IntechOpen

Scientometrics

Edited by Mari Jibu and Yoshiyuki Osabe

SCIENTOMETRICS

Edited by **Mari Jibu** and **Yoshiyuki Osabe**

Scientometrics
http://dx.doi.org/10.5772/intechopen.72488
Edited by Mari Jibu and Yoshiyuki Osabe

Contributors

Nicola Bernabò, Rosa Ciccarelli, Alessandra Ordinelli, Zhao Qu, Maria Teresa Fernández-Bajón, Felix De Moya, Gerardo Tibaná-Herrera, Iñaki Bildosola, Rosa Río-Bélver, Enara Zarrabeitia, Gaizka Garechana, Alexander Maz-Machado, Noelia Jiménez-Fanjul, Wataru Souma, Marcel Clermont, Dirk Tunger, Andreas Meier, Ming Xiao, Zeshun Shi, Shanshan Wang, Esther Ferrandiz, Ana Fernández, M. Dolores León, Takahiro Kawamura, Katsutaro Watanabe, Naoya Matsumoto, Shusaku Egami, Yasuhiro Yamashita, Meen Chul Kim, Yongjun Zhu, Mari Jibu, Yoshiyuki Osabe

Notice

First published in London, United Kingdom, 2018 by IntechOpen
eBook (PDF) Published by IntechOpen, 2019
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Scientometrics
Edited by Mari Jibu and Yoshiyuki Osabe

p. cm.

Print ISBN 978-1-78923-306-3

Online ISBN 978-1-78923-307-0

eBook (PDF) ISBN 978-1-83881-692-6

Meet the editors

Mari Jibu has been a fellow at the Center for Research and Development Strategy, Japan Science and Technology Agency (JST), since 2008. She received her MBA degree from McGill University and her PhD degree in Medicine from Okayama University. Her professional experience includes an assistant from 1987 to 1994, a lecturer from 1994 to 2000, and an associate professor from 2000 to 2005, Notre Dame Seishin University. She also worked as a senior research fellow from 2005 to 2008 at NISTEP, MEXT. She also served as a consultant, an economist, and a policy analyst in the Organisation for Economic Co-operation and Development (OECD) (2013–2016). She received the Best Paper Award in 1997 from the 11th European Meeting on Cybernetics and Systems Research. Her research interests include science and technology policy, scientometrics, and quantum brain dynamics (QBD).

Yoshiyuki Osabe is a deputy director from the Patent Information Policy Planning Division, Japan Patent Office. He graduated from the University of Tokyo, Faculty of Pharmaceutical Science, and Graduate School in the University of Tokyo. With majors in Intellectual Property and Biotechnology, he joined the Japan Patent Office in 2002. Yoshiyuki is a visiting scientist in Catholic University of Leuven, Belgium; the deputy director in Bio-industry Division, METI; an economist/IP analyst in OECD; and an administrative judge in JPO. He has authored *Towards More Inclusive IP Analysis by Frontier Tools* (Intellectual Property Rights). His other authored works include *Innovation Front and Technology Linkage* (Business and Management Studies (2016), Vol. 2, No. 1, pp. 88–94) and *Future Information Technology* (Springer, Lecture Notes in Electrical Engineering, Vol. 309 (2014), pp. 549–554).

Contents

Preface XI

Section 1 Introduction 1

Chapter 1 **Introductory Chapter: Scientometrics 3**
Yoshiyuki Osabe and Mari Jibu

Section 2 Bibliometric Analysis 7

Chapter 2 **Scientometrics of Scientometrics: Mapping Historical Footprint and Emerging Technologies in Scientometrics 9**
Meen Chul Kim and Yongjun Zhu

Chapter 3 **Patterns of Academic Scientific Collaboration at a Distance: Evidence from Southern European Countries 29**
Ana Fernández, Esther Ferrándiz and M. Dolores León

Chapter 4 **Mapping a Research Field: Analyzing the Research Fronts in an Emerging Discipline 49**
Gerardo Tibaná-Herrera, María Teresa Fernández-Bajón and Félix de Moya-Anegón

Chapter 5 **Collaboration and Citation Analysis Within Social Sciences: A Comparative Analysis Between Two Fields 65**
Alexander Maz-Machado and Noelia Jiménez-Fanjul

Chapter 6 **A Scientometric Study on Graphene and Related Graphene-Based Materials in Medicine 83**
Nicola Bernabò, Rosa Ciccarelli, Alessandra Ordinelli, Juliana Sofia Somoes Machado, Mauro Mattioli and Barbara Barboni

Chapter 7 **Technology Roadmapping of Emerging Technologies: Scientometrics and Time Series Approach 99**
Iñaki Bildosola, Rosamaría Río-Bélver, Gaizka Garechana and Enara Zarrabeitia

Chapter 8 **Altmetrics: State of the Art and a Look into the Future 123**
Dirk Tunger, Marcel Clermont and Andreas Meier

Section 3 Patent Analysis 135

Chapter 9 **Patent Research in a Period of Industry Transformation: A Focus on Electromobility 137**
Zhao Qu

Chapter 10 **Exploring Characteristics of Patent-Paper Citations and Development of New Indicators 151**
Yasuhiro Yamashita

Section 4 Content-based Analysis 173

Chapter 11 **Mapping Science Based on Research Content Similarity 175**
Takahiro Kawamura, Katsutaro Watanabe, Naoya Matsumoto and Shusaku Egami

Chapter 12 **The Impact on Citation Analysis Based on Ontology and Linked Data 195**
Ming Xiao, Zeshun Shi and Shanshan Wang

Chapter 13 **Progress of Studies of Citations and PageRank 213**
Wataru Souma and Mari Jibu

Preface

Scientometrics has provided a gentle introduction to empower decision-makers to make sense of science, technology, and innovation data to improve daily decision.

Section 1 argues for bibliometric analysis in the study of science, technology, and innovation structure. The science of scientific publication studies and information visualization explain facts at different levels of science, technology, and innovation system.

Section 2 introduces patent analysis in the study of science, technology, and innovation structure and dynamics. Linkage between scientific publications and patents explains knowledge flows from science to technology.

Section 3 examines content-based analysis and discusses the possible impact of real-time visualization. It includes methodology for analyzing publications based on the combination of word embedding and entropy-based approach.

Those analyses in this book are very useful for the scientometrics community.

Our sincere thanks go to Julian Virag at InTechOpen Limited who ingeniously mastered the many complexities involved in publishing the book *Scientometrics*. We are indebted to our family and friends for providing much inspiration, energy, and loving support.

Mari Jibu
Japan Science and Technology Agency, Japan

Yoshiyuki Osabe
Japan Patent Office, Japan

Introduction

Chapter 1

Introductory Chapter: Scientometrics

Yoshiyuki Osabe and Mari Jibu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78027

1. Introduction

Scientometrics has been defined as the "quantitative study of science, communication in science, and science policy" [1]. Over 20 years have passed since Hess's definition and now it has been used in many different fields. As representative works in the field of scientometrics, we can refer the Science Citation Index (SCI) [2, 3], the first Academic Ranking of World Universities (ARWU) of the Shanghai Jiao Tong University in 2004 [4], the h-index [5], g-index [6], and so on. Among these indicators, the h-index provides a simple impact metric for individual authors that can readily be used in online searching, for example, with Google Scholar, but is also incorporated into the major citation databases such as the Web of Science and Scopus.

The international organizations like OECD and the National Governments have also followed the activity related to scientometrics. For example, the Organization of Economic Cooperation and Development (OECD) has published "Science, Technology and Industry Scoreboard" once every 2 years. In terms of an example of National Government activities, the National Science Board (NSB) in US also publishes "Science & Engineering Indicators" once every 2 years. In these publications, scientometrics indicators contribute to OECD and NSB efforts, especially in terms of standardization of calculations, collection of data, and analysis of a wide range of science, technology, and innovation activities by providing evidence on a selected set of Science and Technology (S&T) output.

Therefore, the concept of scientometrics has already disseminated to our society and has become essential for evidence-based policy makings, especially in the fields of S&T and Innovation.

2. Main points of this book

Technological change is one of the greatest issues in the modern world. As the world faces societal challenges, for example, climate challenges, aging problem, and energy security, technology will contribute to new or better solutions for those problems. New technologies take longer to develop and mature; moreover which tend to be born in the interconnection of multiple technology fields, therefore early detection of emerging technological concepts across multiple disciplines will be a very important issue.

Our goal is to seek to develop automated methods that aid the systematic, continuous and comprehensive assessment of technological emergence using one of the major foresight exercises, scientometrics. There is now a huge flood of scientific and technical information, especially scientific publications and patent information. Using the information patterns of emergence for technological concepts have been discovered and theories of technical emergence have also been developed in several years. We have been developing visualization tools that thousands of technical areas have been interacted with each other and evolved in time. Several indicators of technical emergence have been improved by universities, international organizations, and funding agencies.

This book intends to provide readers a comprehensive overview of the current state-of-the-art in scientometrics, focusing on the systematic, continuous and comprehensive assessment of technological emergence. This book is composed of 12 chapters by cutting-of-edge authors of many different nationalities from Europe to Asia.

Especially the chapter "Mapping Science based on research content similarity" by Dr Kawamura shows an interesting methodology for analyzing publications based on an adaptation of word embedding and paragraph embedding with an entropy-based word clustering methodology. The proposed combination of word embedding and entropy-based approach is very useful for the scientometrics community.

3. Conclusions and future perspective

Last but not least, we would like to mention an expected future landscape of this field. Now it is evolutionary time from basic research phase to implementation phase and scientometrics will be expected to be applied to the fields below at the implementation level.

