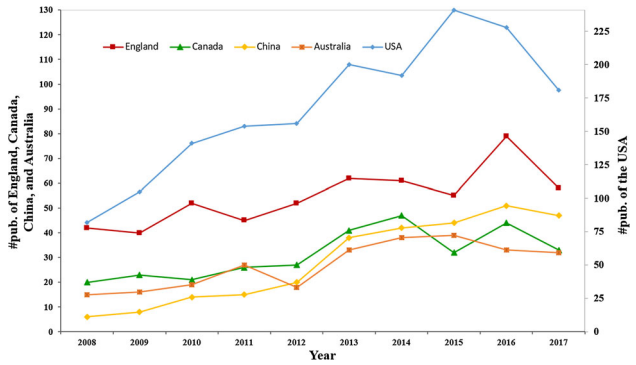

Fig. 3 Publication distributions by year for the top 5 countries

USA except *University of Manchester* from UK, *University of Toronto* from Canada, and *University of Sao Paulo* from Brazil. The top 5 are all from the USA, including *National Institutes Health* (120 publications), *University of Utah* (110 publications), *Vanderbilt University* (93 publications), *Harvard University* (86 publications), and *Mayo Clinic* (84 publications). The first institute percentage for most of the institutes is above 50% except *University of California San Diego* (38.78%) and *University of Texas Health Science Center Houston* (40.91%), indicating the leading position of the top productive institutes. Most institutes collaborate a lot with other institutes with an average collaboration percentage up to 78.98%, especially *Salt Lake City VA Health Care System* (96.08%).

3.2 Collaboration analysis

3.2.1 Collaboration degree

Figure 4 presents the annual collaboration degrees at three perspectives. The auctorial collaboration degree increases apparently, up to 5.29. In contrast, institutional and international collaboration degrees are steady and relatively low, especially the international collaboration degree. This reflects that the authors tend to collaborate more with those within the same country or institute. The three average degrees are 4.51, 2.26, and 1.30, respectively, that is, 4.51 authors, 2.26 institutes, and 1.30 countries participate in one publication averagely.

Table 3 The most productive authors in the research

R	Authors	C	TP	FP	LP	R	Authors	C	TP	FP	LP
1	Denny, Joshua C.	USA	52	10	17	10	Samore, Matthew H.	USA	20	0	7
1	Xu, Hua	USA	52	7	26	12	Darmoni, Stefan J.	FR	19	1	14
3	Savova, Guergana K.	USA	38	8	9	12	South, Brett R.	USA	19	3	3
4	Liu, Hongfang	USA	33	1	14	14	Jonnalagadda, Siddhartha R.	USA	18	6	4
5	Lu, Zhiyong	USA	32	1	23	14	Kohane, Isaac S.	USA	18	0	3
6	Chapman, Wendy W.	USA	26	2	11	14	Meystre, Stephane M.	USA	18	9	6
7	Chute, Christopher G.	USA	21	2	10	17	Carrell, David S.	USA	17	4	0
7	Rindflesch, Thomas C.	USA	21	1	15	18	Elhadad, Noemie	USA	16	0	8
7	Uzuner, Ozlem	USA	21	6	10	18	Shen, Shuying	USA	16	0	0
10	Pakhomov, Serguei V. S.	USA	20	11	0	18	Yu, Hong	USA	16	0	14

R, rank; C, country of author (USA, America; FR, France); TP, total publications; FP, number of publications as the first author; LP, number of publications as the last author

Table 4 The most productive first authors and last authors in the research

R	First authors	C	#pub.	R	Last authors	C	#pub.
1	Pakhomov, Serguei V. S.	USA	11	1	Xu, Hua	USA	26
2	Denny, Joshua C.	USA	10	2	Lu, Zhiyong	USA	23
3	Meystre, Stephane M.	USA	9	3	Denny, Joshua C.	USA	17
4	Botsis, Taxiarchis	USA	8	4	Rindflesch, Thomas C.	USA	15
4	Savova, Guergana K.	USA	8	5	Darmoni, Stefan J.	FR	14
4	Xu, Rong	USA	8	5	Liu, Hongfang	USA	14
7	Roberts, Kirk	USA	7	5	Yu, Hong	USA	14
7	Speier, William	USA	7	8	Khorasani, Ramin	USA	13
7	Xu, Hua	USA	7	8	Weng, Chunhua	USA	13

R, rank; C, country of author (USA, America; FR, France)

Table 5 The most productive institutes in the research

R	Institutes	C	TP	FP (%)	CP (%)	R	Institutes	C	TP	FP(%)	CP(%)
1	National Institutes Health	USA	120	71 (59.17)	81 (67.50)	11	University of Sao Paulo	BR	57	30 (52.63)	39 (68.42)
2	University of Utah	USA	110	80 (72.73)	92 (83.64)	12	Brigham and Women's Hospital	USA	56	35 (62.50)	52 (92.86)
3	Vanderbilt University	USA	93	62 (66.67)	67 (72.04)	13	Salt Lake City VA Health Care System	USA	51	39 (76.47)	49 (96.08)
4	Harvard University	USA	86	43 (50.00)	79 (91.86)	14	University of California San Diego	USA	49	19 (38.78)	41 (83.67)
5	Mayo Clinic	USA	84	47 (55.95)	65 (77.38)	14	University of Pittsburgh	USA	49	28 (57.14)	42 (85.71)
6	Columbia University	USA	69	47 (68.12)	42 (60.87)	16	Indiana University	USA	46	28 (60.87)	38 (82.61)
7	University of Washington	USA	64	36 (56.25)	42 (65.63)	17	Massachusetts General Hospital	USA	45	25 (55.56)	38 (84.44)
8	Stanford University	USA	62	36 (58.06)	42 (67.74)	18	University of Texas Health Science Center Houston	USA	44	18 (40.91)	41 (93.18)
9	University of Manchester	UK	58	33 (56.90)	46 (79.31)	19	University of Michigan	USA	39	23 (58.97)	28 (71.79)
9	University of Toronto	CA	58	37 (63.79)	44 (75.86)						

R, rank; C, country of institute (USA, America; UK, England; CA, Canada; BR, Brazil); TP, total publications; FP(%), number and proportion of publications as the first institute; CP(%), number and proportion of collaborated publications

