Scientometrics (2021) 126:2971–2989
https://doi.org/10.1007/s11192-020-03858-y

The use of citation context to detect the evolution of research topics: a large-scale analysis

Chaker Jebari[1] · **Enrique Herrera-Viedma**[2] · **Manuel Jesus Cobo**[3]

Received: 28 May 2020 / Accepted: 27 December 2020 / Published online: 5 February 2021

Abstract
With the exponential increase in the number of published papers, discovering how topics evolve becomes increasingly important for anybody involved in research, including researchers, institutes, research funding bodies, and decision-makers. This study proposes a large-scale analysis of the evolution of biomedical and life sciences using the citation contexts of the collected papers, or more precisely their citing sentences. Using 64,350 papers published in PubMed Central between 2008 and 2018, we determined the research trends for ten research topics. Moreover, we studied how these topics evolve across countries and across the most common journals in biomedical and life sciences.

Keywords Citation context · Research trends · Topic modeling · Biomedical and life sciences

Introduction

We are currently witnessing a rapid growth in the number of research publications produced each year in different disciplines (Larsen and von Ins 2010). These publications are stored in predefined scientific databases that contain millions of research papers, such as Web of Science, Scopus, Google Scholar, Dimensions, etc. It has been estimated that the amount of scientific literature increases by 8–9% annually (Bornmann and Mutz 2015). This leads to an evolving research environment in which new research topics emerge regularly while others become obsolete. Hence, detecting the evolution of research topics becomes a very important task for anybody involved in research, such as researchers, institutes, research funding bodies, and decision-makers. For instance, academic researchers should know the evolution of topics within journals in order to target the most suitable one for their publications. They also need to know the evolution of research topics

✉ Chaker Jebari
jebarichaker@yahoo.fr

[1] Department of Information Technology, University of Technology and Applied Sciences, Ibri, Sultanate of Oman

[2] Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada, Spain

[3] Department of Computer Science and Engineering, University of Cádiz, Cádiz, Spain

in different countries to establish collaborations with other researchers. Research funding bodies should know the most evolving and promising topics for which they can provide research funds.

To detect the evolution of research topics, many works have proposed to study particular fields such as biomedical informatics (Kim et al. 2011), marketing (Murgado Armenteros et al. 2015), information retrieval (Chen et al. 2017), fuzzy theory (Yu et al. 2018), agricultural science (Sagar et al. 2013), Internet of Things (MacDonald and Dressler 2018), transportation (Cobo et al. 2014; Sun and Yin 2017), etc., while others have conducted their analysis by focusing on one journal such as *Knowledge-Based Systems* (Cobo et al. 2015; Zhang et al. 2017). It is worth noting that these studies are indeed very helpful to detect research trends within those particular fields and journals.

To detect research trends, bibliometric methods have also been used to evaluate scientific manuscripts (Zitt et al. 2005; Li et al. 2009). Thus, by modeling a set of scientific manuscripts as a complex system, a bibliographic network can be built to understand the social, conceptual, or intellectual aspects of a field (Kim et al. 2011; Murgado Armenteros et al. 2015; Chen et al. 2017; Yu et al. 2018; Sagar et al. 2013; MacDonald and Dressler 2018; Cobo et al. 2014; Sun and Yin 2017).

It has been shown that the citation context provides valuable data about its content (Zhang et al. 2013), and it has thus been considered as a fundamental step in many natural language processing applications such as publication summarization (Qazvinian and Radev 2010), survey article generation (Mohammad et al. 2009), information retrieval (Ritchie 2009), sentiment analysis (Athar 2014), author co-citation analysis (Bu et al. 2018), etc.

In this paper, we propose to use the citation context to study the evolution of the ten most common topics in biomedical and life sciences using dynamic topic modeling. A large-scale analysis is carried out using a large collection of papers gathered from the open repository PMC[1]. We choose to study the biomedical and life sciences because they are witnessing an explosion of information and applications in various disciplines of science, engineering, and medicine. This revolution in biomedical and life sciences has impacted academicians, scientists, and industrialists.

Using a collection of 64,350 papers published in different journals, our analysis reveals the evolution of the ten most important topics within a period of 11 years (from 2008 to 2018). The results show that some topics are evolving, while others are receding. In addition, we analyzed how these topics evolved in 16 journals and over 12 countries as well.

The remainder of this paper is organized as follows: Sect. 2 presents the preliminaries of our study. This section is divided into two parts: the first part presents previous studies to identify research trends, while the second part explains citation context in detail. Section 3 explains the methodology used, then Sect. 4 presents and discusses the obtained results.

Preliminaries

Related studies

Previous studies have proposed many methods to analyze the evolution of research topics and thereby detect emerging research topics, including both qualitative and

[1] https://www.ncbi.nlm.nih.gov/pmc/

quantitative measures. These methods are mainly clustering based and can be classified into three categories: (i) citation-based methods, (ii) text-based methods, and (iii) hybrid methods.

In the case of citation-based methods, different bibliometric traits have been used to identify research trends, such as the annual growth in the number of publications (Small et al. 2014; Bengisu 2003) and the relationship between publications, such as direct citation relations [e.g., Shibata et al. (2011); Kajikawa and Takeda (2008)], bibliographic coupling relations [(e.g., Morris et al. (2003)], and co-citation relations [e.g., Upham and Small (2010)]. In particular, co-citation relationships connecting different papers are commonly used to create a network of related papers which can be used to create clusters of similar papers (Small 1973; Kajikawa and Takeda 2008; Chen et al. 2012). These clusters are used to detect emerging research topics.

In the case of text-based methods, text mining techniques are used to identify the invisible connections between publications and thus enhance the detection of research topics. Text mining techniques have been applied to detect emerging technologies in a particular field using the frequencies of terms and phrases (Smalheiser 2001). To consider the semantics of these terms and phrases, more sophisticated techniques have been used, such as latent semantic analysis (LSA) (Gordon and Dumais 1998) and latent Dirichlet analysis (LDA) (Grifiths and Steyvers 2004).