3.1. IP landscape

Recently "IP landscape" has been referred in the field of intangible assets. IP landscape provides not only a snapshot but also a strategic analysis of the IP trends of a specific technology field within either a given company or a given country. It is said that the techniques or tools in scientometrics are very useful for the needs of IP landscape as following:

(i) understanding of IP for products and technologies,

(ii) building a simple model,

(iii) identification of key technology players,

(iv) discover of white areas where no one achieves a field yet, and

(v) understanding of stakeholders (e.g., competitors, upstream and downstream partners, potential acquisition target).

3.2. Data-driven innovation

Recently, creating a new business and solving social problems utilizing big data have been expected to increase. The Ministry of Economy, Trade and Industry in Japan is supporting business creation through data utilization, and enterprises are developing advanced measures in the fields such as agriculture and medical care. On the other hand, new cooperation beyond industrial barriers between a present entity and a new entity created sharing data is still limited. For the economic development in near future, so-called "Data-Driven Innovation" will be necessary for firms: for example firms will utilize data sharing beyond entities, creating new added value. Since companies, especially SMEs, have rarely data scientists who deal with big data, scientometrics indicator or tools thereof can contribute to enhancement of the data-driven innovation.

3.3. Fields close to scientometrics

Although "scientometrics" is mainly a study of relations between text of articles or patents and their authors/institutions, it is also highly corresponded to "science of sociology" which is mainly a study of relations between authors/institutions and text networking, or AI-related fields like "semantic search" and "machine translation." Interdisciplinary research with other fields is expected.

A reconstruction or remodeling of S&T fields above mentioned reinforces the knowledge-based development in terms of society and economy. Scientometrics will be able to foster a development of science, technology, and innovation by a quantitative perception and evidence-based policy making. Further study and development of scientometrics are expected in future.

Author details

Yoshiyuki Osabe[1]* and Mari Jibu[2]

*Address all correspondence to: osabe-yoshiyuki@jpo.go.jp

1 Patent Information Policy Planning Division, Japan Patent Office, Japan

2 Center for Research and Development Strategy, Japan Science and Technology Agency, Japan

References

[1] Hess DJ. Science Studies: An Advanced Introduction. New York: New York University Press; 1997

[2] Garfield E. Citation Indexing: Its Theory and Application in Science, Technology, and Humanities. New York: John Wiley; 1979

[3] Wouters P. The citation culture [unpublished Ph.D. thesis]. Amsterdam: University of Amsterdam; 1999

[4] Shin JC, Toutkoushian RK, Teichler U. University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education. Dordrecht: Springer; 2011

[5] Hirsch JE. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America. 2005; **102**(46):16569-16572

[6] Egghe L. Theory and practise of the g-index. Scientometrics. 2006;**69**(1):131-152

Section 2

Bibliometric Analysis

Chapter 2

Scientometrics of Scientometrics: Mapping Historical Footprint and Emerging Technologies in Scientometrics

Meen Chul Kim and Yongjun Zhu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77951

Abstract

Scientometrics is the study of quantitative aspects of science, technology, and innovation. This chapter identifies thematic patterns and emerging trends of the published literature in scientometrics using a variety of tools and techniques, including CiteSpace, VOSviewer, and dynamic topic modeling. Using 8098 bibliographic records of published scientometrics research, we explored domain-level citation paths, subject category assignment, keyword co-occurrence, topic models, and document co-citation network to map and characterize the intellectual landscapes of scientometrics. Findings reveal that the domain is multi-disciplinary in that a wide range of disciplines contribute to the growth of literature, but only partially interdisciplinary as some works heavily cites from similar domains. Early literature was interested in measuring the impact of a science and evaluating research performance and productivity. Modeling scientometrics laws and indicators is also of greatest interest. Later work explored applications of scientometrics to a variety of domains such as material sciences, medicine, environmental sciences, and social media analytics. Impact measure and science mapping are among the topics receiving consistent attention.

Keywords: scientometrics, science mapping, domain analysis, visual analytics, intellectual structure, emerging technologies

1. Introduction

Scientometrics is the quantitative study of science. It aims to analyze and evaluate science, technology, and innovation. Major research includes measuring the impact of authors, publications, journals, institutes, and countries as referenced to sets of scientific publications such as

articles and patents. It also aims to understand the behavior of scientific citations as a mean of scholarly communication and map intellectual landscapes of a science. Other effort focuses on the production of indicators for use in the evaluation of performance and productivity [1]. In practice, there is a significant overlap between scientometrics and other neighboring domains such as bibliometrics, informetrics, webometrics, and cybermetrics. Bibliometrics, one of the canonical research domains in library and information science, studies quantitative aspects of written publications. Informetrics is the study of quantitative aspects of information [2], regarded as an umbrella domain overarching the rest of them. Björneborn and Ingwersen [3] describe the relationships between these domains as abstracted in **Figure 1**.

Driven by a variety of research communities, the volume of published literature in these domains has exponentially grown. Given the increasing publications and the scientific diversity in disciplines, a systematic investigation of the intellectual structure is in need to identify not only emerging trends and new developments but also historic areas of innovation and current challenges. The motivation of the present chapter lies in our intention to identify the intellectual structure of scientometrics in a systematic manner. Toward that end, we explore epistemological characteristics, thematic patterns, and emerging trends of the field, using scientometrics approaches. In particular, we operationalize scientometrics as encompassing closely related domains such as informetrics, bibliometrics, cybermetrics, and webometrics. In the rest of this manuscript, we use the term "scientometrics" inclusively. The present chapter aims to trace the evolution and applications of scientific knowledge in scientometrics. Thus, we also operationalize emerging trends and recent developments uncovered throughout the present chapter as "emerging technologies" in scientometrics.

The contributions of the present chapter include followings. First, it helps the scientometrics community to be more self-explanatory as it has a detailed publication-based profile. Secondly, researchers in the field can benefit from this systematic domain analysis by

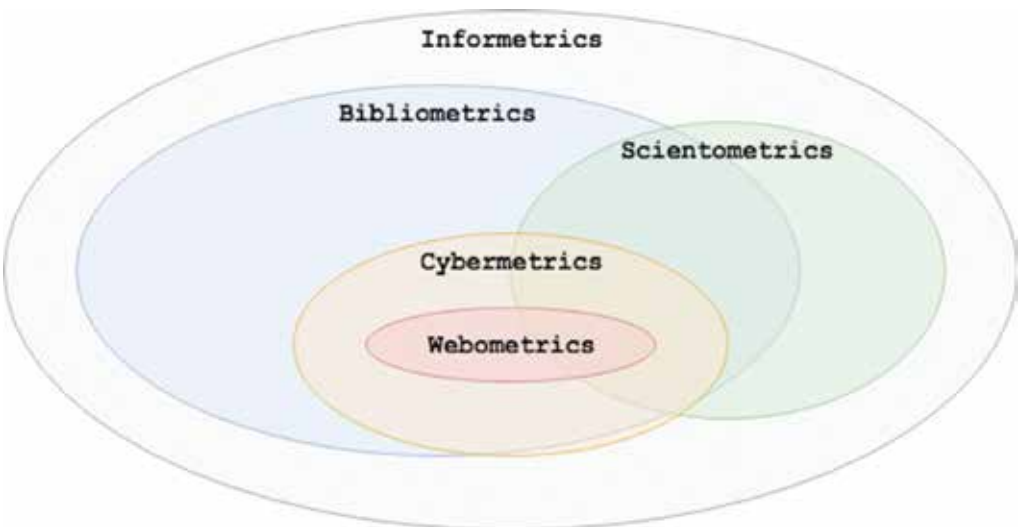


Figure 1. Relationships between metrics sciences re-cited from [3].

identifying emerging technologies, better positioning their research, and expanding research territories. Finally, it guides those interested in the field to learn about historic footprint and current issues.

The rest of the chapter is organized as follows. We introduce the methodology of the study. Then, the intellectual landscapes of scientometrics is described. We conclude this chapter with discussion into findings, implications, and limitations.

2. Methodology

This section details our data collection method and analytical approaches. **Figure 2** pipelines the research procedure.

2.1. Data collection

The present chapter explores the intellectual structure of published literature in scientometrics. Considering the aforementioned operationalization of scientometrics, we conducted a topic search on the web of science (WoS). The search query consisted of seven terms as follows: Bibliometric* OR scientometric* OR informetric* OR webometric* OR altmetric* OR cybermetric* OR entitymetric*. The wildcard character "*" captures any relevant variations of a term such as bibliometrics and bibliometric analysis. A bibliographic record is considered as relevant if any of the terms appear in its title, abstract, or keywords. As of December 31, 2017, the query returned 8098 bibliographic records written in English between 1990 and 2017. The subscription of the authors' institutes covered from 1980s at the time of querying, but in many

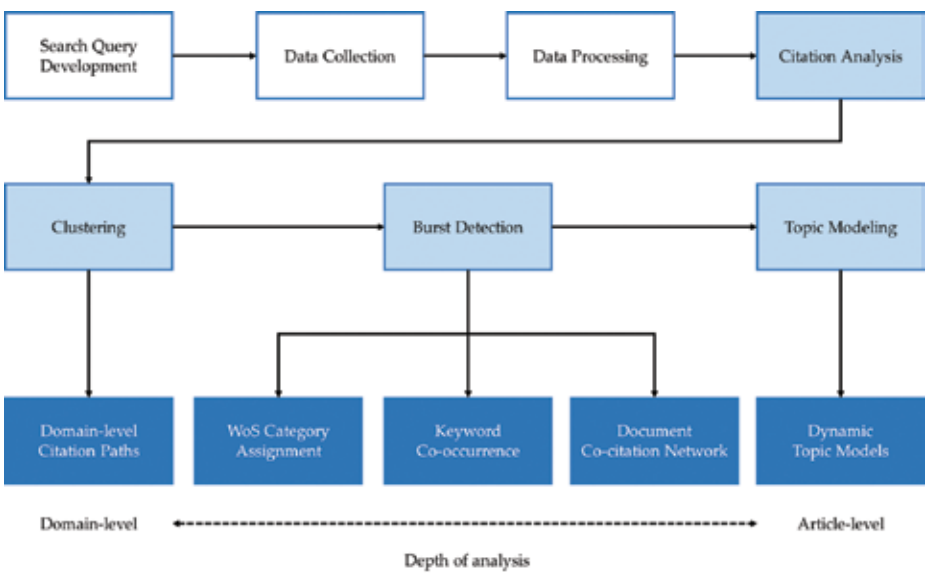


Figure 2. Research procedure.

cases text fields were omitted. Thus, we excluded data before 1990. The brief statistics of the retrieved data set is described in **Table 1**.