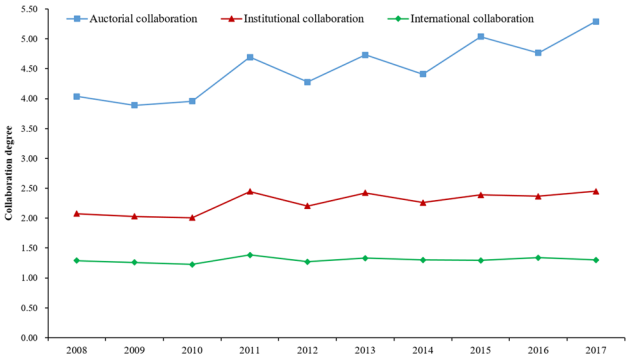

Fig. 4 Annual collaboration degree distributions

3.2.2 Collaboration visualization

We further visualize the collaborations in three perspectives using SNA. A collaboration network[2] for 88 countries/regions with 88 nodes and 516 edges is shown in Fig. 5. The USA (the largest node in brown color) has the most collaborations with other countries/regions. The USA–England collaboration (the thickest line) ranks at the first, followed by the USA–China and the USA–Canada collaborations. The collaboration network[3] among 67 institutes with the number of publications $\geq$ 20 is shown in Fig. 6 with 67 nodes and 564 edges. Forty of the 67 institutes come from the USA, and the collaboration network among them (the nodes in blue color) is very dense. The collaboration network[4] of 81 authors with publications $\geq$ 10 is as Fig. 7. The node count and edge count are 81 and 291. Among the nodes, 8 are sparse nodes including "Xu, Rong", "Botsis, Taxiarchis", "Khorasani, Ramin", "Stewart, Robert", "Darmoni, Stefan J", "Nenadic, Goran", "Dai, Hong-Jie", and "Zweigenbaum, Pierre" due to the lack of collaborations with other author nodes. Most of the authors (74.07%) come from the USA, and the collaboration network among them (the nodes in blue color) is very dense.

3.3 Topic modeling analysis

Terms with TF-IDF values more than the threshold 0.1 are employed in the topic modeling analysis. Table 6 lists the top 20 frequent terms. Apparently, terms listed in the table are more specific terminology of text mining in medical research issues. There are several nursing-related terms with high occurrence numbers such as "Nursing" (1002) and "Nurse" (871), suggesting the significance of nursing research using text mining techniques. Terms such as "Breast", "Depression", "Sexual", and "Obesity" reflect specific medical issues in the research. "Chinese" (422) is the only country appear-

[2] http://www.zhukun.org/haoty/resources.asp?id=JSC_cocountry.

[3] http://www.zhukun.org/haoty/resources.asp?id=JSC_coaffiliation.

[4] http://www.zhukun.org/haoty/resources.asp?id=JSC_coauthor.

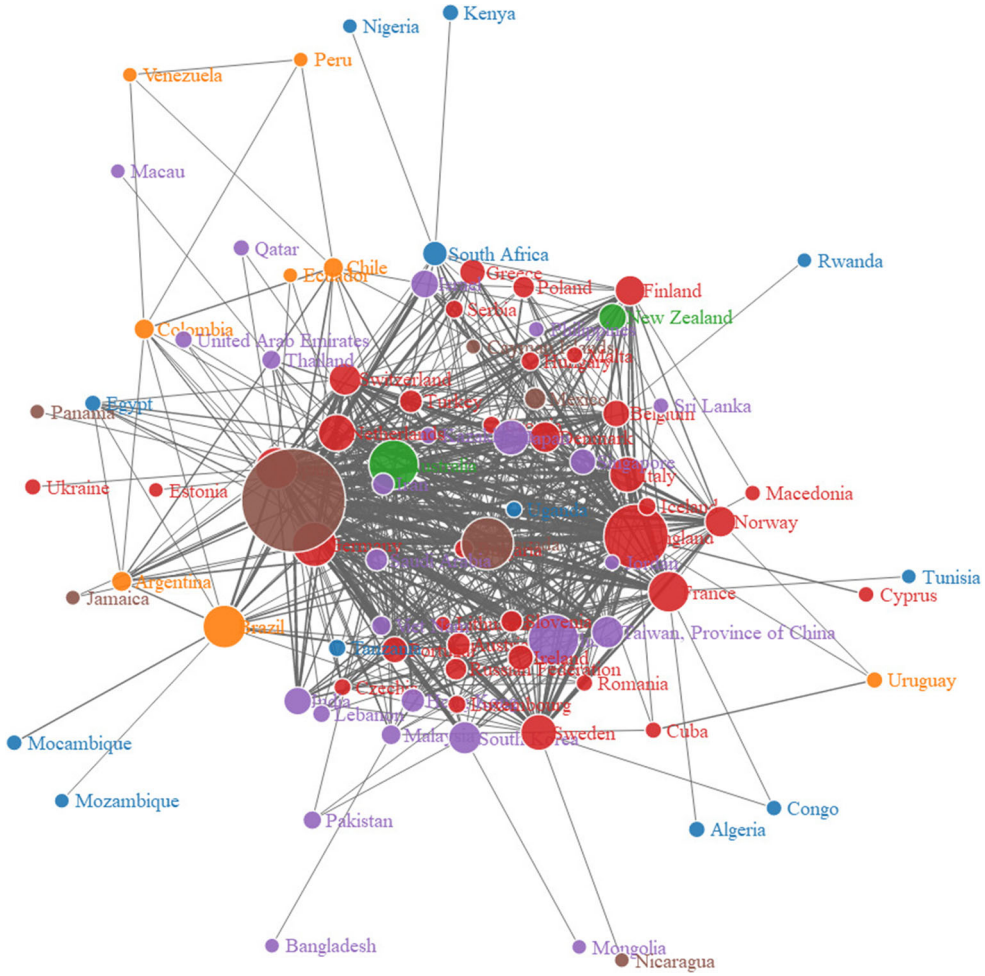

Fig. 5 Collaboration network of 88 countries/regions (the orange nodes represent countries/regions from South America, blue for Africa, green for Oceania, red for Europe, purple for Asia, and brown for North America)

ing in the table, indicating that China has been focusing on text mining in medical research during these years.

3.3.1 Topic generation

We employ LDA model to reveal the latent intellectual topics in the literature corpus based on terms selected by TF-IDF. To fit the model, we should determine the parameters including the number of topics and the α. Hence, we compute the perplexities of a set of models with different numbers of topics to find a minimum in the tenfold cross-validation. Figure 8 presents the perplexities of models with different numbers of topics. The result indicates that the data are best accounted for by a model incorporating 40 topics. The α is set to the mean value 0.01857649 in the cross-validation fitted using VEM. Using the parameters, we estimate the LDA model using Gibbs sampling.

We assign potential theme to each topic by semantics analysis of representative terms in each topic, as well as reviewing text intention of the corresponding publications. The order of topics is determined based on Hellinger distance. Specifically, Topic 31 is the best matching topic, and Topic 22 ranks at 2nd. Due to space limitation, Table 7 only displays the top 10 best matching topics with the most frequent terms. Each publication is assigned to the most likely topic based on posterior probability. We then obtain a topic distribution by integrating topic proportions for all the publications. The 4 most frequent research topics are: Topic 16 (3.91%), Topic 24 (3.31%), Topic 9 (3.29%), and Topic 31 (3.14%), while the 4 least frequent research topics are: Topic 25 (1.91%), Topic 14 (1.88%), Topic 33 (1.88%), and Topic 37 (1.88%).