Another text-mining technique, namely co-word analysis, is based on the assumption that keywords represent the best description of the paper content (Callon et al. 1983). This technique uses the keywords shared by publications and their interactions across publications to represent the structure of a research field and to map research topics (Hui and Fong 2004; Wang et al. 2012; Chen et al. 2016). Co-word analysis has been used by many researchers to analyze and map research fields. Dehdarirad et al. Dehdarirad et al. (2014) proposed a co-word method based on keywords from funded projects to map research trends. Hu et al. Hu et al. (2013) used 959 full-text articles downloaded from the *Scientometrics* journal to map the intellectual structure of the scientometrics field using text-mining and co-word analysis.

Co-word analysis has been used in many specific research fields such as library and information science (Hu et al. 2013), a recommendation system in China (Hu and Zhang 2015), robot technology in Korea (Lee and Jeong 2008), consumer behavior (Muñoz-Leiva et al. 2012), intelligence models in management and business (López-Robles et al. 2019), rheumatology (Perez-Cabezas et al. 2018), complementary and alternative medicine (Moral-Munoz et al. 2018), etc.

In the case of hybrid methods, bibliometric and text-based techniques are used (Kostoff 2001). In hybrid analysis, citation-based methods are applied to reduce the number of papers before applying text-mining techniques (Kostoff et al. 2001; Chen 2006). Glänzel and Thijs (Glänzel and Thijs 2012) combined the co-citation and term-based techniques to analyze research trends in regenerative medicine. Some studies have analyzed the "core document" in each cluster to identify emerging topics (e.g., Reiss et al. (2013); Small (2006); Ohniwa et al. (2010)), while others have used the frequency of keywords as an indicator of emerging topics (e.g., Guo et al. (2011)). Recently, Small et al. Small et al. (2017) proposed a method to identify discoveries in biomedical science. The proposed method uses the citation context information, more precisely citing sentences, drawn from the PubMed Central database. The proposed method focuses on the use of specific terms in the citing sentences and the joint appearance of cited references.

Citation context

Since the seminal work published by Garfield Garfield (1963), citation context analysis has attracted increasing attention from many researchers in the bibliometric field (Alvarez and Gómez 2016). A citation context is the text surrounding a reference marker that cites other papers. It has been stated by Schwartz and Hearst Schwartz and Hearst (2006) that the citation context can better describe the main points of an article than does its abstract. Therefore, it can be considered as a good summary of the main contributions of the cited papers.

A citation context is formed by one or more consecutive sentences that constitute the main ideas of the cited papers (Small 2011). A citation sentence can be classified as either explicit or implicit. An explicit citation sentence is a sentence that contains one or more citation references (Athar and Teufel 2012). An implicit citation sentence appears next to an explicit citation sentence and does not include any citation reference but supplies additional information on the content of the cited paper (Abu-Jbara and Radev 2012). The following example illustrates these two types of citation sentence, where the sentence in bold is an explicit citation and the italic sentence is an implicit citation:

> ...**In order to improve sentence-level evaluation performance, several metrics have been proposed, including** *ROUGE − W* **and** *METEOR* **[4]**. *METEOR is essentially a unigram based metric, which prefers the monotonic word alignment between MT output and the references by penalizing crossing word alignments....*

It is worth noting that extracting implicit citation sentences is not trivial, thus all previous works on citation context extraction either explored explicit citation sentences only (Qazvinian and Radev 2010) or used a fixed-size text window to recognize citation context (Athar 2011). It has been stated by Zhang et al. Zhang et al. (2013) that the fixed-size window method can efficiently detect implicit citation sentences but creates a lot of noise. Recently, a new approach proposed by Jebari et al. Jebari et al. (2018) to extract implicit citations from scientifi papers. The new approach uses topic modeling to identify the topic of the cited paper and word embeddings to represent the implicit citing sentences.

Recently, citation context has been used as a source of information in many natural language processing applications such as literature retrieval (Liu et al. 2014), citation sentiment (Yan et al. 2020), citation recommendation (Ma et al. 2020), etc.

As stated by He and Chen He and Chen (2018), citation context provides information on how different aspects of past scientific papers are described by researchers in their publications. These descriptions are different and may change as science evolves over the years. For this reason, we hypothesize that citation contexts can provide relevant words that can help to detect research trends.

In this study, we use explicit citing sentences to study how research topics evolve over time and across different journals and countries. To the best of our knowledge, we are the first to use citation contexts to analyze the evolution of biomedical and life sciences.

Methodology

To detect research trends in biomedical and life sciences, a bibliometric analysis is carried out. Using citation contexts extracted from the body of scientific publications, our methodology can detect the evolution of research topics using dynamic topic modeling. As shown

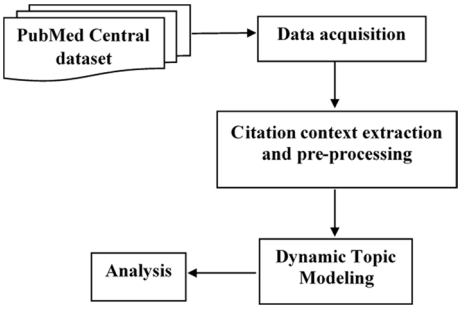


Fig. 1 Methodology workflow

in Fig. 1, our methodology consists of four tasks: (1) data acquisition, (2) citation context extraction and preprocessing, (3) dynamic topic modeling, and (4) analysis.

Data acquisition

To collect our data, we used the PubMed Central Open Access Subset (PMC OAS) dataset, an open-access XML-formatted full-text document repository of biomedicine and life sciences maintained by the US National Library of Medicine (US NLM). Each paper has a PMID index (called also PubMed reference number) assigned by the National Institutes of Health to papers indexed in the PubMed dataset. In PubMed dataset, we can download papers in XML format, which can facilitates the processing of their contents. In this study, we collected only papers published from 2008 to 2018 and we ignored papers that did not use PMID indices to cite other papers.

Citation context extraction and preprocessing

To extract the citation sentences that cite a given paper, we replaced the reference enclosed between <xref ref-type="bibr"></xref> with its PMID. To ensure that all papers are cited by other papers published within the same period (from 2008 to 2018), we decided to remove papers that are not cited at all by papers from 2008 to 2018. After that, for each paper, we extracted the citation sentences where the paper is cited. Table 1 presents the different citations of the paper with PMID=18988837 by four other papers published in different years.

Using Python natural language toolkit NLTK[2], we split the citation sentences into words and removed stop-word characters. To retain only nouns, we labeled each remaining word with its Part-Of-Speech tag. Besides the citation sentences, we extracted for each paper the publication year, the name of the journal in which it is published, and the affiliations of the authors. To detect the affiliation country from the university name, we used geopy[3] and geotex[4] Python packages.