Figure 3 renders the record distribution over time in our data collection. As illustrated, there has been exponentially increasing interest in scientometrics from the community.

Table 2 describes the contributing terms to the data retrieval and corresponding number of records to each term. As shown, the literature has used "bibliometric*" the most frequently.

2.2. Investigating the intellectual structure in scientometrics

Scientometrics depicts the intellectual landscapes of a science with a variety of bibliographic units such as authors, keywords, texts, and citations and networks of those entities. The present chapter systematically mapped historical footprint and emerging technologies from published research in scientometrics. In particular, we investigated citation paths at a disciplinary level, co-occurrence of WoS categories and keywords, and networks of co-cited references. Network clustering and topic modeling were also used to find homogeneous sets of literature and coherent streams of research. In so doing, we captured emerging trends, recent developments, and current challenges in the domain. Especially, we employed a top-down approach in analyzing data going from macro-level to micro-level. It had us add richer interpretations as we gradually moved on to lower-level units of analysis such as journal-level citation paths, subject categories, keywords, titles and abstracts to cited references. To this end, this chapter is mainly guided by

Duration	Total	Articles	Procs.	Reviews	Authors	Keywords	Refs.
1990–2017	8098	7013	413	672	23,791	98,493	328,096

Table 1. Data statistics.

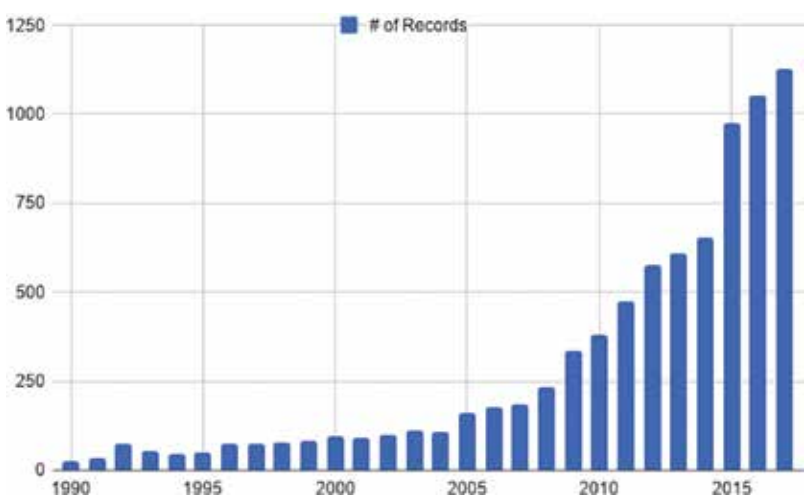


Figure 3. The distribution of records over time.

Term	Duration	Total	Articles	Procs.	Reviews
bibliometric*	1990–2017	6352	5449	313	590
scientometric*	1990–2017	1779	1577	93	109
informetric*	1990–2017	382	334	28	20
webometric*	1997–2017	288	254	25	9
altmetric*	2012–2017	261	237	7	17
cybermetric*	1999–2015	28	27	1	—
entitymetric*	2013–2015	3	3	—	—

Table 2. Querying terms (the wildcard character "*" captures any relevant variations of a term).

two suites of software, namely CiteSpace [4–6] and VOSviewer [7]. The input is a collection of bibliographic records relevant to a topic of interest. Given the records, the toolkits detect and render thematic patterns and emerging trends in science as networked in a variety of bibliographic units. As argued by preceding papers [8, 9], this chapter's approaches have several methodological merits over a conventional domain analysis. First, a much more inclusive range of topically relevant literature can be examined. Second, an inquiring individual does not need prior expertise to analyze a domain of interest. Finally, this kind of survey can be conducted as frequently as in need given the fast growth of a science. The underlying techniques and findings of the present chapter could be more clearly delivered as we introduce followings:

- Network reduction: In network analysis, investigating the entire nodes and edges between them is computationally challenging. It may not intuitively communicate the topological structure to the audience as well for it is visually overwhelming with many links. To handle this, we select up to 100 frequently occurring entities such as keywords and cited references within a one-year time slice.
- Clustering: Clustering is unsupervised learning which uncover latent groups of entities sharing homogeneous characteristics. We employ a network clustering technique called smart local moving [10] to capture thematically similar clusters on a document co-citation network.
- Burst detection: Proposed by [11], burst detection models the burstiness of features which rise sharply in frequency. An entity has bursting activities when it intensively appears during a specific span of time. We can overcome the limitation coming from considering cumulative, snapshot metrics as impact measures.
- Cluster labeling: CiteSpace labels clusters with extracted terms from titles and abstracts of citing articles. There are three algorithms to serve cluster labeling: (1) latent semantic analysis (LSA), (2) log-likelihood ratio (LLR), and (3) mutual information (MI). LSA captures unknown semantic relationships over all the documents while LLR and MI reflect a unique aspect of a cluster [5].
- Topic modeling: Topic modeling is unsupervised machine learning which aims to discover latent semantic structure occurring in a text body. We employ dynamic topic modeling

(DTM) which is a generative technique extended from Latent Dirichlet Allocation (LDA). DTM captures the evolution of latent topics in a collection of documents whereas it was oblivious to the preceding model [12].

3. Results

3.1. Domain-level research patterns

Citation paths at a disciplinary level are depicted in the visual representation called a dual-map overlay [6] (see **Figure 4**). The left regions represent where the collected literature publishes while the right regions render where it cites from. Citing literature and cited literature are also called research frontier and knowledge base respectively. The base map consists of the journal/conference-level citation relationships among over 10,000 venues. Major clusters are labeled by terms chosen from the titles of venues in corresponding clusters. First, all of the terms' log-likelihood ratios are calculated based on their frequency in clusters. The use of LLR achieves to represent those terms' uniqueness in clusters. Then, top three terms are selected to tag clusters, based on their LLR values in descending order. Citation trajectories are colored based on the citing regions. The width of the paths is proportional to the z-score-scaled citation frequency.

Table 3 describes these trajectories in descending order of the third column, namely Z-score. The color of each row is corresponding to the path. Findings indicate that scientometrics has been largely driven by social sciences and medicine as represented by "psychology, education, health" and "medicine, medical, clinical" respectively at the first column. Literature from social sciences heavily cites from "psychology, education, social", "systems, computing, computer", "health, nursing, medicine", "economics, economic, political", and "molecular, biology, genetics", yielding five citation paths. Research frontiers from medicine are based on "health, nursing, medicine" and "molecular, biology, genetics", having two additional trajectories. These observations

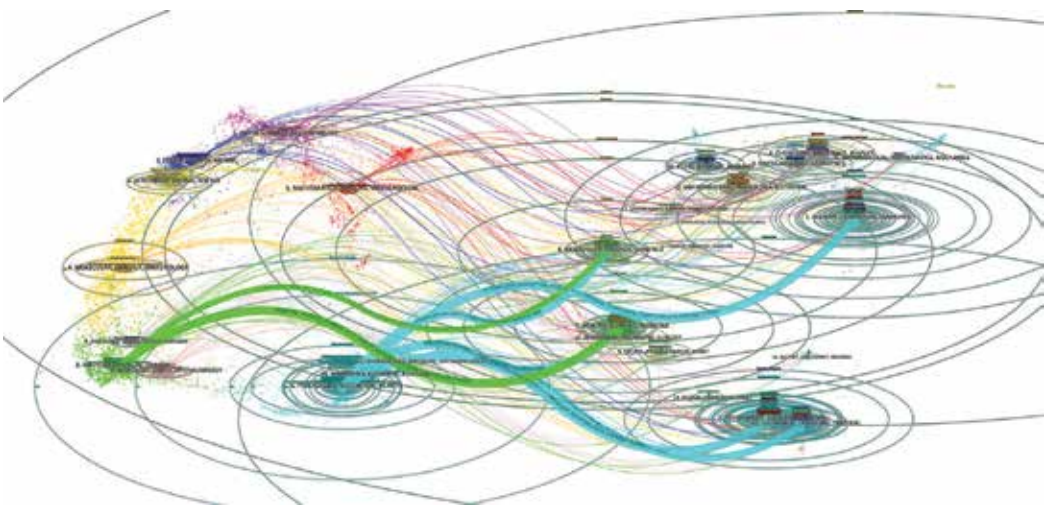


Figure 4. Citation paths at a disciplinary level.

Research frontier	Knowledge base	Z-score
Psychology, education, health	Psychology, education, social	8.841
Psychology, education, health	Systems, computing, computer	4.766
Medicine, medical, clinical	Health, nursing, medicine	4.052
Psychology, education, health	Health, nursing, medicine	3.313
Psychology, education, health	Economics, economic, political	2.724
Psychology, education, health	Molecular, biology, genetics	2.461
Medicine, medical, clinical	Molecular, biology, genetics	1.984

Table 3. Domain-level citation trends.