3.3.2 Topic cluster analysis and trend analysis

We use the hierarchical cluster analysis to perform the cluster analysis of the 40 topics. One way of measuring topic similarity is based on term-level similarity, meaning that topics

Fig. 6 Collaboration network of 67 institutes (different colors of nodes represent different countries/regions, e.g., the blue nodes represent institutes from the USA, orange for England, purple for Australia)

may contain some of the same terms. Another way of topic similarity measuring is by document-level similarity, meaning that topics may appear in some of the same documents. The clustering results based on cosine similarity for the two measurements are shown in Figs. 9 and 10. In the figures, lower location of connecting line means that topics are more similar.

Identifying emerging research topics can provide valuable insights into the development of the research field (Jiang et al. 2016). Therefore, we then explore the annual publication proportions of the 40 research topics, as shown in Fig. 11. We use a nonparametric trend test called MannKendall test (Mann 1945) to examine whether increasing or decreasing trends are existing in the 40 topics. Test results show that fourteen topics, including Topic 2, Topic 4, Topic 11, Topic 13, Topic 15, Topic 20, Topic 22, Topic 25, Topic 26, Topic 27, Topic 32, Topic 33, Topic 36, and Topic 40, present a statistically significant increasing trend at the two-sided $p = 0.05$ level.

4 Discussion

Scientific literature related to text mining in medical research is an abundant and reliable data pool, from which we can understand the major academic concerns about the research field and hence deploy a proper development strategy. Based on the 4189 publications collected from the WoS and PubMed databases, the analysis focuses on literature characteristics, geographical publication distribution, collaboration relations, as well as research topic. Results of this exploration present a comprehensive overview and an intellectual structure of the research, especially research topics, from 2008 to 2017.

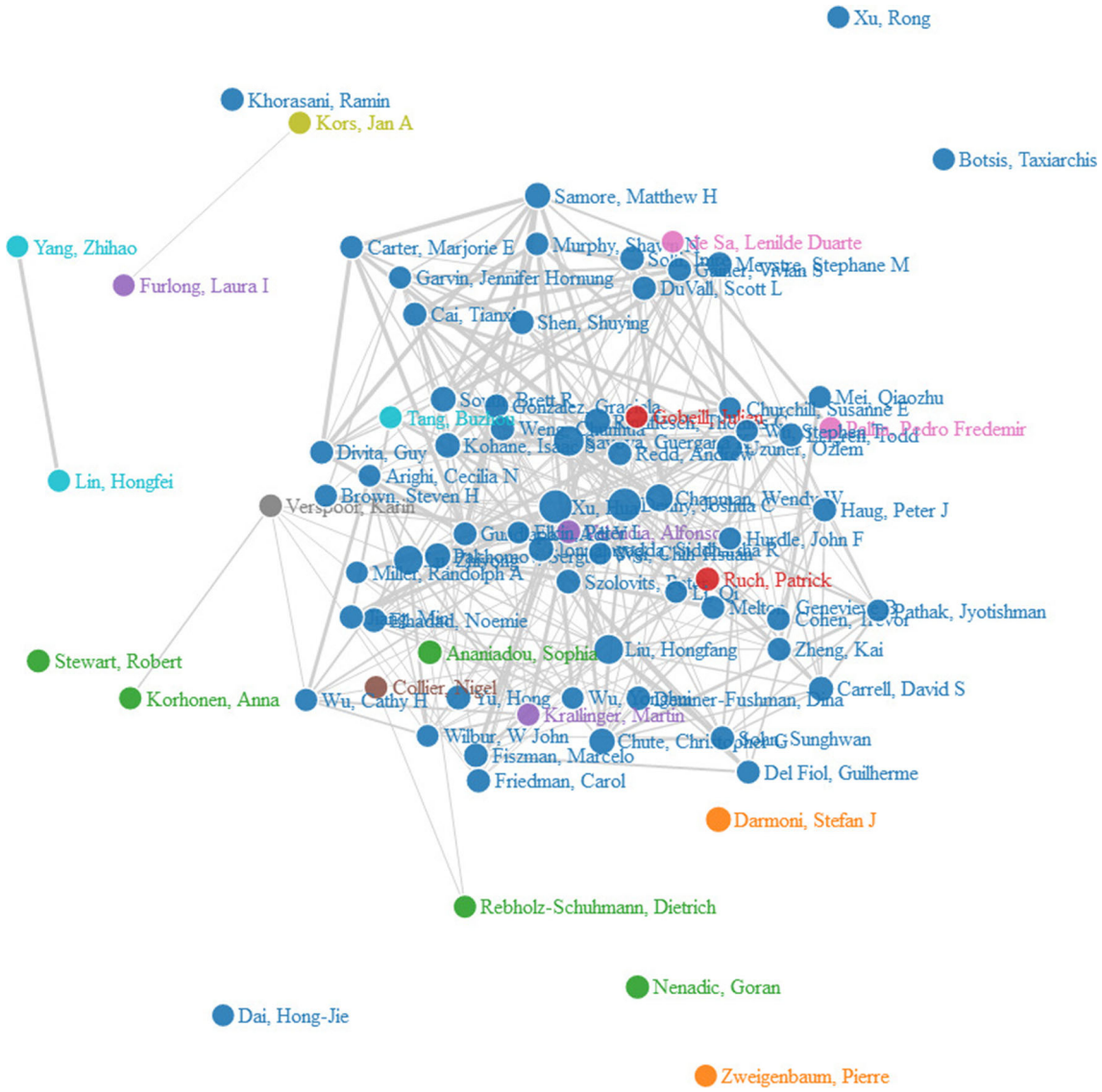

Fig. 7 Collaboration network of 81 authors (different colors of nodes represent different countries/regions, e.g., the blue nodes represent institutes from the USA, green for England, light blue for China)

Table 6 Top 20 most frequent terms

R	Stemmed terms	Frequency	R	Stemmed terms	Frequency
1	Child	1110	11	Chinese	422
2	Nursing	1002	12	Infant	409
3	Nurse	871	13	Breast	392
4	Speech	793	14	Reading	382
5	Student	759	15	Depression	364
6	Radiology	556	16	Sexual	339
7	Sentiment	496	17	Parent	333
8	Segmentation	469	18	Translation	327
9	Men	466	19	Twitter	315
10	Memory	459	20	Obesity	310