[2] https://www.nltk.org/
[3] https://pypi.org/project/geopy/
[4] https://pypi.org/project/geotext/

Table 1 Citation context example

PMID	Year	Citation sentence
19683024	2010	It is well known that linkage analysis is powerful in detecting rare and high risk alleles but has limited power in identifying common genetic variants with low-penetrance [<xref rid="R29">15516958</xref>,<xref rid="R30">18988837</xref>].
19818800	2011	Faster and more efficient genotyping methods have propelled studies that seek to identify QTL in many different systems, including humans (<xref rid="R1">18988837</xref>)
20639796	2014	A key assumption in GWAS is what is known as the common disease/common variant hypothesis [<xref rid="R18">18988837</xref>]
20552648	2016	Determining the genetic basis of complex genetic diseases is one of the main challenges in human genetics [<xref rid="R2">18988837</xref>

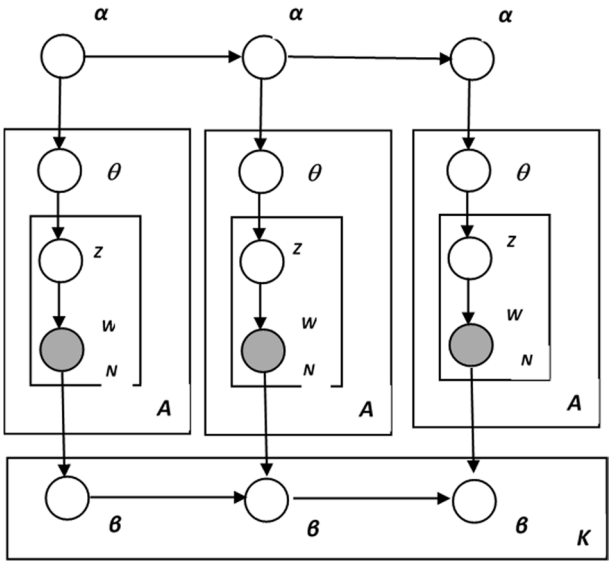


Fig. 2 A graphical representation of DTM

Table 2 DTM notation

Notation	Description
D	A corpus of M documents
T	A set of time slices
D_t	The set of documents belonging to time slice t
α_t	The per-document topic distribution at time slice t
$\beta_{t,k}$	The word distribution of topic k at time slice t
$\eta_{t,d}$	The topic distribution for document d in time slice t
$Z_{t,d,n}$	The topic for word n in document d in time slice t
$W_{t,d,n}$	A specific word

Dynamic topic modeling

To discover the evolution of research topics over time, Dynamic Topic Modeling (DTM) algorithm is used. DTM is a statistical generative model used to analyze the evolution of latent topics over time slices. It is an extension to the Latent Dirichlet Allocation (LDA) proposed by David Blei and John Lafferty Blei and Lafferty (2006). In contrast to LDA, where the distribution of topics is time independent, DTM assumes that the topics of documents from a time slice t are evolved from the topics of the previous time slice $t-1$. The graphical representation of DTM is illustrated in Fig. 2.

To explain DTM, we used the notations listed in the following table (Table 2):

The generative process at time slice t is explained as follows:

1. Draw the K topics $\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \rho^2 I)$
2. Draw mixture model $\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \rho^2 I)$
3. For each document:

3.1 Draw $\eta_{t,d} \sim N(\alpha_{t,k}, \rho^2 I)$
3.2 For each word:

3.2.1 Draw topic $Z_{t,d,n} \sim Mult(\pi(\eta_{t,d}))$
3.2.2 Draw topic $Z_{t,d,n} \sim Mult(\pi(W_{t,d,n}))$

Where $\pi(x)$ is a mapping from the natural parameterization x to the mean parameterization, calculated as follows:

$$\pi(x_i) = \frac{exp(x_i)}{\sum_i exp(x_i)} \tag{1}$$

To infer the model parameters (α_t, $\beta_{t,k}$, $\eta_{t,d}$, and $Z_{t,d,n}$), variational Kalman filtering and variational wavelet regression (Blei and Lafferty 2006) are used.

To apply the DTM algorithm, we grouped our papers into 11 time slices (from 2008 to 2018) according to their publication year. In this work, we used the Python Gensim implementation of DTM[5] with LDA initialization. The DTM algorithm requires the number of topics K as an input, which is set to 10 in this study. The hyperparameter α_i controls the sparsity of the document topics for the LDA models in each time slice and is set to 0.01. The Gaussian parameter defined in the beta distribution to dictate how the beta values evolve over time is set to 0.005.

Analysis

The analysis is carried out using 10 topics identified in the period from 2008 to 2018. The evolution of the topics is studied across publication year, journal name, and the author's country. To do so, we used the distribution $\eta^k_{t,d}$ of the topic k in the document d and at time slice t. Based on this distribution, we proposed the following three measures:

The *topic distribution over time*, $\Theta_t(k)$, represents the proportion of topic k at time slice t and is defined as follows:

$$\Theta_t(k) = \frac{\sum_{d \in D_t} \eta^k_{t,d}}{|D_t|} \tag{2}$$

The *journal topic distribution over time*, $\Theta^j_t(k)$, represents the proportion of topic k in journal j at time slice t and is defined as follows:

$$\Theta^j_t(k) = \frac{\sum_{d \in D_t} \eta^k_{t,d} \times h^j(d)}{\sum_{d \in D_t} h^j(d)} \tag{3}$$

where $h^j(d) = 1$ if paper d is published by journal j, and 0 otherwise.

The *country topic distribution over time*, $\Theta^c_t(k)$, represents the proportion of topic k at time slice t published by authors from country c and is defined as follows:

[5] https://radimrehurek.com/gensim/models/dtmmodel.html.

Table 3 Used dataset

Year	#papers	Year	#papers
2008	702	2014	5404
2009	11,533	2015	1557
2010	14,649	2016	875
2011	11,634	2017	1813
2012	8087	2018	665
2013	7431		

$$\Theta_t^c(k) = \frac{\sum_{d \in D_t} n_{t,d}^k \times h^c(d)}{\sum_{d \in D_t} h^c(d)} \tag{4}$$

Where $h^c(d) = 1$ if the country of one of the authors is c, and 0 otherwise.