WoS category	Year	Frequency	Density
Information science & library science	1990	3880	138.571
Computer science	1990	3260	116.429
Computer science, interdisciplinary applications	1990	2284	81.571
Computer science, information systems	1990	925	33.036
Business & economics	1992	653	25.115
Management	1992	374	14.385
Engineering	1992	292	11.231
Public administration	1992	199	7.654
Planning & development	1992	179	6.885
Education & educational research	1992	165	6.346
Social sciences – other topics	1992	160	6.154
Science & technology – other topics	1993	462	18.480
Multidisciplinary sciences	1993	348	13.920
Business	1994	242	10.083
Neurosciences & neurology	1996	159	7.227
Environmental sciences & ecology	1997	261	12.429
General & internal medicine	1999	145	7.632
Surgery	2000	162	9.000
Public, environmental & occupational health	2003	201	13.400
Environmental sciences	2006	189	15.750

Table 4. Top 20 frequently assigned WoS categories.

show scientometrics is multidisciplinary and partially interdisciplinary; Multidisciplinary since scientometrics research has been published in multiple disciplines; Partially interdisciplinary for literature published in "psychology, education, health" has a variety of intellectual bases while "medicine, medical, clinical" largely cites from neighboring domains.

We considered WoS category assignment to literature as another important indicator representing domain-level thematic concentration. The top 20 frequently assigned WoS categories to the records are described in **Table 4**. It shows the year it was first assigned, and the density of how many times per year a specific category has been given, from its first year. The table is sorted in ascending order of the year. Results show that three categories have been assigned more than 2000 times – "information science & library science" (n = 3880), "computer science" (n = 3260), and "computer science, interdisciplinary applications" (n = 2284). These categories were first assigned from the beginning in the data set, demonstrating the greatest densities. The most frequently assigned category to be added to the top four list is "computer science, information systems." This category also demonstrates a relatively high density (33.036), given its first year of assignment was 1990. This finding suggests that literature under these four categories has had the largest influence on the emergence and development of scientific knowledge in scientometrics. In turn, research with scientific foci in social sciences, engineering, medical & health sciences, and environmental sciences brought along a multidisciplinary grasp to the domain.

3.2. Trending keywords

Given by authors and indexers, keywords reflect representative concepts underlying published literature. The top 20 frequently occurring keywords in the data set are described in **Table 5**. It shows the year it first appeared, and the density of how many times on average a specific keyword has appeared, from its first year. Findings indicate that in the beginning, "bibliometrics" and "scientometrics" focused on employing "citation analysis" to examine the "impact" of a "science". We assume that "journal" and "publication" were considered as units of analysis. Another effort focused on evaluating research "performance" and "productivity" and examining the "pattern" of scientific "collaboration." The other stream of research had interest in devising a "bibliometric indicator" such as journal "impact factor", which led to the recent development of the widely accepted author-level metric "h-index."

Keyword	Year	Frequency	Density
Science	1991	1613	59.741
Bibliometric analysis	1991	871	32.259
Journal	1991	815	30.185
Citation	1991	803	29.741
Bibliometrics	1992	1914	73.615
Impact	1992	969	37.269
Citation analysis	1992	814	31.308
Publication	1992	700	26.923
Scientometrics	1992	646	24.846
Indicator	1992	596	22.923
Performance	1992	348	13.385

Keyword	Year	Frequency	Density
Productivity	1992	270	10.385
Collaboration	1993	353	14.120
Bibliometric indicator	1993	290	11.600
Pattern	1993	273	10.920
Network	1994	357	14.875
Impact factor	1996	527	23.955
Index	2002	324	20.250
h-index	2007	386	35.091
Scopus	2008	280	28.000

Table 5. Top 20 frequently occurring keywords.

Figure 5 displays the keyword co-occurrence in the data set. We used a technique called a density visualization guided by VOSviewer. The font size of a keyword is proportional to its occurrence frequency. The more frequently a pair of keywords co-occurs, the closer the pair is located to the red spots. The visualization resulted in 484 keywords which occurred more than or equal to 18 times. As depicted, "bibliometrics" frequently co-occurred with "impact" which is consistent with the finding above. It also determined that devising an "impact factor" for "journal ranking" was among the important themes in scientometrics.

Table 6 lists 20 keywords which have surged during a specific duration of time. The investigation of keyword bursts adds temporal contexts in understanding historic footprint and emerging technologies in scientometrics which were oblivious to the snapshot metrics. The keywords were sorted in ascending order of the beginning years of bursts. "physics" is one of the keywords with the longest bursts, ending in 2010. It also has the second strongest bursts when not including "science." It indicates applications of scientometrics to physics and/or knowledge transfer from physics to scientometrics had intensively been conducted from the early years. The widely accepted author-level metric, namely h-index, was also derived from physics. The second longest bursts from 1992 is led by "law", also demonstrating a relatively high value of bursts. It shows the identification of laws existing in scientometrics phenomena was among the important initiatives. "publication output" is the keyword with the third longest and strongest bursts. It is argued that the evaluation of research performance and productivity was one of the key themes in the domain. The strongest burst episode from 1992 is associated with "indicator." In consideration with other keywords such as "stationary distribution", "model", and "informetric distribution", we argue modeling an indicator of impact measure was of greatest interest in scientometrics.

3.3. Temporal topic models

We analyzed another text fields, namely titles and abstracts since more informational points of content can be examined than only exploring keywords. We aimed to uncover the evolution of latent topics in the records over time. Toward that end, we removed stop words from

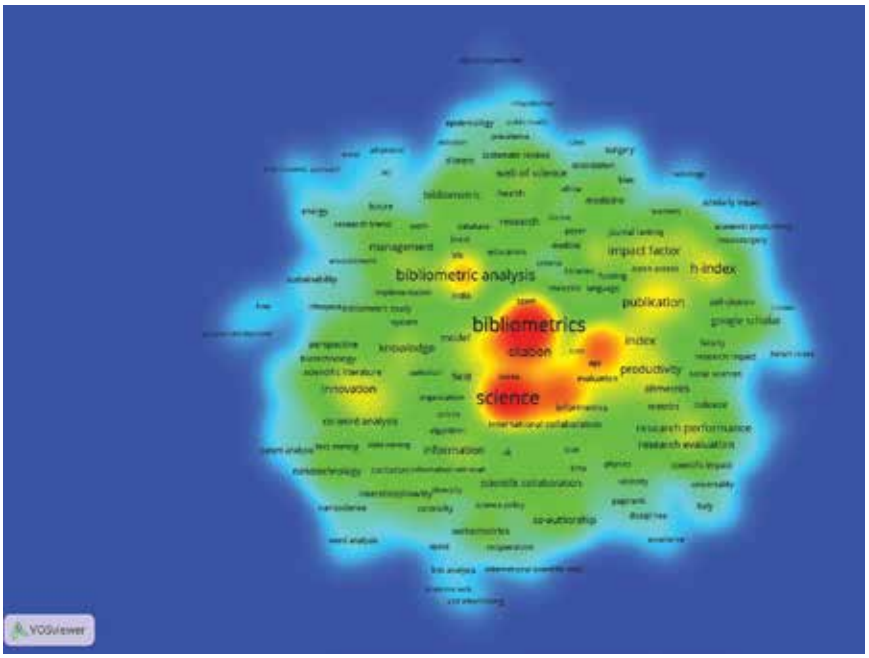


Figure 5. Keyword co-occurrence network (n = 484).

text, using a list of stop words in Python NLTK. The text was lowercased, tokenized, and de-accented. Then, we lemmatized the tokens and extracted noun phrases by bigram indexing. Text pre-processing and topic modeling were driven by gensim, a robust text mining toolkit in Python. **Table 7** describes 20 topics and 10 corresponding terms per topic. The terms were sorted in descending order of the average probabilities over the 28 years. Results show that most of the terms having high probabilities are unigram-formed.

Figure 6 illustrates the topical trends from 1990 till 2017 using a visualization technique called a bump chart. The topics are sorted in descending order of normalized probability distributions in the beginning year. We further discuss nine prominent topics, Topics 9, 17, 7, 4, 1, 5, 11, 16, and 0, due to their relatively high probability distributions. We categorized these topics into four trends: (1) rising, (2) rising-falling, (3) falling, and (4) static.

1. Rising topics: Topics 9, 17, 7, and 1 are consistently rising. Topic 9 we labeled "applications of scientometrics to material sciences" has received the greatest attention over time. Topic 17 which has sharply increased is named "publication-based scholarly communication." Topics 7 and 1 have been always in the top topic list and recently received increasing attention. We labeled them "evaluation of funded research" and "applications of scientometrics to medical education" respectively. Findings indicate that applications of scientometrics to domains other than biomedical sciences are of increasing concerns in the scientific community.
2. Rising-falling topics: Topics 4, 16, and 0 repeat rising and falling. Topic 4 can be named "literature-based research in healthcare." Topics 16 and 0 can be understood as "applications

of scientometrics to biomedicine" and "literature-based research in medicine" respectively. Knowledge discovery in healthcare and biomedical sciences has been among the greatest interest in scientometrics. We assume that this stream of research has ups and downs based on the change of scientific foci.

3. Falling topics: Topic 5 has fallen. We labeled it "history and philosophy of scientometrics." It is obvious that a study of theory and practice tends to be prominent in early years of a science. As staging into the maturation, this kind of topic naturally moves way from interest. It has also decreased in scientometrics.

4. Static topics: Topic 11 has been statically distributed over time. Based on the extracted terms, Topic 11 is interpreted as "mapping intellectual structure using citation and network analysis." This is one of the canonical research themes in scientometrics receiving consistent attention from the beginning of the domain.