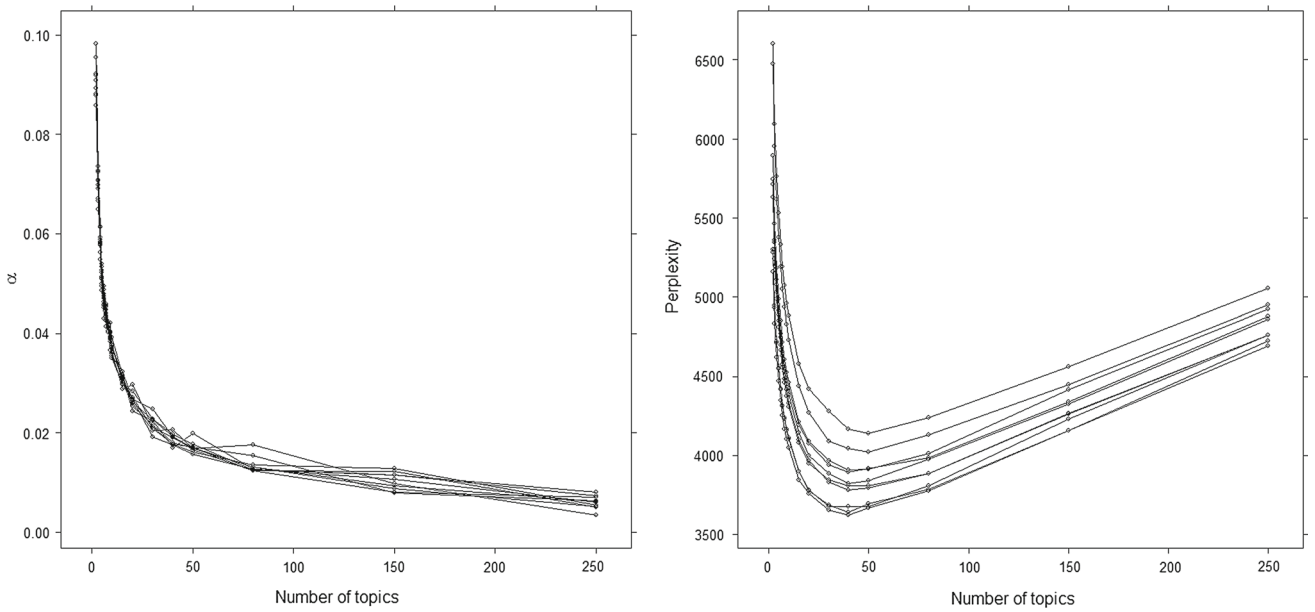

Fig. 8 Left: estimated α value for the models fitted using VEM. Right: perplexities of the test data for the models fitted by using Gibbs sampling. Each line corresponds to one of the folds in the tenfold cross-validation

Table 7 Top 20 most frequent terms for the top 10 best matching topics

Topic	Potential theme	Percentage	Top frequent terms
31	Speech event	3.14	Speech; Segmentation; Infant; Cue; Aphasia; Prosodic; Child; Prosody; Transitional; Listener; Distal; Consonant; Vowel; Sound; Biological Expression Language; Clause; Newborn; Phonological; Rhythm; Aphasic
22	Sexual event	2.39	Men; Sexual; Hiv; Masculinity; Sexuality; Hiv/Aids; Africa; Gay; African; Prostate; Reproductive; Youth; Masculine; Lesbian; Condom; Contraceptive; Gist; Tanzania; Violence; Help-seeking
5	Alzheimer event	2.78	Memory; Schizophrenia; Dementia; Alzheimer; Connectivity; Fluency; Alzheimer Disease; Functional Magnetic Resonance Imaging; Short-term; Epilepsy; Executive; Episodic; Speech; Associative; Neuron; Repetition; Mild; Spiritual; Seizure; Schizophrenic
21	Parenting event	2.59	Child; Parent; Caregiver; Mother; Air; Pollution; Parental; Distress; Neonatal; Infant; Parenting; Attention-deficit Hyperactivity Disorder; Bereavement; Newborn; Attachment; Father; Hyperactivity; Respiratory; Preterm; Neonatal Intensive Care Unit
16	Nursing event	3.91	Nursing; Nurse; Violence; Domestic; Victim; Intensive Care Units; Forensic; Intimate; Newspaper; Justice; Ward; Crime; Aggression; Delirium; Manager; Nurse-patient; Abuse; Caregiver; British; Pornography
9	Education event	3.29	Student; Curriculum; Professionalism; Teacher; Interprofessional; Enzyme; Dilemma; Emotion; Problem-based Learning; Cytochrome; Residency; Graduate; Problem-based; Essay; Taiwanese; Governmentality; Cytochromep450; The Comprehensive Enzyme Information System; Game; Reform
18	Reading event	3.02	Reading; Eye; Movement; Music; Chinese; Fixation; Readability; Emotion; Saccade; Segmentation; Ancestry; Continental; Asian; Scene; Embedding; Literacy; Decoding; Musical; FMRI; Diagram
6	Disease diagnosis	2.78	Radiology; Computed Tomography; Pulmonary; Pneumonia; Nodule; Systematic Review; Abdominal; Magnetic Resonance Imaging; Permutation Entropy; Lung; Venous; Incidental; Venous Thromboembolism; Angiography; Discrepancy; Radlex; Thromboembolism; Comparative Effectiveness Research; D-dimer; Postoperative
27	Heart disease	2.32	Heart; Stroke; Coronary; Cognitive Disorders; Artery; Readmission; Rehabilitation; Fraction; Ejection; Ventricular; Coronary Artery Disease; Heart Failure; Echocardiography; Bowel; Coronary Heart Disease; Congestive Heart Failure; Alopecia Areata; Inflammatory Bowel Disease; Comorbidities; Stenosis
30	Parturition event	2.01	Pregnancy; Birth; Mother; Midwifery; Midwife; Maternity; Pregnant; Breastfeeding; Childbirth; Maternal; Infant; Feeding; Postpartum; Milk; Baby; Fetal; Zealand; Gestational; Prenatal; Antenatal