Results and discussion

This section firstly describes the used dataset. Secondly, it presents the top ten discovered topics and discusses their evolution. Thirdly, it discusses the evolution of these ten topics across journals and countries.

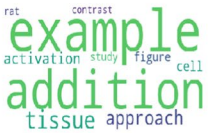
Dataset

As explained in "Methodology" section, the dataset used in this paper is collected from the PubMed Central (PMC) Open Access Subset (OAS), which is an open-access XML-formatted full-text document repository of the biomedicine and life sciences maintained by the US NLM. After the acquisition step explained in "Methodology" section, we created a dataset comprising 64,350 papers grouped by publication year (See Table 3). These papers are published by 10,171 journals, by authors from 123 countries. After the preprocessing step explained in "Methodology" section, we obtained a total of 59,244 terms.

Discovered topics

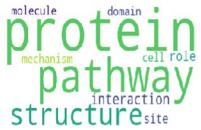
After running the DTM model with 10 topics and 11 time slices (from 2008 to 2018), we obtained the word distribution for each topic. Using the 50 top terms in each topic, we presented these distributions as a wordcloud, where the size of the word is proportional to its probability (See Fig. 3). To assign labels to the topics, we used Medical Subject Headings (MeSH), a comprehensive controlled vocabulary created and updated by the US NLM to facilitate searching. Most publications in MEDLINE/PubMed are manually assigned a set of descriptors from MeSH by biomedical experts at the US NLM. The descriptors or subject headings are arranged in a hierarchy. When performing a MEDLINE search via PubMed, entry terms are automatically mapped to the corresponding descriptors. In this study, we mapped the seven top terms in each topic to the corresponding descriptor which is used as a topic label.

(a) Topic 1: mast cells

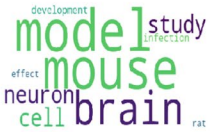


(b) Topic 2: risk factors

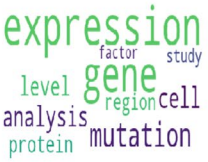
(c) Topic 3: protein structure, tertiary



(d) Topic 4: models, theoretical

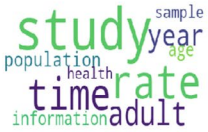


(e) Topic 5: gene expression profiling



(f) Topic 6: immunity

(g) Topic 7: time and motion studies



(h) Topic 8: behavioral research

(i) Topic 9: clinical trial

(j) Topic 10: receptors, vascular endothelial growth factor

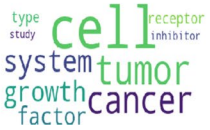


Fig. 3 Discovered topics

Topic trends

To analyze topic trends, we studied how topics evolve over years. To do so, we calculate for each topic the proportion of papers in each year using equation (2). The evolution of all the discovered topics is illustrated in Figs. 4. From Fig. 4, we observe that topic 1 (mast cells) and topic 5 (gene expression profiling) have the lowest proportions in all years. This is due to the generality of these topics, which can be cited in many biomedical and life science publications. We also observe that topic 3 (protein structure, tertiary) has the highest proportion in almost all the years. This can be explained by the fact that this topic is more specific and therefore it is published in more specific papers and journals. With respect

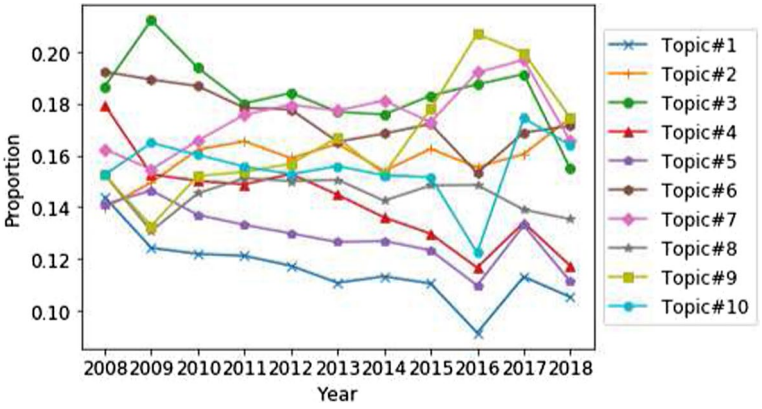


Fig. 4 The evolution of topics

to the topic trends, we observe that, from 2008 until 2012, topic 3 (protein structure, tertiary) and topic 6 (immunity) are dominant. From 2012, topic 7 (time and motion studies) started to be the most widely discussed topic along with topic 3 (protein structure, tertiary). From 2016 until 2018, topic 9 (clinical trial) become the most important topic, followed by topic 7 (time and motion studies) and topic 3 (protein structure, tertiary). It is worth noting that the proportion of all topics declines in 2018. This can be explained by the fact that the papers published in 2018 are relatively new and therefore have less number of citations.

Figure 5 shows the growth rate of each topic. From this figure, we can see that topics 7 (time and motion studies) ,9 (clinical trial), and 2 (risk factors) are growing quickly. For topics 7 and 9, the growth is explained by the fact that these topics are related to medical and life technologies that are continuously developing. The rise of topic 2 is due to the increase in many effects that increase the likelihood of developing a disease or injury. Some examples of the most important risk factors are low weight, unsafe sex, high blood pressure, tobacco and alcohol consumption, and unsafe water, sanitation, and hygiene. Topics 1 (mast cells),3 (protein structure, tertiary), 4 (models, theoretical), 5 (gene expression profiling), and 6 (immunity) are declining because they have already been studied in depth for many years. Topic 8 (behavioral research) and 10 (receptors, vascular endothelial growth factor) do not show clear trends because they are related to many variables; For example, the behavioral research topic explores cognitive processes within and between organisms, which mainly depends on observations.

Topic trends across journals

To study the evolution of topics across journals, we used the 16 most important journals having at least 1000 papers in biomedical and life sciences (See Table 4). To calculate the proportion of each topic in a given journal, we used equation (3). The evolution of the research topics across journals is illustrated in Fig. 6. From this Figure, we can observe that the topic proportions differ from one journal to another, as shown by the different colors.