Keyword	Burst	Begin	End	1990 - 2017
science	30.062	1990	2001	
british science	3.371	1990	1997	
stationary distribution	6.931	1992	2003	
model	10.688	1992	2002	
fact	4.116	1992	1993	
citation impact	4.850	1992	1997	
library	10.410	1992	2004	
physics	13.560	1992	2010	
co-citation analysis	3.430	1992	1993	
informetric distribution	6.299	1992	2006	
relative citation impact	3.430	1992	1993	
newest version	4.116	1992	1993	
indicator	17.159	1992	2000	
country	10.903	1992	2001	
law	10.395	1992	2007	
publication output	11.439	1992	2006	
figure	4.116	1992	1993	
cooperation	4.189	1993	2005	
combined co-citation	3.607	1994	2005	
journal	4.427	1994	1996	

Table 6. Top 20 keywords with the greatest intensive burstiness.

Topic 0	**Topic 1**	**Topic 2**	**Topic 3**
article	psychology	publication	productivity
journal	education	cancer	faculty
author	nursing	document	publication
article published	Brazilian	drug	index
number	research	research	gender
literature	study	descriptor	result
study	psychiatry	Korean	study
research	theses	Latin American	conclusion
medicine	school	literature	woman
publication	aids	drug	year
Topic 4	**Topic 5**	**Topic 6**	**Topic 7**
health	science	research	research
research	history	country	evaluation
publication	scientometrics	science	impact
public health	book	collaboration	funding
literature	reception	publication	assessment
medicine	removal	output	policy
method	philosophy	physics	researcher
result	nature	university	project
disease	colleague	study	scientist
health care	sport	productivity	work
Topic 8	**Topic 9**	**Topic 10**	**Topic 11**
performance	technology	research	structure
indicator	literature	field	analysis
research	patent	analysis	map
bibliometric indicator	nanotechnology	information	network
quality	serial	study	mapping
evaluation	indexing	science	citation
group	application	development	data
measure	development	data	cluster
data	material	paper	database
peer review	core	knowledge	method
Topic 12	**Topic 13**	**Topic 14**	**Topic 15**
study	distribution	information	paper
population	model	web	research
method	data	library	publication
country	index	link	literature

disease	two	use	country
data	one	online	journal
research	theory	library information	period
result	paper	search	number
health	number	internet	sci
water	function	study	study
capacity	system	subject	bibliometric analysis
Topic 16	**Topic 17**	**Topic 18**	**Topic 19**
research	communication	journal	ecology
rehabilitation	bibliometrics	citation	species
stem cell	scholarly communication	analysis	geography
neuroscience	dss	impact	climate change
credit	publishing	study	city
guideline	science	impact factor	conservation
paper	library information	paper	knowledge
study	media	reference	biodiversity
transplantation	theory	science	tourism
article	impact	author	study

Table 7. 20 generated topics.

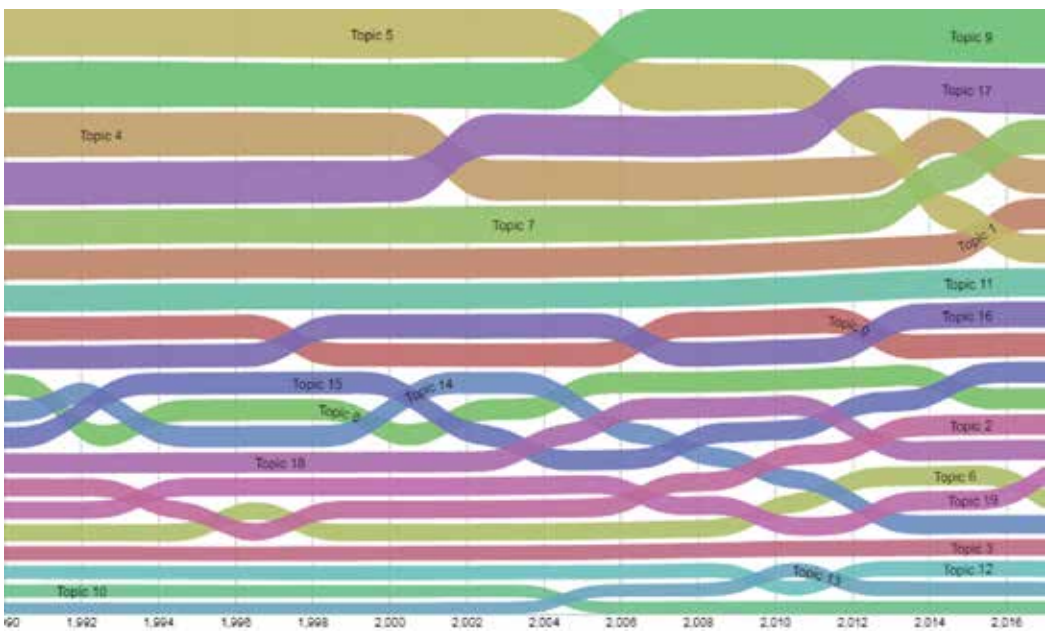


Figure 6. Topical trends.

3.4. Document co-citation network

Previous section utilized titles and abstracts to investigate topical trends without any bound context. This section examined those fields in a context of document-level co-citation relationship. **Figure 7** visualizes the document co-citation network in the data set. Each node is a cited reference extracted from the reference sections of the records and the size of the node is proportional to its cumulative frequency of received citations. Nodes with inner circles in

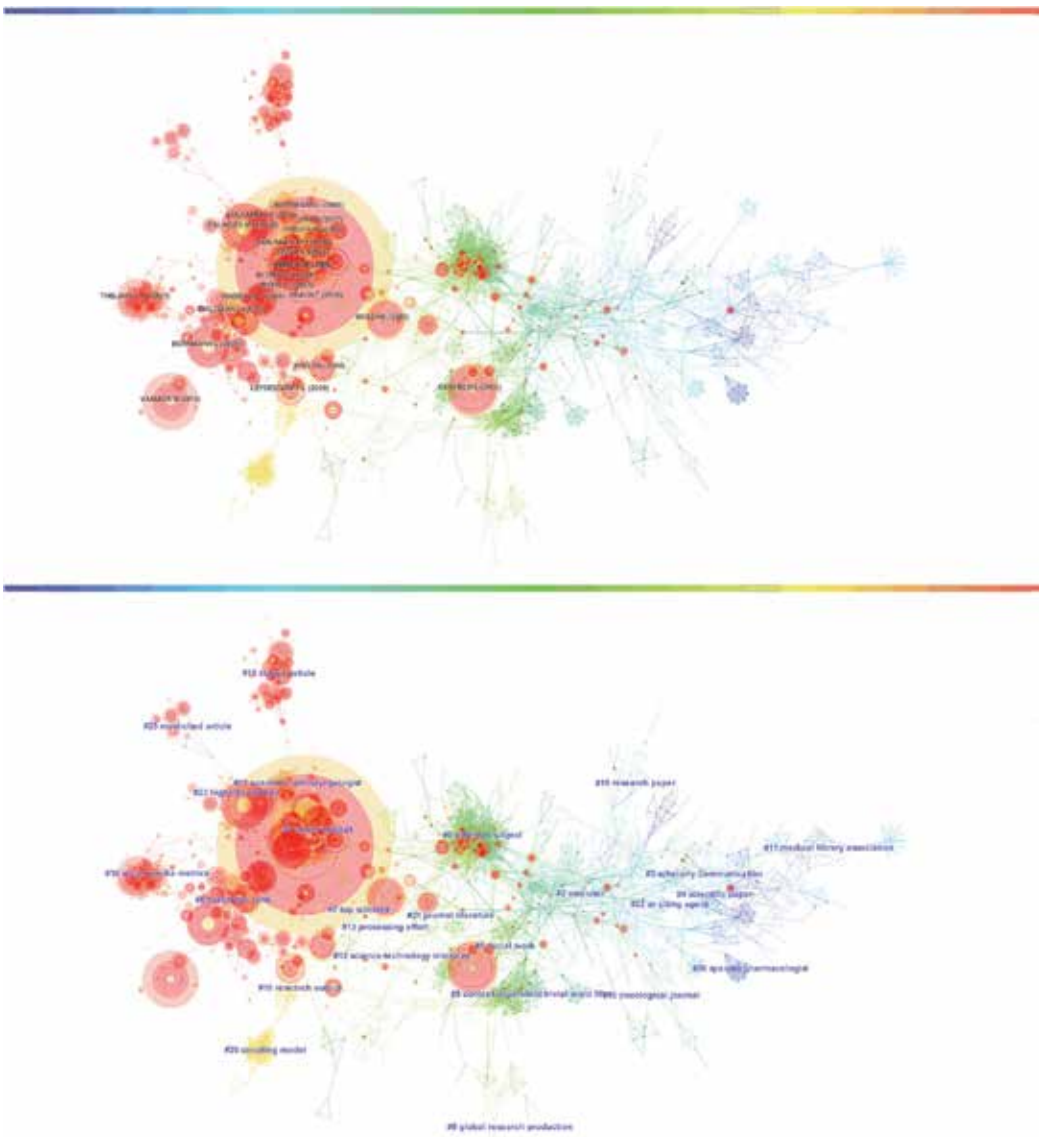


Figure 7. Document co-citation networks with truncated labels of first authors' names and published years (upward) and cluster labels (downward) (n = 1856, e = 6127).

red represent articles with citation bursts. We labeled the most highly cited 20 articles in black following a truncated form of <LAST NAME> < ABBREVIATED FIRST NAME> (<YEAR>) so as to only display first authors' names and published years (see the upward in **Figure 7**). They are cited more than or equal to 95 times locally, meaning in the data set. The color legend at the top of the display indicates links and citations in cooler colors happen more closely to 1990 whereas hotter ones occur in closer years to 2017. Based on the color scheme, we can keep track of the evolution of the document network. Findings show that most of the landmark articles were published relatively recently. Cumulative citations and citation bursts also intensively happened with these articles. Next, we conducted clustering and labeled the clusters in blue, using LLR (see the downward in **Figure 7**). Clusters are numbered in such a way that higher rankings are given to the clusters containing more references. In order to add richer contexts in interpreting the clustering results, we generated another visualization called a timeline visualization (see **Figure 8**).