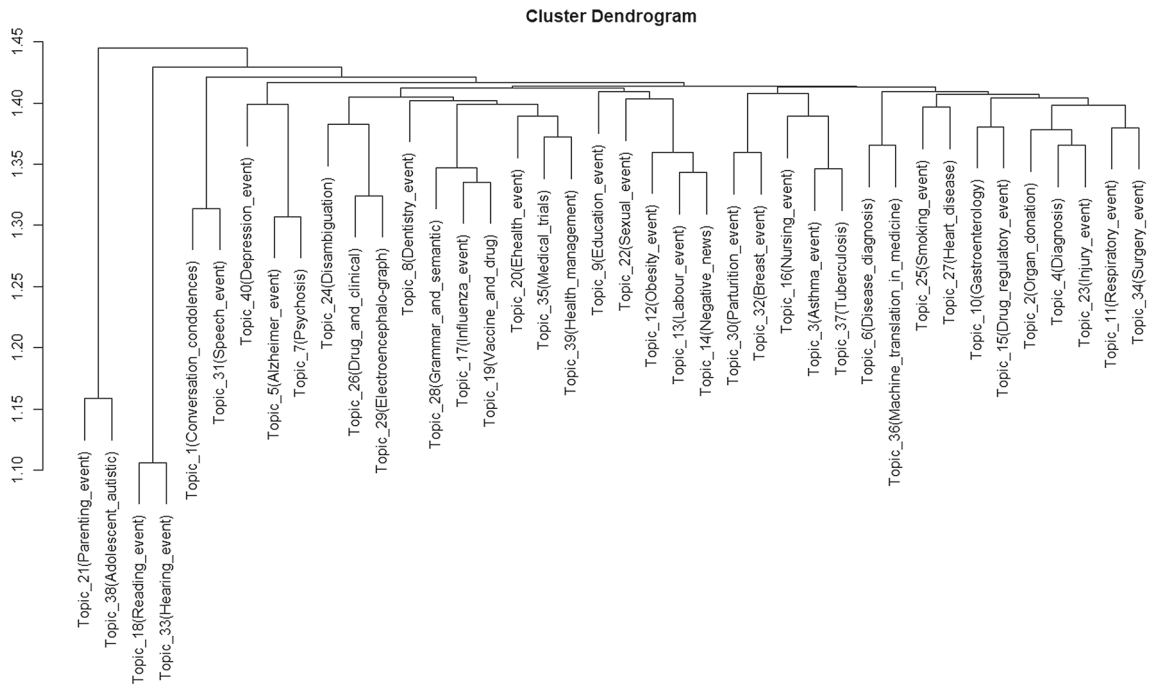

Fig. 9 Dendrogram of the term-level similarity clustering

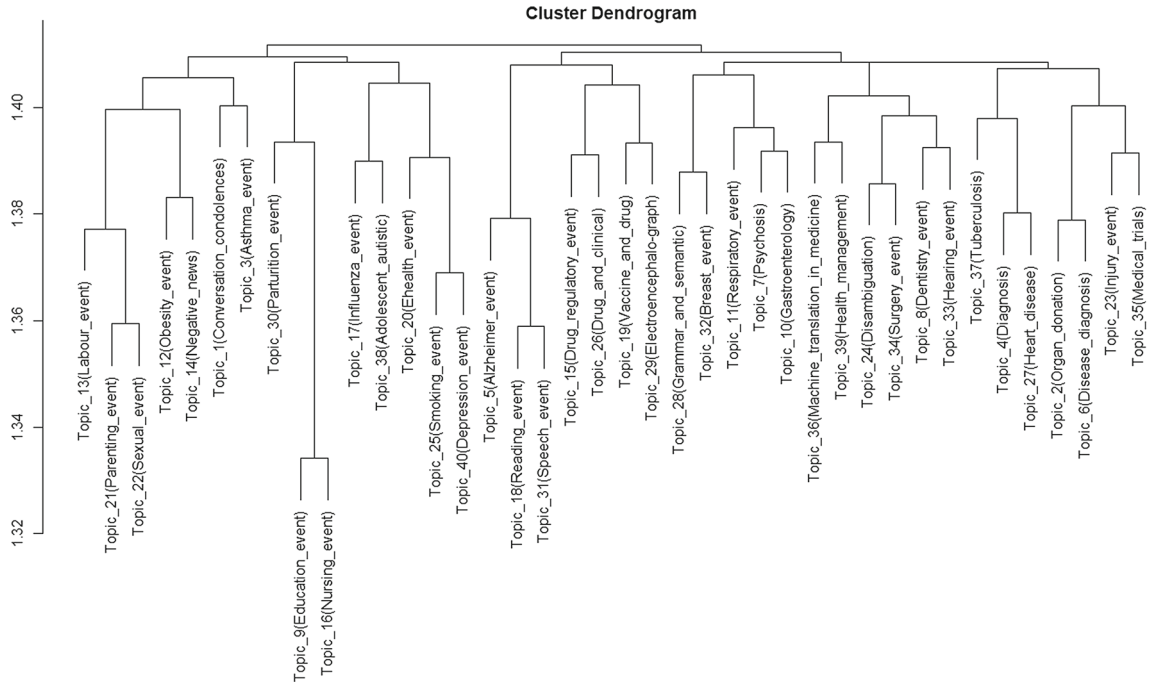

Fig. 10 Dendrogram of the document-level similarity clustering

The rapid growth of relevant research publications reveals the vigorous development of text mining in medical research in recent years. The top 20 productive publication sources contribute 38.46% of the total publications, with *Journal of Biomedical Informatics* as the most productive one. The USA dominates in the field with a publication number far more

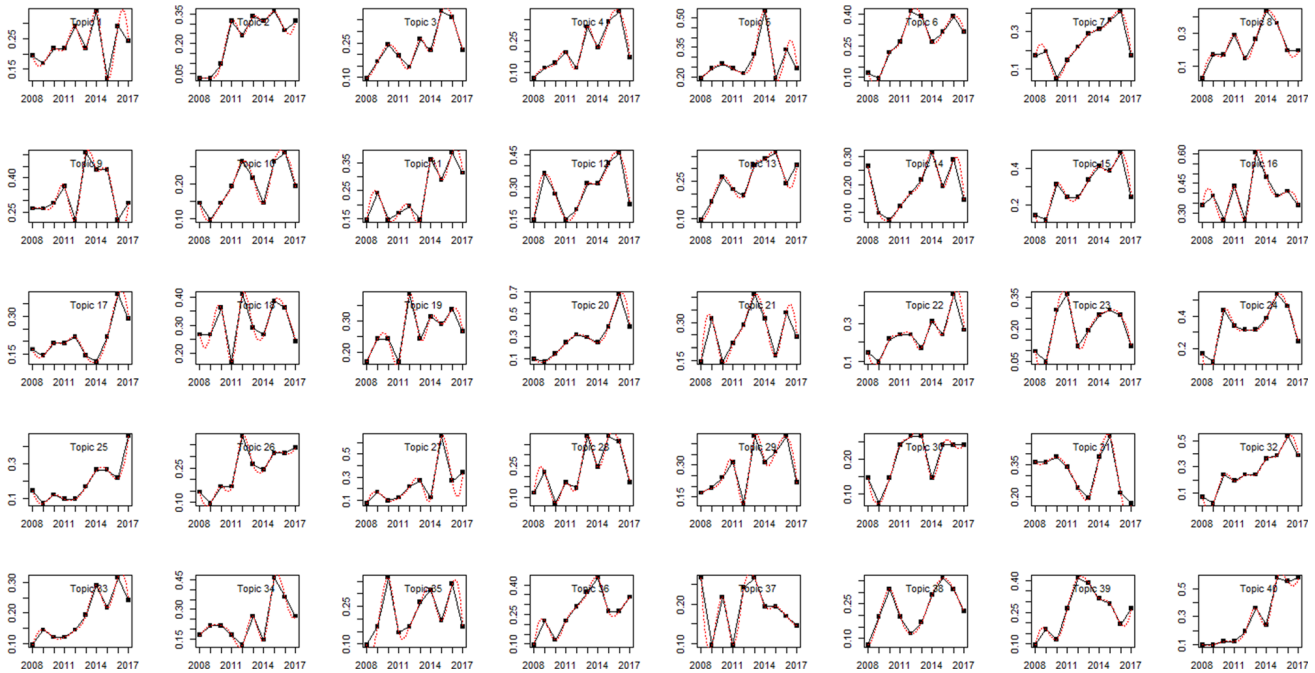

Fig. 11 Trends of the 40 research topics during the year 2008–2017 (*x*-coordinate as year, *y*-coordinate as proportion %)