To clearly show which topics are published by which journal, we summarize the results in Table 5. The symbol "$\surd$" indicates that the topic is covered in the journal, while the symbol "+" indicates that the topic is the most covered, i.e., has the highest value of $\Theta_t^j(k)$. From this table, we observe that topics 2, 3, 4, 5, and 6 are covered by all 16 journals,

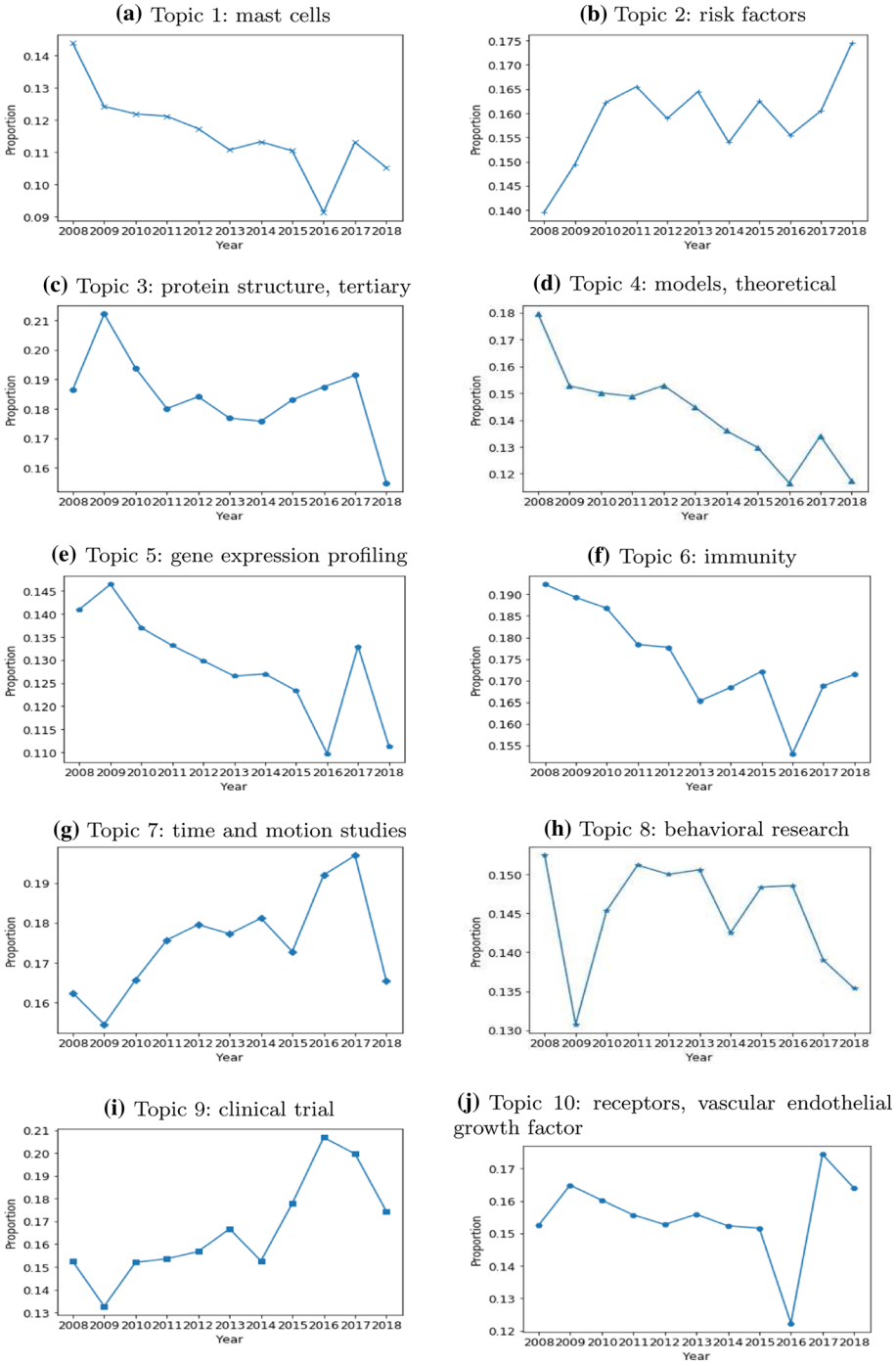


Fig. 5 Evolution of each topic

Table 4 Journals in the dataset with more than 1000 papers

Journal name	#papers
Journal of Immunology	3488
Journal of Neuroscience	3284
Journal of the American Chemical Society	3280
Biochemistry	2837
Cancer Research	2566
Nature	2049
Clinical Cancer Research	1759
Biochemical et Biophysical Acta	1655
Neuroimage	1603
Science	1450
Cell	1440
Neuron	1320
Journal of Molecular Biology	1279
Analytical Chemistry	1274
Cell Reports	1271
Neuroscience	1234
Methods in Molecular Biology	1130
Cancer	1124
Cancer Epidemiology, Biomarkers and Prevention	1112
Circulation Research	1105
Brain Research	1100
Biochemical and Biophysical Research Communications	1092
Oncogene	1070
Molecular Cell	1069
Circulation	1050
Organic Letters	1025
Developmental Biology	1019
Journal of Medicinal Chemistry	1009
Biomaterials	1008

while the rest of the topics are specific to some journals. This information is very useful for researchers who are searching for the most appropriate journal to publish their research papers.

Topic trends across countries

Similar to the analysis performed for the journals, we studied the evolution of topics across countries. To do so, we used only the countries in which at least 1000 scientists published a paper belonging to our dataset (See Table 6).