In **Figure 8**, we re-grouped all the nodes on multiple lines so that the cluster memberships can be more accessibly identified. As depicted in the figure, emerging trends can further be captured by examining Clusters 1, 6, 10, 16, 17, 18 given cluster sizes, recency, cumulative citations, and citation bursts. **Table 8** summarizes these clusters in terms of cluster size, three types of labels, and mean year of citees, i.e. cluster age. Of the selected clusters, Cluster 1 is the largest and oldest. In consideration with Cluster 6, results show that impact measure is still among the important themes in scientometrics. The third largest and newest group of literature is Cluster 10. It indicates practical applications of social media analytics to scientometrics is receiving the most recent attention. Other emerging topics include international collaboration (Cluster 16) and applications to medicine (Cluster 17) and environmental sciences and policy (Cluster 18).

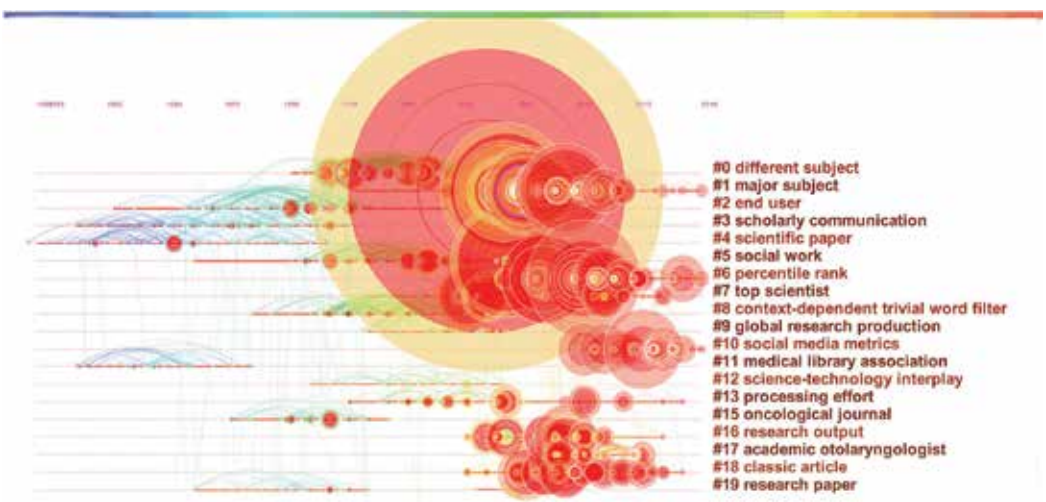


Figure 8. Timeline visualization with LLR cluster labels.

Cluster	Size	Age	Labels		
			LSA	LLR	MI
1	142	2007	h-Index	Major subject	Productivity incentive
6	74	2010	References	Percentile rank	Average citation
10	47	2013	Papers	Social media metrics	Practical application
16	34	2008	China	Processing effort	Worldwide research productivity
17	32	2011	Documents	Academic otolaryngologist	Peer-reviewed ophthalmology
18	30	2009	Water	Classic article	National policy intervention

Table 8. Cluster summary.

4. Discussion

4.1. Epistemological characteristics

The domain-level investigation revealed the following characteristics of published research in scientometrics. First, scientometrics research is multidisciplinary. Multiple disciplines such as "psychology, education, health" and "medicine, medical, clinical" are engaged in advancing knowledge in the domain. In particular, computer and information sciences had the largest influence on the emergence and development of scientific knowledge. The assignment of WoS categories also evidenced the multidisciplinarity of scientometrics as a variety of domains such as social sciences, engineering, medical and health sciences, and environmental sciences have contributed to the growth of the field. Second, scientometrics is not yet fully interdisciplinary as shown in the finding that research frontiers from "medicine, medical, clinical" largely cites from similar domains. Examining domain-level citation patterns in consideration with the WoS category assignment obtained a solid overview of the publication profile of the field. It revealed the growth of the domain by visualizing the distribution of citation trajectories at a disciplinary level, adding richer contexts with examining the distribution of WoS category assignment. Finally, most of the landmark articles were published relatively recently, namely after 2004 in spite of the long history of the domain. We argue that the domain's maturation is still ongoing.

4.2. Historic footprint and emerging technologies

The analysis of keywords, topic models, and document clusters identified the following thematic patterns in scientometrics research. In the beginning some researchers focused on employing citation analysis to measure the impact of a science. Another effort focused on the evaluation of performance and productivity of research, employing scientometrics approaches. The identification of patterns in scientific collaboration was also among the important themes. The other effort had interest in modeling scientometrics laws and proposing scientometric indicators and impact measures. Recently, applications of scientometrics

approaches to a variety of domains such as material sciences, medicine, and environmental sciences have received increasing attention. In reverse, practical applications of social media analytics to scientometrics is also receiving the most recent interest. Impact measure and science mapping are among the canonical research themes receiving consistent attention from the beginning of the domain.

5. Conclusion

The present chapter aimed to explore epistemological characteristics, historic areas of innovation, and emerging trends in scientometrics. We achieved this by investigating domain-level citation paths, WoS category assignment, keyword co-occurrence, temporal topic models, and document clusters. The findings indicate the domain of scientometrics is multidisciplinary and partially interdisciplinary. Social sciences and biomedicine have published to the field, but not yet cited from each other. We argue that the maturation of scientometrics as a scientific field is still ongoing. Next, early studies tried to measure a science's impact and performance and productivity of published research. Successive effort investigated laws and indicators in scientometrics and explored scientific collaboration. Recent literature is paying attention to topics such as applying scientometrics approaches to different domains and bringing social media analytics in scientometrics.

The approaches of the present study provide advantages in investigating intellectual structure of a science as follows. First, we tried to make our data collection inclusive by investigating closely neighboring domains. Conventional studies of domain analysis often cover only a fraction of published literature. Our method provides a systematic way to explore the broader coverage of a scientific discipline. Second, we investigated the domain from a multi-faceted point of view. Domain-level citation trajectories, subject category assignment, networks of subject categories and keywords, bursting keywords, topic models, and document co-citation networks were identified in this study. Sub-sections in Results triangulated each other, adding richer interpretations from macro units of analysis to micro ones. Finally, the analytical procedure and tools employed in the present work enabled us to explore time-aware research trends in the domain. In addition, one can conduct this kind of domain analysis of his or her concern as frequently as needed without prior knowledge or experience. Thus, the proposed approaches have a relatively higher reproducibility and lower cost for conducting studies at a larger scale, especially as in the era of mass publication.

There are several limitations in our work. First, the topic search we conducted on WoS may have missed relevant records. It is acknowledged that the vocabulary mismatch presents a challenge for keyword-based search. We may be able to overcome this drawback by employing citation indexing or iterative search query development as an alternative strategy in order to capture a much broader context. Second, WoS as our source of data may have underrepresented conference proceedings. It is also recognized as an issue for disciplines such as social sciences and arts and humanities [13]. At the time of data retrieval, the authors' institutes only subscribed to the core collection of WoS. Thus, it was inevitable not to miss some

relevant records accordingly. Additional sources such as Scopus are recommended for future refinements of this type of analysis. In addition, some findings or sub-sections in Results may seem too general to characterize emerging technologies in scientometrics when considered independently from the entire context. We argue that that is not because of the limitation of our approaches and tools but due to the characteristics of bibliographic records. That means textual fields that can be used only include titles, abstracts, and keywords which are often abstract to be inclusive. To overcome this, we employed not only frequency-based metrics such as citation counts and latent semantic analysis but also burst detection and probability-oriented techniques such as LLR, MI, and DTM. Then, we tried to triangulate the findings from each sub-section, adding richer interpretations as moving between different units of analysis. We argue that our approaches be more strengthened if we can have access to more informational sources such as full text. Finally, we selected 100 highly cited references to generate the intellectual landscapes. Although this data reduction is in part intuitive, we can strengthen our approach by choosing cited articles based on more refined indicators such as h-index or g-index. It may be worth conducting a separate study of the theoretical implications of using a variety of conceivable selection criteria. We also plan to apply the present chapter's approaches to much more comprehensive records that cover a various type of publication materials.

Conflict of interest

There are no conflicts of interest.