than other countries. The majority of productive institutes and authors come from the USA. Collaboration degree analysis reveals that authors tend to collaborate more with those within the same institute or country.

A topic modeling-based bibliometric exploration regarding the global research trend of text mining in medical research field is also conducted. The 40-topic model has been successfully applied to discover the latent thematic patterns in the corpus. In light of our prior knowledge about text mining in medical research, most topics identified using LDA method are recognizable and easy-to-understand, as they are related to major issues in the research field. This topic modeling-based bibliometric exploration directly contributes to our understanding of what academic concerns of text mining in medical research field are in the past decade. We provide interpretations of the top 5 best matching topics as follows.

Topic 31 pertains to be *Speech related event mining* with the highest frequent term "Speech". Terms like "Prosodic", "Prosody", "Listener", "Consonant", "Vowel", "Sound", "Phonologicaland", and "Rhythm" are also included. Some researchers concern about the study of Aphasia, e.g., speech segmentation in Aphasia (Peñaloza et al. 2015); thus, terms such as "Aphasia" and "Aphasic" are also contained in Topic 31. Other study focuses include semantic processing in connected speech (Ahmed et al. 2013), automatic speech-recognition systems development for spoken clinical questions (Liu et al. 2011).

Topic 22 contains terms like "Men", "Sexual", "Hiv", "Sexuality", "Hiv/aids", "Gay", "Lesbian", and "Condom", and thus apparently refers to *Sexual related event mining*. Although improvements in the medical management of HIV have reduced the rate of perinatal transmission from mothers to their children, youth still continue to acquire HIV through risky behaviors such as unprotected sex and injection-drug use (Leonard et al. 2010). This attracts widespread attention from all circles of the society. Researchers in academia also concern much about sexual risk reduction through strengthening prevention efforts and clinical behavioral interventions.

Topic 5 centers around *Alzheimer event related mining*. Thus, terms like "Memory", "Schizophrenia", "Dementia", "Alzheimer", "Short-term", and "Spiritual" are contained in the topic. As one of the leading causes of death and one of the most financially costly diseases, Alzheimer has been always a worldwide concern. It is estimated that by 2050, one new case of Alzheimer's is expected to develop every 33 seconds, resulting in nearly 1 million new cases per year (Alzheimer's 2015). Many researchers devote themselves to Alzheimer's study using text mining techniques (e.g., Pistono et al. 2016; Oscar et al. 2017).

Topic 21 contains words like "Child", "Parent", "Mother", "Caregiver", "Parental", "Neonatal", "Infant", and "Parenting" and thus discusses *Parenting for child and infant*. Parenting or child rearing is the process of promoting and supporting the physical, emotional, social, and intellectual

development of a child from infancy to adulthood.[5] Relevant researches focus on events such as parenting stress (Kantrowitz-Gordon et al. 2016), and parenting and disability (Fraser and Llewellyn 2015).

Topic 16 focuses on *Nursing event mining* with terms like "Nursing", "Nurse", "Intimate", "Delirium", "Nurse-patient", "Abuse", and "Caregiver". Relevant studies include nursing education (Shin et al. 2015), nursing practices (Fey and Jenkins 2015), professionalism, and ethical dilemmas for nursing students (Rees et al. 2015; Kim et al. 2015), mental health nursing (Mårtensson et al. 2014), and the like.

The 40 identified topics are further clustered based on term-level similarity and document-level similarity to find latent relations and emerging interdisciplinary fields of these topics. As can be seen from Fig 9, Topic 18 and Topic 33 as well as Topic 21 and Topic 38 have high term-level similarity and are far distant from other topics. From the topic interpretations, Topic 18 and Topic 33 concern with reading and hearing issues, and both Topic 21 and Topic 38 contain "Child" as high frequent term. Other topics with less term-level similarity are mapped in the middle of the dendrogram. For instance, both Topic 5 and Topic 7 discuss about psychosis issues, and both Topic 1 and Topic 31 focus on speech issues. In summary, the dendrogram shown in Fig. 9 clearly presents the term usage similarity structure of the research topics.

Different from term-level similarity clustering, the goal of document-level similarity clustering is to describe the interaction structure of the research topics. As shown in Fig. 10, Topic 9 and Topic 16 have a high document-level similarity, meaning that publications with a high topic proportion of Topic 9 often have a high topic proportion of Topic 16 simultaneously. Document-level measure of topic similarity has the same meaning of interdisciplinary analysis (Lu and Wolfram 2012). If two topics frequently appear in the same publications, there is a big potential to foster a novel interdisciplinary research field. Almost all topic pairs have high term-level similarity but low document-level similarity, such as Topic 18 and Topic 33, or high document-level similarity but low term-level similarity, such as Topic 9 and Topic 16. The differences between term-level similarity and document-level similarity also reflect the intellectual structure of the research field in the past decade.

Increasing and decreasing topics are also recognized through statistic test. We provide brief explanations for some of the emerging topics. Topic 2 discusses *Organ transplantation*; Topic 13 focuses on the *Labour related event*; Topic 15 addresses *Drug regulatory related event*; Topic 22 focuses on *Sexual event mining*; Topic 25 addresses *Smoking event*; Topic 26 is about *Aging event*; Topic 27 centers around *Heart disease*; and Topic 40 relates with *Depression event*. As can be seen from Fig. 11, some topics, such as Topic 1, Topic 5, and Topic 9, show a trend with relatively sharp fluctuations. Topic 8 and Topic 14 show an increasing trend before 2014 and a decreasing trend after 2014.