To detect the countries of the authors who participated in a given paper, we used the affiliation of the authors. To do so, we used the python GeoText[6] package to extract the

[6] https://pypi.org/project/geotext/

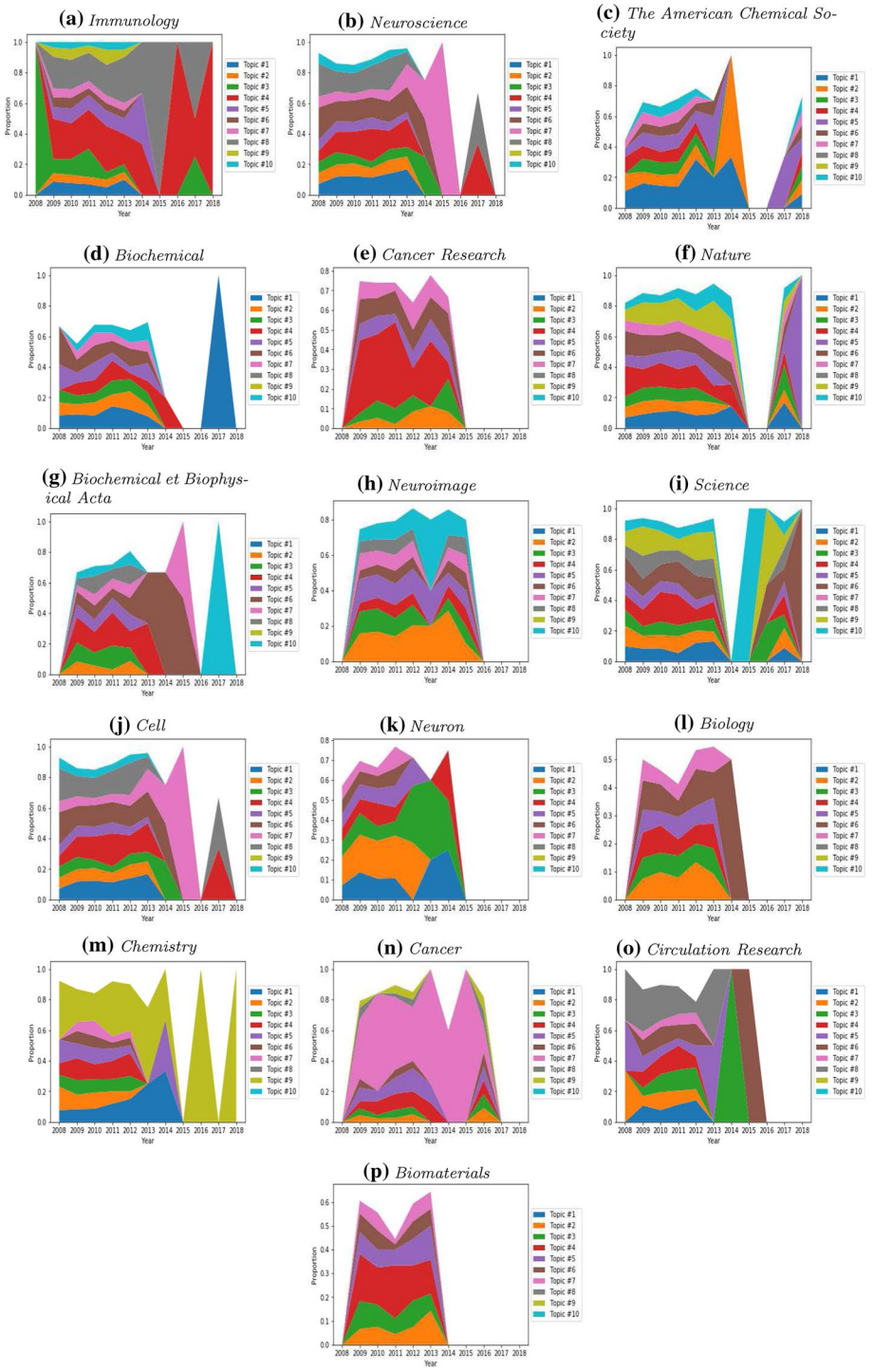


Fig. 6 Topic participation in each journal

Table 5 Topic participation in the 16 most important journals

Journal name	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Immunology	√	√	+	+	√	√	√	√	√	√
Neuroscience	√	√	√	√	√	√	√	√		√
The American Chemical Society	√	√	√	√	√	√	√			√
Biochemistry	+	√	√	√	√	√	√	√		√
Cancer Research		√	√	√	√	√	√	√		
Nature	√	√	√	√	√	√	√		√	√
Biochemical et Biophysical Acta		√	√	√	√	√	√	√		√
Neuroimage		+	√	√	+	√	√	√		√
Science	√	√	√	√	√	+		√	√	√
Cell	√	√	√	√	√	√	√	√		√
Neuron	√	√	√	√	√	√	√			
Biology		√	√	√	√	√	√			
Chemistry	√	√	√	√	√	√		+		
Cancer		√	√	√	√	√	+	√		
Circulation Research	√	√	√	√	√	√	√			
Biomaterials		√	√	√	√	√	√			

Table 6 Countries in the dataset with more than 1000 scientists

Country name	#papers
USA	108,334
Mali	5785
UK	2689
South Africa	2609
Germany	2222
Canada	1705
Brazil	1503
China	1485
Spain	1244
Japan	1239
Uruguay	1117
Italy	1024

country name from the author's affiliation. The evolution of topics in 12 countries is illustrated in Fig. 7. We can conclude that topic 7 is the most studied topic in all countries, followed by topic 10. For USA, all the topics are studied almost equally. In China, topic 5 is the most studied. In Mali, all topics are studied, with the most studied being topics 7, 9, and 10. This can be explained by the fact that Mali is witnessing many diseases (i.e., malaria, hepatitis, Ebola virus, etc.) and many other health and environmental problems. It is also clear that Uruguay focused more on topic 1, especially in 2015. We can notice also that UK, Japan, and Italy show similar topic trends. Germany, Spain, and Brazil also have similar topic trends. In 2015, we can observe that all countries witnessed a sharp decline,