Author details

Meen Chul Kim[1]* and Yongjun Zhu[2]

*Address all correspondence to: meenchul.kim@drexel.edu

1 Drexel University, Philadelphia, PA, USA

2 Sungkyunkwan University, Seoul, South Korea

References

[1] Leydesdorff L, Milojević S. Scientometrics. In: International Encyclopedia of the Social & Behavioral Sciences. 2nd ed. Oxford, UK: Elsevier; 2015

[2] Wolfram D. Applied Informetrics for Information Retrieval Research. Westport, CT: Libraries Unlimited; 2003

[3] Björneborn L, Ingwersen P. Toward a basic framework for webometrics. Journal of the American Society for Information Science and Technology. 2004;**55**(14):1216-1227

[4] Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology. 2006;**57**(3):359-377

[5] Chen C, Ibekwe-SanJuan F, Hou J. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. Journal of the American Society for Information Science and Technology. 2010;**61**(7):1386-1409

[6] Chen C, Leydesdorff L. Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. Journal of the American Society for Information Science and Technology. 2014;**65**(2):334-351

[7] Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;**84**(2):523-538

[8] Kim MC, Zhu Y, Chen C. How are they different? A quantitative domain comparison of information visualization and data visualization (2000-2014). Scientometrics. 2016;**107**(1):123-165

[9] Zhu Y, Kim MC, Chen C. An investigation of the intellectual structure of opinion mining research. Information Research. 2017;**22**(1): paper 739. http://www.informationr.net/ir/22-1/paper739.html

[10] Waltman L, Eck V, NJ. A smart local moving algorithm for large-scale modularity-based community detection. European Physical Journal B. 2013;**86**(11):471

[11] Kleinberg J. Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery. 2003;**7**(4):373-397

[12] Blei DM, Lafferty J D. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 113-120

[13] Mongeon P, Paul-Hus A. The journal coverage of web of science and Scopus: A comparative analysis. Scientometrics. 2016;**106**(1):213-228

Chapter 3

Patterns of Academic Scientific Collaboration at a Distance: Evidence from Southern European Countries

Ana Fernández, Esther Ferrándiz and
M. Dolores León

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77370

Abstract

The main objective of this chapter is to examine the trends of academic scientific collaboration (SC) at a distance among public universities located in peripheral countries: Spain, Italy, Greece, and Portugal. The data to capture scientific collaboration consists of a set of co-authored articles published between 2001 and 2010 by universities located in the mentioned Southern countries, indexed by the Science Citation Index expanded (SCI Expanded) of the Information Sciences Institute (ISI) Web of Science (WoS) database. We link this data to institution-level information provided by the EUMIDA dataset. In addition, we retrieved regional data on economic variables from Eurostat. The methodology relies on a descriptive analysis of the evolution of co-publications at different notions of proximity. Our results show a trend toward collaboration over longer distances, although we find heterogeneity by countries and disciplines. Building on our results, we provide some policy implications.

Keywords: scientific collaboration (SC), co-authorship, proximity dimensions, geographical distance

1. Introduction

In the last decades, there has been an increasing trend toward scientific collaboration (SC) [1, 2]. Getting more insights about trends in scientific collaboration (SC) is important because SC is assumed to enhance the quality of the research for a number of stemming benefits largely discussed in the literature [3–5]. It brings together complementary knowledge and expertise. The presence of co-authors often implies a higher internal quality control than

single-authored papers; learning, social networks creation, knowledge diffusion, and cross-fertilization across individuals and/or disciplines are enhanced. From an economic viewpoint, SC also provides benefits including access to a wide variety of resources and new foundations or instruments. These benefits, together with the well-known role of knowledge creation and diffusion as the main sources for sustainable economic growth in the long run [6, 7], have shaped the European policy. The European government initiative aimed to convert Europe into the "the most competitive and dynamic knowledge-based economy" [8] giving priority to invest more in knowledge and innovation and to give Europe a new "fifth liberty," the free circulation of knowledge in order to construct a European research area [9].

The contribution of this research is twofold. First, we provide a comprehensive analysis of the evolution of geographical, cognitive, institutional, social, and organizational proximity on scientific collaboration. Apart from these, we also add economic distance as suggested in the recent literature [10, 11]. Second, we provide a joint analysis of trends in SC in all disciplines included in the Science Citation Index (SCI) of the Web of Science (WoS), and a separated analysis for *Chemistry & Chemical*, *Life Sciences* and *Physics and Astronomy* in order to examine whether there are differences across disciplines. We have chosen these disciplines because, jointly with *Medicine & Biomedicine*, they have the highest publication and collaboration share[1]. For our purpose, we use an original dataset containing information on 152,140 collaborations in publications in Science and Engineering (excluding social sciences) indexed in the Science Citation Index (SCI) provided by WoS and co-authored among academics from different universities. Our analysis includes 175 public universities from peripheral countries in Southern Europe: Spain, Greece, Italy, and Portugal. Focusing on peripheral countries is relevant because they usually include universities and regions far from core centers of knowledge with the lower level of resources and fewer opportunities to integrate in collaboration networks.

The remainder of the chapter is organized as follows. In Section 2, we review the relevant literature. Section 3 describes the data and explains the methodology. Section 4 provides the results. The main conclusions and policy implications are obtained at the end of the paper.

2. Literature review

The French school of proximity dynamics was pioneer to consider other notions of proximities beyond the geographical [12–14]. Drawing upon this line of research, Boschma [15], from a theoretical point of view, identified five kinds of proximities: geographical, cognitive, institutional, social, and organizational. Recent research has also highlighted the relevance of economic differences as an explanatory factor of SC [5, 10, 11, 16]:

- *Geographical distance* among actors hinders SC because face-to-face interactions that facilitate knowledge flows and tacit knowledge sharing become costly as distance increase

[1]Note that we do not perform a detailed analysis for *medicine & biomedicine* because some of the publications may be associated with university hospitals, which may have been or not co-authored by academics. Publications, for which we could not establish a clear link with an academic institution, have been excluded from our sample. Thus, our study may underestimate the scientific output in this discipline.

(e.g. [4, 17, 18]). Despite some authors claimed the death of distance due to ICT development, Hoekman et al. [18] found that physical distance still impedes research collaboration, with no evidence of a declining effect in the period 2000–2007.

- *Cognitive proximity*, that is, the degree of the shared knowledge base of organizations, facilitates knowledge transfer by contributing to building absorptive capacity that enables actors to identify, acquire, understand, and exploit knowledge available from others [19]. Nevertheless, recent studies have shown a certain degree of cognitive distance as a potential source of complementarities in order to improve knowledge base [20, 21]. Thus, the challenge is to collaborate with actors that provide access to heterogeneous sources of knowledge to generate sufficiently diverse complementarities, while ensuring the absorption capacity enabled by the shared knowledge base.

- *Institutional proximity* is defined by the degree of similarity in formal institutions, such as laws and rules, and informal institutions, like culture norms and habits, may enable knowledge flows by facilitating trust and reducing uncertainty and risks [15, 22]. Hoekman et al. [18] found that SC is more likely to occur within the same sub-national region, within the same country, and within the same linguistic area. Hennemann et al. [23] look in detail at the spatial structures of scientific activity (epistemic communities) showing that intra-country collaboration is more likely to occur than international collaboration.

- *Social proximity*, that is, socially embedded relations based on friendship, kindship and past experience between agents at the micro-level, is expected to stimulate interactive learning due to the trust and commitment [15]. It is commonly accepted to measure social proximity based on prior collaborations or previous research experiences [24–26].

- *Organizational proximity* can be understood as a variable capturing organization that share the same or similar regulation and routines at a micro-level. In that sense, a certain degree of organizational proximity is desirable to reduce uncertainty and opportunism in knowledge creation within and between organizations. In research collaboration literature, this dimension has been often included by a variable capturing whether partners to the same institutional arrange, for example, by belonging to the same corporation [27]. In this research, difficulties to consider organizational proximity in Boschma's sense, arises due to the absence of hierarchical relations among universities. However, they cannot be considered homogenous organizations because research institutions differ in their norms, structure, size, and strategy [28, 29].

- *Economic distance* (differences in economic resources among geographic areas) may determine the spatial patterns in SC, as derived from the center-periphery hypothesis applied to research collaboration [10, 11]. According to this literature, scientists in peripheral countries are willing to collaborate with core countries to gain access to resources, while core areas seek for complementarities [16]. However, empirical evidence provided by Acosta et al. [10] using data on a sample of co-authored papers among regions in EU-15 showed that differences in per capita income do not affect collaboration, while having similar levels of resources devoted to R&D play a positive role. They argue that having access to greater resources increase opportunities for mobility and attendance to international conferences, which enables establishing and reinforcing personal contacts for future collaborations.

3. Methodology and data

The empirical data used in this chapter consists of a set of 152,140 collaborations by scientists affiliated to different universities and published in journals indexed by the Science Citation Index Expanded (SCI Expanded) provided by the Thomson Reuters Web of Science (WoS). Socio-economic and humanities disciplines are excluded from our analysis. Our period of analysis is 2001–2010. This dataset was built following a similar procedure to Acosta et al. [10, 30]. Since our focus is at the university level, we had to harmonize the name variations of universities, mainly stemming from the use of the native versus the English name or the use of different acronyms. Then, papers were assigned to universities following the full counting process (crediting one publication to each co-author institution). Next, data on academic collaboration was placed into a symmetrical matrix containing all co-publications between university *i* and university *j* and, therefore, excluded intra-university collaboration. Publications where classified into 12 scientific disciplines following the Centre for Science and Technology Studies (CWTS) classification, using again the full counting method for those publications included in journals related to more than one discipline.

In a further step, we matched this dataset with EUMIDA dataset (Data Collection 1) in order to get information about organizational characteristics of the universities. EUMIDA data is the result of an initiative of the European commission to provide a complete census of European universities and provides information at the university level including organizational details such as education offered and staff employed[2]. Our final sample includes only those universities that were present in both datasets, that is, 175. Consequently, there are potentially (175*174) ÷ 2 = 15,225 collaboration links (observations). Additional information about regional Gross Domestic Product (GDP) and R&D expenditures was extracted from Eurostat.

In order to estimate the influence of different proximity dimensions on university SC, we put forward several variables:

- Geographical distance (*Geodist*) is measured as the Euclidean distance between universities *i* and *j*.
- Cognitive distance (*Cogndist*) is captured as the correlation index calculated as Paci and Usai [31] for the 12 discipline composition of scientific papers in university *i* and university *j* for the period 2001–2005. This coefficient ranges between zero (minimum distance, identical specialization) and one (maximum distance).
- Institutional proximity is measured by two binary variables. *Region* is a dummy variable, which takes value 1 when universities *i* and *j* are in the same region, 0 otherwise. *Country* is a dummy variable, which takes value 1 when universities i and j are in the same country, 0 otherwise.

[2]A description of data and the collection procedure is provided in EUMIDA 2010. Feasibility Study for Creating a European University Data Collection [Contract No. RTD/C/C4/2009/0233402]. Data collection 1 is available at http://ec.europa.eu/research/era/areas/universities/universities_en.htm. (Accessed at 18/10/2012). Data collection 2, which contains more detailed data, was not available to us by the time of this research.