We highlight this study at its improvements comparing with the existing similar works with the adoption of bibliometrics. According to our investigation, some deficiencies of the existing bibliometric works are found as follows. First, most relevant studies used either WoS or PubMed as the publication retrieval database for studying medical-related topics (e.g., Khan et al. 2017; Nafade et al. 2018; Baker et al. 2018). However, the difference in database coverage might lead to insufficiency of analyzing results when only one of them was used. Second, the existing bibliometric studies focusing on theme discovery seldom included terms in title and abstract fields as the analysis elements, which might lead to insufficient analysis. Last but not the least, although in a few studies such as Yeung et al. (2017), key terms in title and abstract fields were included for analysis, but with equal importance. However, it is more reasonable to bestow weighing for terms from different fields. Therefore, giving the deficiencies in the existing researches, this study uses both WoS and PubMed as the publication resource databases. We not only include key terms extracted from free text by using a self-developed NLP module, but assign weights based on experiment to terms from different fields. We also employ various analyzing techniques such as geographic visualization, collaboration degree, social network analysis, and topic modeling analysis for a more comprehensive analysis.

There are some limitations in this study. First, we treat journal and conference publications equally important in the analysis. Generally, the quality of a journal publication is higher than a conference publication. Therefore, in the future, we will seek persuasive way to bestow weighing for publications of different types. Second, citation data available from WoS have not been employed in the analysis since PubMed does not provide citation data as WoS. Citation data are indeed valuable to describe relations between scientific publications. Thus, further investigation is required to take citation data into consideration, with an in-depth understanding of the citing rationale. Last but not least, as for topic cluster analysis, the clustering is based on cosine similarity, and the clustering results might be vulnerable to choices of similarity measurement method. Therefore, in our future work, we will conduct comparison on different calculation methods for further exploration.

Notwithstanding its limitations, this study is the first to thoroughly assess research output of text mining in medical research field in statistical perspective. The findings in the study can potentially benefit relevant researchers, especially newcomers in understanding the research performance and recent development of the research field, optimizing research topic decision, and monitoring new scientific or technological activities.

[5] https://en.wikipedia.org/wiki/Parenting.

5 Conclusions

This study presents a bibliometric analysis of the text mining in medical research area during the year 2008–2017. Our work is the first in-depth study on keeping track of the current advances in the research area from quantitative perspective. The result shows that the developed methods are universal and can help researchers comprehensively understand the knowledge of a certain field hidden in a large amount of scientific literature. The rapid growth of scientific literature reveals the vigorous development of text mining in medical research in recent years. Collaboration degree analysis and social network analysis reveal scientific collaboration characteristics. Latent Dirichlet allocation exploration presents a comprehensive overview and an intellectual structure of the research, especially research topics. The clustering analysis and trend analysis can help process the derived topics to provide an architecture overview of a certain field in more detail.

For further studies, we will employ the author-topic model, a probabilistic model for linking authors to observed words in the scientific literature of the research field. This will provide a general framework for exploration, discovery, and query-answering in the context of the relations of author and topics.

Acknowledgements The work was funded by the grant from National Natural Science Foundation of China (No. 61772146) and Guangzhou Science Technology and Innovation Commission (No. 201803010063).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix

See Table 8.

Table 8 The list of keywords related to the "text mining" determined by relevant domain experts in the field

"Natural language processing" OR "semantic analysis" OR "word sense disambiguation" OR "named entity recognition" OR "sentiment analysis" OR "information extraction" OR "syntactic analysis" OR "dependency parsing" OR "syntactic parsing" OR "morphological segmentation" OR "part-of-speech tagging" OR "sentence boundary disambiguation" OR "word segmentation" OR "terminology extraction" OR "machine translation" OR "natural language generation" OR "natural language understanding" OR "question answering" OR "recognizing textual entailment" OR "relationship extraction" OR "topic segmentation" OR "topic recognition" OR "natural-language generation" OR "natural-language understanding" OR "natural-language processing" OR "automatic summarization" OR "discourse analysis" OR "text mining" OR "information retrieval" OR "named entity recognition" OR "text categorization" OR "text clustering" OR "entity extraction" OR "concept extraction" OR "sentiment analysis" OR "document summarization" OR "entity relation modeling"

References

Ahmed S, de Jager CA, Haigh AM, Garrard P (2013) Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. Neuropsychology 27(1):79

Alzheimer's A (2015) 2015 Alzheimer's disease facts and figures. Alzheimers Dement 11(3):332

Apte C, Damerau F, Weiss SM, Apte C, Damerau F, Weiss S (1998) Text mining with decision trees and decision rules. In: Proceedings of the conference on automated learning and discorery, Workshop 6: learning from text and the web, Citeseer

Baker NC, Ekins S, Williams AJ, Tropsha A (2018) A bibliometric review of drug repurposing. Drug Discov Today 23:661–672

Batet M, Sánchez D, Valls A (2011) An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 44(1):118–125

Blei DM, Lafferty JD (2007) A correlated topic model of science. Ann Appl Stat 1:17–35

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Bouyssou D, Marchant T (2011) Ranking scientists and departments in a consistent manner. J Am Soc Inf Sci Technol 62(9):1761–1769

Chen X, Chen B, Zhang C, Hao T (2017a) Discovering the recent research in natural language processing field based on a statistical approach. In: International symposium on emerging technologies for education, Springer, pp 507–517

Chen X, Weng H, Hao T (2017b) A data-driven approach for discovering the recent research status of diabetes in China. In: International conference on health information science, Springer, pp 89–101

Chen X, Ding R, Xu K, Wang S, Hao T, Zhou Y (2018a) A bibliometric review of natural language processing empowered mobile computing. Wirel Commun Mob Comput. https://doi.org/10.1155/2018/1827074

Chen X, Xie H, Wang FL, Liu Z, Xu J, Hao T (2018b) A bibliometric analysis of natural language processing in medical research. BMC Med Inform Decis 18(1):14

Chou WyS, Prestin A, Kunath S (2014) Obesity in social media: a mixed methods analysis. Transl Behav Med 4(3):314–323

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

Fey MK, Jenkins LS (2015) Debriefing practices in nursing education programs: Results from a national study. Nurs Educ Perspect 36(6):361–366

Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics, Association for Computational Linguistics, pp 363–370

Fraser V, Llewellyn G (2015) Good, bad or absent: discourses of parents with disabilities in Australian news media. J Appl Res Intellect Disabil 28(4):319–329