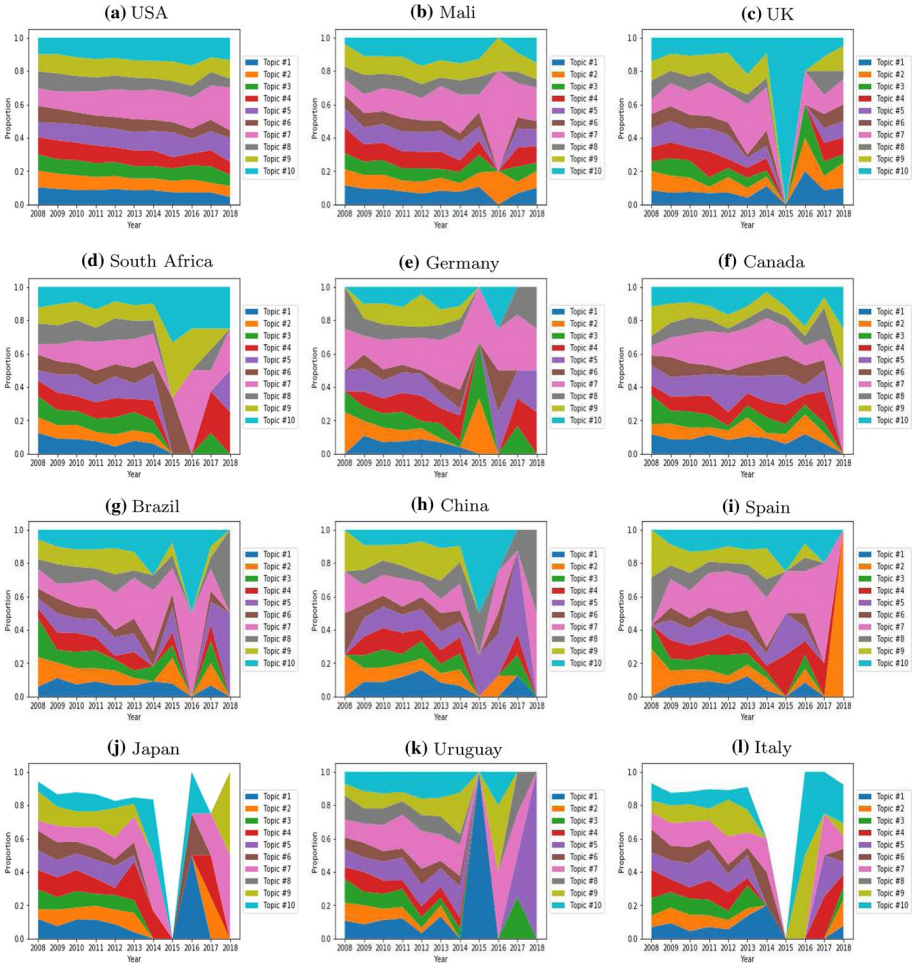


Fig. 7 Topic participation in each country

except Uruguay. This results can help to detect how a topic is evolving in a given country and thereby help research centers and academic institutes that wish to start research collaborations with a particular country in biomedical and life sciences.

Conclusion

In this paper, we presented an empirical work aimed at discovering research topics and their trends in biomedical and life sciences. We applied DTM on citation context extracted from papers published between 2008 to 2018. DTM can reveal the evolutionary nature of topics and does not require any prior knowledge. We discovered the top most studied topics and we studied their variation across years, journals, and countries.

By aggregating topics distributions at journal level, we found that most journals essentially cover different topics, while some focus more on a specific topic. This will help

researchers to target the most appropriate journal for their publications. By aggregating topics distributions at country level, we found that some countries pay more attention to some topics at the expense of others. This will reflect actual demand and what each country seeks through research. Based on this country-wise analysis, research funding agencies can evaluate the potential of different topics and prioritize their funding support.

Finally, we can say that this study provides a tool for anyone interested in biomedical and life sciences to obtain a better understanding of the topics discussed in these fields. However, it should be noted that the definition and concept of the discovered topics may change over time.

Acknowledgements This work has been supported by the Spanish Ministry of Science and Innovation under Grants PID2019-105381GA-I00 (iScience) and PID2019-103880RB-I00.

References

Abu-Jbara, A. and Ezra, J. and Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 596–606.

Abu-Jbara, A. and Radev, D. R. (2012). Reference scope identification in citing sentences, Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, Montreal, Canada, pp. 80–90.

Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts. *Information Retrieval*, *13*, 101–131.

Alvarez, M. H., & Gómez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, *22*, 327–349.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of ACL conference (student session)* (pp. 81–87).

Athar, A. (2014). Sentiment analysis of scientific citations, Technical Report, University of Cambridge, Computer Laboratory, (UCAM-CL-TR-856).

Athar, A., & Teufel, S. (2012). Context-enhanced citation sentiment detection. In *Proceedings of HLT-NAACL*, *597–601*,

Bengisu, M. (2003). Critical and emerging technologies in materials, manufacturing, and industrial engineering: A study for priority setting. *Scientometrics*, *58*, 473–487.

Blei, D. M. and Lafferty, J. (2006). Dynamic topic models, Proceedings of the 23rd International Conference on Machine Learning (ICML), 113–120.

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

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *JASIST*, *66*, 2215–2222.

Bu, Y., Wang, B., Huang, W. B., Che, S., & Huang, Y. (2018). Using the appearance of citations in full text on author co-citation analysis. *Scientometrics*, *116*, 275–289.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*, 191–235.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, *57*, 359–377.

Chen, X., Chen, J., Wu, D., Xie, Y., & Li, J. (2016). Mapping the research trends by co-word analysis based on keywords from funded project. *Procedia Computer Science*, *91*, 547–555.

Chen, S. H., Huang, M. H., Chen, D. Z., & Lin, S. G. (2012). Detecting the temporal gaps of technology fronts: A case study of smart grid field. *Technological Forecasting and Social Change*, *79*, 1705–1719.

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. *Informetrics*, *11*, 1175–1189.

Cobo, M. J., Chiclana, F., Collop, A., Oña, J., & Herrera-Viedma, E. (2014). A bibliometric analysis of the intelligent transportation systems research based on science mapping. *IEEE Trans. Intelligent Transportation Systems*, *15*, 901–908.