Fu Hz, Ho Ys, Sui Ym, Li Zs (2010) A bibliometric analysis of solid waste research during the period 1993–2008. Waste Manag 30(12):2410–2417
Glanzel W (2003) Bibliometrics as a research field: a course on theory and application of bibliometric indicators. Course Handouts. http://www.norslis.net/2004/ib_Module_KUL.pdf
Grün B, Hornik K (2011) Topicmodels: an R package for fitting topic models. J Stat Softw 40(13):1–30
He P, Deng Z, Wang H, Liu Z (2016) Model approach to grammatical evolution: theory and case study. Soft Comput 20(9):3537–3548
He P, Deng Z, Gao C, Wang X, Li J (2017) Model approach to grammatical evolution: deep-structured analyzing of model and representation. Soft Comput 21(18):5413–5423
Hearst M (2003) What is text mining. SIMS, UC, Berkeley
Hoek J, Gifford H, Maubach N, Newcombe R (2014) A qualitative analysis of messages to promote smoking cessation among pregnant women. BMJ Open 4(11):e006716
Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., pp 289–296
Jiang H, Qiang M, Lin P (2016) A topic modeling based bibliometric exploration of hydropower research. Renew Sustain Energ Rev 57:226–237
Jin J, Yan X, Li Y, Li Y (2016) How users adopt healthcare information: an empirical study of an online Q&A community. Int J Med Inform 86:91–103
Kantrowitz-Gordon I, Altman MR, Vandermause R (2016) Prolonged distress of parents after early preterm birth. Jognn J Obst Gyn Neo 45(2):196–209
Khan MS, Ullah W, Riaz IB, Bhulani N, Manning WJ, Tridandapani S, Khosa F (2017) Top 100 cited articles in cardiovascular magnetic resonance: a bibliometric analysis. J Cardiovasc Magn Reson 18(1):87
Kim K, Han Y, Js Kim (2015) Korean nurses ethical dilemmas, professional values and professional quality of life. Nurs Ethics 22(4):467–478
Knight R, Shoveller JA, Oliffe JL, Gilbert M, Frank B, Ogilvie G (2012) Masculinities, guy talkand manning up: a discourse analysis of how young men talk about sexual health. Sociol Health Ill 34(8):1246–1261
Leonard AD, Markham CM, Bui T, Shegog R, Paul ME (2010) Lowering the risk of secondary HIV transmission: insights from HIV-positive youth and health care providers. Perspect Sex Reprod Health 42(2):110–116
Li W, Zhao Y (2015) Bibliometric analysis of global environmental assessment research in a 20-year period. Environ Impact Assess Rev 50:158–166
Li Z, Nie F, Chang X, Yang Y (2017) Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis. IEEE Trans Knowl Data Eng 29(10):2100–2110
Lin W, Xu S, He L, Li J (2017) Multi-resource scheduling and power simulation for cloud computing. Inform Sci 397:168–186
Liu F, Tur G, Hakkani-Tür D, Yu H (2011) Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. J Am Med Inform Assoc 18(5):625–630
Lu K, Wolfram D (2012) Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches. J Assoc Inf Sci Technol 63(10):1973–1986
Lucini FR, Fogliatto FS, da Silveira GJ, Neyeloff JL, Anzanello MJ, Kuchenbecker RdS, Schaan BD (2017) Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform 100:1–8
Luo M, Chang X, Li Z, Nie L, Hauptmann AG, Zheng Q (2017) Simple to complex cross-modal learning to rank. Comput Vis Image Underst 163:67–77
Mann HB (1945) Nonparametric tests against trend. Econometrica 13:245–259
Mårtensson G, Jacobsson JW, Engström M (2014) Mental health nursing staff's attitudes towards mental illness: an analysis of related factors. J Psychiatr Ment Health Nurs 21(9):782–788
Mazloumian A (2012) Predicting scholars' scientific impact. Plos One 7(11):e49246
Merigó JM, Gil-Lafuente AM, Yager RR (2015) An overview of fuzzy research with bibliometric indicators. Appl Soft Comput 27:420–433
Meystre S, Haug PJ (2005) Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 5(1):30
Nafade V, Nash M, Huddart S, Pande T, Gebreselassie N, Lienhardt C, Pai M (2018) A bibliometric analysis of tuberculosis research, 2007–2016. Plos One 13(6):e0199706
Nichols LG (2014) A topic model approach to measuring interdisciplinarity at the national science foundation. Scientometrics 100(3):741–754
Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K (2017) Machine learning, sentiment analysis, and tweets: an examination of Alzheimers disease stigma on twitter. J Gerontol B Psychol 72(5):742–751
Peñaloza C, Benetello A, Tuomiranta L, Heikius IM, Järvinen S, Majos MC, Cardona P, Juncadella M, Laine M, Martin N et al (2015) Speech segmentation in aphasia. Aphasiology 29(6):724–743
Pistono A, Jucla M, Barbeau EJ, Saint-Aubert L, Lemesle B, Calvet B, Köpke B, Puel M, Pariente J (2016) Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimers disease. J Alzheimers Dis 50(3):687–698
Rees CE, Monrouxe LV, McDonald LA (2015) My mentor kicked a dying woman's bedanalysing UK nursing studentsmost memorable professionalism dilemmas. J Adv Nurs 71(1):169–180
Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 60(5):503–520
Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. Expert Syst Appl 33(1):135–146
Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
Saraswathi K, Tamilarasi A (2016) Ant colony optimization based feature selection for opinion mining classification. J Med Imaging Health Inform 6(7):1594–1599
Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
Shin S, Park JH, Kim JH (2015) Effectiveness of patient simulation in nursing education: meta-analysis. Nurse Educ Today 35(1):176–182
Tan H, Gao Y (2017) Patch-based principal covariance discriminative learning for image set classification. IEEE Access 5:15001–15012
Tan H, Gao Y, Ma Z (2018) Regularized constraint subspace based method for image set classification. Pattern Recogn 76:434–448
Teh YW, Jordan MI, Beal MJ, Blei DM (2005) Sharing clusters among related groups: hierarchical Dirichlet processes. In: Advances in neural information processing systems, pp 1385–1392
Wang H, Wang W, Cui Z, Zhou X, Zhao J, Li Y (2018) A new dynamic firefly algorithm for demand estimation of water resources. Inform Sci 438:95–106
Wei Y, Mi Z, Zhang H (2013) Progress of integrated assessment models for climate policy. Syst Eng Theory Pract 33(8):1905–1915
Yau CK, Porter A, Newman N, Suominen A (2014) Clustering scientific documents with topic modeling. Scientometrics 100(3):767–786
Yeung AWK, Goto TK, Leung WK (2017) The changing landscape of neuroscience research, 2006–2015: a bibliometric study. Front Neurosci Switz 11:120

Yu D, Xu Z, Wang W (2018) Bibliometric analysis of fuzzy theory research in china: a 30-year perspective. Knowl Based Syst 141:188–199

Zhang K, Wang Q, Liang QM, Chen H (2016) A bibliometric analysis of research on carbon tax from 1989 to 2014. Renew Sustain Energy Rev 58:297–310

Zhang S, Yang Z, Xing X, Gao Y, Xie D, Wong HS (2017) Generalized pair-counting similarity measures for clustering and cluster ensembles. IEEE Access 5:16904–16918

Zhong S, Geng Y, Liu W, Gao C, Chen W (2016) A bibliometric review on natural resource accounting during 1995–2014. J Clean Prod 139:122–132