Cobo, M. J., Martínez, M. A., Gutiérrez-Salcedo, M., Fujita, H., & Herrera-Viedma, E. (2015). 25 years at knowledge-based systems: A bibliometric analysis. *Knowledge-Based Systems*, *80*, 3–13.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2014). Research trends in gender differences in higher education and science: A co-word analysis. *Scientometrics*, *101*, 273–290.
Garfield, E. (1963). Science citation index. *Science Citation Index*, *1*.
Garfield, E. (1962). Can citation indexing be automated. *Essays of an Information Scientist*, *1*, 84–90.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, *178*, 471–479.
Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, *91*, 399–416.
Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. *Journal of the American Society for Information Science*, *49*, 674–685.
Grifiths, T.L. & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of national academy of sciences* 101 (Suppl. 1), USA, (pp. 5228–5235).
Guo, H., Weingart, S., & Börner, K. (2011). Mixed-indicators model for identifying emerging research areas. *Scientometrics*, *89*, 421–435.
He, J., & Chen, C. (2018). Temporal representations of citations for understanding the changing roles of scientific publications. *Frontiers in Research Metrics and Analytics*, *3*.
He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, C. L. (2010). Context-aware citation recommendation. *Proceedings of WWW Conference*, *421–430*,
Hu, C. P., Hu, J. M., Deng, S., & Liu, Y. (2013). A co-word analysis of library and information science in China. *Scientometrics*, *97*, 369–382.
Hui, S. C., & Fong, A. C. M. (2004). Document retrieval from a citation database using conceptual clustering and co-word analysis. *Information Review*, *28*, 22–32.
Hu, J., & Zhang, Y. (2015). Research patterns and trends of Recommendation System in China using co-word analysis. *Information Processing Management*, *51*, 329–339.
Jebari, C., Cobo, M. J., & Herrera-Viedma, E. (2018). *A new approach for implicit citation extraction, proceedings of IDEAL conference* (pp. 121–129). Spain: Madrid.
Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics*, *6*, 391–406.
Kajikawa, Y., & Takeda, Y. (2008). Structure of research on biomass and bio-fuels: A citation-based approach. *Technological Forecasting and Social Change*, *75*, 1349–1359.
Kim, H., Jiang, X., & Ohno-Machado, L. (2011). Trends in biomedical informatics: most cited topics from recent years. *JAMIA*, *18*, 166–170.
Kostoff, R. N. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, *68*, 223–253.
Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, E. O., & Ramirez, A. M. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal American Society Information Sciences Technology*, *52*, 1148–1156.
Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*, 575–603.
Lee, B., & Jeong, Y. (2008). Mapping Korea's national R&D domain of robot technology by using the co-word analysis. *Scientometrics*, *77*, 3–19.
Li, L., Ding, G., Feng, N., Wang, M., & Ho, Y. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics*, *80*, 39–58.
Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Li, Y. (2014). Literature retrieval based on citation context. *Scientometrics*, *101*, 1293–1307.
López-Robles, J. R., Otegi-Olaso, J. R., Gómez, I. P., & Cobo, M. J. (2019). 30 years of intelligence models in management and business: A bibliometric review. *International Journal of Information Management*, *48*, 22–38.
MacDonald, K. I., & Dressler, V. (2018). Using citation analysis to identify research fronts: A case study with the internet of things. *Science and Technology Libraries*, *37*, 171–186.
Ma, S., Zhang, C., & Liu, X. (2020). A review of citation recommendation: from textual content to enriched context. *Scientometrics*, *122*, 1445–1472.
Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., et al. (2009). (pp. 584–592) USA:
Moral-Munoz, J. A., Arroyo-Morales, M., Piper, B. F., Cuesta-Vargas, A. I., Díaz-Rodríguez, L., Cho, W. C. S., et al. (2018). Thematic trends in complementary and alternative medicine applied in cancer-related symptoms. *Journal Data Information Science*, *3*, 1–19.
Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, *54*, 413–422.

Muñoz-Leiva, F., Viedma-del-Jesús, M. I., Sánchez-Fernández, J., & López-Herrera, A. G. (2012). An application of co-word analysis and bibliometric maps for detecting the most highlighting themes in the consumer behaviour research from a longitudinal perspective. *Quality & Quantity*, *46*, 1077–1095.

Murgado Armenteros, E. M., Gutiérrez Salcedo, M., Torres Ruiz, F. J., & Cobo, M. J. (2015). Analysing the conceptual evolution of qualitative marketing research through science mapping analysis. *Scientometrics*, *102*, 519–557.

Ohniwa, R., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. *Scientometrics*, *85*, 111–127.

Perez-Cabezas, V., Ruiz-Molinero, C., Carmona-Barrientos, I., Herrera-Viedma, E., Cobo, M. J., & Moral-Munoz, J. A. (2018). Highly cited papers in rheumatology: Identification and conceptual analysis. *Scientometrics*, *116*, 555–568.

Qazvinian, V. & Radev, D. R. (2010). Identifying non-explicit citing sentences for citation based summarization. In *Proceedings of the 48th annual meeting* ACL. Uppsala, Sweden, pp. 555–564.

Reiss, T., Vignola-Gagne, E., Kukk, P., Glänzel, W., & Thijs, B. (2013). *ERACEP- Emerging research topics and their coverage by ERC-supported projects*. European Research Council: Technical Report.

Ritchie, A. (2009). *Citation context analysis for information retrieval*. UK: University of Cambridge.

Sagar, A., Kademani, B. S., & Bhanumurthy, K. (2013). Research trends in agricultural science: A global perspective. *Journal of Scientometric Research*, *2*, 185–201.

Schwartz, A. S. and Hearst, M. (2006). Summarizing key concepts using citation sentences, Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis, ser. BioNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 134–135.

Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2011). Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. *Technological Forecasting and Social Change*, *78*, 274–282.

Smalheiser, N. R. (2001). Predicting emerging technologies with the aid of text-based data mining: the micro approach. *Technovation*, *21*, 689–693.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269.

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics*, *68*, 595–610.

Small, H. (2011). Interpreting maps of science using citation context sentiments: A preliminary investigation. *Scientometrics*, *87*, 373–388.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*, 1450–1467.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*, 46–62.

Sugiyama, K., Kumar, T., Kan, M. Y., & Tripathi, R. C. (2010). *Identifying citing sentences in research papers using supervised learning, Proceedings of International Conference on Information Retrieval and Knowledge Management (CAMP), Shah Alam* (pp. 67–72). Malaysia: Selangor.

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

Teufel, S. and Siddharthan, A. and Tidhar, D. (2006). An annotation scheme for citation function, Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, 80–87.

Upham, S., & Small, H. (2010). Emerging research fronts in science and technology: patterns of new knowledge development. *Scientometrics*, *83*, 15–38.

Wang, Z. Y., Li, G., Li, C. Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics*, *90*, 855–875.

Yan, E., Chen, Z., & Li, K. (2020). The relationship between journal citation impact and citation sentiment: A study of 32 million citances in PubMed Central. *Quantitative Science Studies*, *1*, 1–11.

Yu, D., Xu, Z., & Wang, W. (2018). Bibliometric analysis of fuzzy theory research in China: A 30-year perspective. *Knowledge-Based Systems*, *141*, 188–199.

Zhang, Y., Chen, H., Lu, J., & Zhang, G. (2017). Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016. *Knowledge-Based Systems*, *133*, 255–268.

Zhang, G., Ding, Y., & Milojevic, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*, 1490–1503.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*, 373–401.