- Social proximity (*Socialprox*) is represented by a dummy variable which takes value 1 if universities "i" and "j" have collaborated for the five-years previous period 2001–2005. However, this indicator does not allow us to provide evidence on trends in social distance since we did not have data on previous collaborations for the period 2001–2005.
- Organizational proximity is captured by two variables. *Educprox* is the correlation coefficient between the nine education fields, as identified in EUMIDA, corresponding to university *i* and university *j*. *Staffdist* is the absolute difference in total staff of universities *i* and *j*. These variables refer to year 2008, which is the reference year for EUMIDA dataset.
- Economic distance is measured by three variables. *GDPdist* is the absolute difference in the average GDP in 2004–2008 between regions, where universities *i* and *j* are located. *R&Ddist* is calculated similarly but using the absolute difference in higher education R&D expenditures as % of the GDP. *Convergence* is a dummy variable that equals one if the two universities are located in convergence regions; zero otherwise.

Note that the description of the variables refers to data for all 12-disciplines. For separated descriptive by disciplines, collaborations, and previous collaborations refer to the respective counts for that specific discipline. At the discipline level, $Cogndist_{ij}$ represents the dissimilarity in specialization in a certain discipline. Since it is not possible to calculate it as a correlation coefficient or Paci and Usai index [31], it was calculated following a different procedure for models by disciplines: first, we calculated for each university the share of publications in each discipline over its total number of publications; second, we obtained the absolute difference in this indicator for each pair of universities.

It is worth noting that organizational proximity measures attempt to capture a complex phenomenon difficult to measure. Then, we choose the differences in educational profiles and size as factors capturing organizational characteristics that may shape their culture or orientation. In addition, we did not have access to data on R&D funding information at the level of institutions, so we have included the amount of R&D expenditures in the region in which the university is located.

In order to identify trends in scientific collaboration, we calculate the descriptives of distances for those pairs collaborating during 2001–2005 and, then, for those pairs collaborating during 2006–2010. **Table 1** shows some descriptives on collaborations in our sample: the number

	01–05	**06–10**
Pairs (a)	15,225	15,225
Collaborating pairs (b)	3669	4775
Total Collaborations (c)	60,522	91,618
b/a	24.10%	31.36%
Collaboration intensity (c/b).	16.50	19.19

Source: ISI Web of Science. Own elaboration.

Table 1. Number of collaborations and collaboration intensity 2001–2010.

of collaborating pairs has increased from 3669 to 4775 and total collaboration has substantially increased by 51.38%. From all possible pairs of universities, 24.10% has collaboration in 2001–2005, while it increases to 31.36% in 2006–2010. The intensity of collaboration (number of average collaborations among pairs) has increased from 16.50 to 19.19.

4. Results

In order to analyze the evolution of collaboration across distance, we obtain the mean and standard deviation of each proximity dimension in the period 2001–2005 and 2006–2010 for Spain, Greece, Italy, and Portugal. **Table 2** displays the results including data for all disciplines. **Tables 3–5** show the descriptives for *Chemistry & Chemical*, *Life Sciences* and *Physics & Astronomy*, respectively.

The following conclusions are drawn from the results of **Table 2**:

- The average geographical distance among partners has increased over time by 9.42%. This result holds for the four countries in our sample, with an increase ranging from 2.74% in Portugal to 9.53% in Italy. Therefore, we can identify a strong pattern of increasing collaboration among universities throughout longer geographical distance.
- The mean cognitive distance has slightly increased (0.65% on average for peripheral countries), suggesting a trend toward collaboration with universities specialized in different fields of research.
- The coefficients of *region and country* strongly decrease over the period of analysis (by 15.30 and 10.74%, respectively), suggesting that institutional proximity decays over time. Thus, there is a trend toward interregional and international collaboration. Focusing on detailed country data, Spain shows the strongest decrease in intra-regional collaboration and intra-national collaboration (25.12–15.49%, respectively). It is also remarkable that Portugal, despite showing a similar decrease in intra-regional collaboration, displays a smaller decrease in intra-national collaboration, and suggesting differences in international openness across these countries.
- The coefficients of variables capturing similarities in educational profile and differences in size decrease in the period 2006–2010 by 1.88–3.14%, respectively. Based on these results, we cannot distinguish a clear trend in organizational distance. Country data shows that similarities in educational profile decrease in Greece, Italy, and Portugal but increase in Spain. Differences in size decrease in all countries, with the exception of Portugal.
- The co-efficient of *GDPdist* remains almost steady for peripheral countries as a whole, while differences in R&D slightly decrease over time. However, there are differences by countries: *GDPdist* increases in all countries but in Greece, where it decreases by 6.35%; and *R&Ddist* increases in Spain, Greece, and Portugal, but decreases in Italy. When focusing in collaboration among convergence regions, it arises that universities tend to collaborate more over time with other universities also located in convergence regions, suggesting that economic distance is increasing its importance as a barrier to SC. Italy is the country, where collaboration among convergence regions experienced the strongest increase (by 5.41%).

	All Countries			ES			GR			IT			PT		
	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%
Geographical distance															
$Geodist_{ij}$	8.6937	9.5126	9.42	10.4664	11.2584	7.57	14.2690	15.2108	6.59	9.8263	10.7625	9.53	11.0458	11.3480	2.74
	(7.34)	(7.84)		(7.21)	(7.59)		(10.81)	(11.14)		(7.16)	(7.36)		(9.55)	(9.50)	
Cognitive distance															
$Cogndist_{ij}$	0.7817	0.7868	0.65	0.7834	0.7869	0.45	0.7970	0.8034	0.80	0.7769	0.7819	0.64	0.7898	0.7962	0.81
	(0.05)	(0.05)		(0.04)	(0.04)		(0.05)	(0.06)		(0.05)	(0.05)		(0.04)	(0.04)	
Institutional distance															
$Region_{ij}$	0.0621	0.0526	−15.30	0.0430	0.0322	−25.12	0.0638	0.0581	−8.93	0.0353	0.0289	−18.13	0.0760	0.0569	−25.13
	(0.24)	(0.22)		(0.20)	(0.18)		(0.24)	(0.23)		(0.18)	(0.17)		(0.27)	(0.23)	
$Country_{ij}$	0.6173	0.5510	−10.70	0.4313	0.3645	−15.49	0.3161	0.2940	−6.99	0.5330	0.4642	−12.91	0.2496	0.2267	−9.17
	(0.49)	(0.50)		(0.50)	(0.48)		(0.47)	(0.46)		(0.50)	(0.50)		(0.43)	(0.42)	
Social proximity															
$Socialprox_{ij}$	—	0.6848	—	—	0.6554	—	—	0.4513	—	—	0.6876	—	—	0.5248	—
		(0.46)			(0.48)			(0.50)			(0.46)			(0.50)	
Organizational proximity															
$Educprox_{ij}$	5.7874	5.6786	−1.88	6.1203	6.0237	4.64	3.8997	3.7790	−3.10	5.7466	5.6445	−1.78	6.3322	6.2153	−1.85
	(1.80)	(1.82)		(1.57)	(1.61)		(2.16)	(2.04)		(1.80)	(1.82)		(1.30)	(1.41)	
$Staffdist_{ij}$	1939.01	1878.22	−3.14	2037.58	1947.88	−1.58	1767.02	1744.74	−1.26	2024.37	1985.35	−1.93	1620.53	1644.70	1.49
	(1676.61)	(1631.56)		(1717.26)	(1650.28)		(1621.03)	(1579.96)		(1734.05)	(1692.64)		(1399.37)	(1414.75)	
Economic distance															
$GDPdist_{ij}$	0.0062	0.0063	1.61	0.0060	0.0061	1.67	0.0063	0.0059	−6.35	0.0066	0.0068	3.03	0.0066	0.0067	1.52
	(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)	

	All Countries			ES			GR			IT			PT		
	01–05	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**
$R\&Ddist_{ij}$	0.1092	0.1086	−0.55	0.0957	0.0969	1.25	0.1064	0.1091	2.54	0.1340	0.1336	−0.30	0.0854	0.0855	0.12
	(0.09)	(0.09)		(0.07)	(0.07)		(0.09)	(0.09)		(0.09)	(0.09)		(0.07)	(0.08)	
$Convergence_{ij}$	0.4044	0.4241	4.87	0.4126	0.4227	2.45	0.4800	0.4972	3.58	0.3792	0.3997	5.41	0.5091	0.5316	4.42
	(0.05)	(0.49)		(0.49)	(0.49)		(0.50)	(0.50)		(0.49)	(0.49)		(0.50)	(0.50)	
N°. Obs.	3669	4775	30.14	1929	2612	35.41	329	534	62.31	2210	2807	27.01	605	966	59.67

Source: ISI Web of Science, EUMIDA and Eurostat. Own elaboration.

Table 2. Change in average distance of collaborations per country. About 12 disciplines (Mean and standard deviation).

http://dx.doi.org/10.5772/intechopen.77370

	All Countries			ES			GR			IT			PT		
	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%
Geographical distance															
$Geodist_{ij}$	6.8711	7.9993	16.42	8.9700	9.7482	8.68	13.0388	15.2434	16.91	7.5760	8.9951	18.73	8.3953	9.8989	17.91
	(6.30)	(7.04)		(6.80)	(7.06)		(11.23)	(11.54)		(6.30)	(6.85)		(8.27)	(8.78)	
Cognitive distance															
$Cogndist_{ij}$	0.0924	0.0988	0.07	0.1133	0.1166	2.91	0.0821	0.0847	3.17	0.0833	0.0894	7.32	0.0926	0.0991	7.02
	(0.08)	(0.08)		(0.09)	(0.09)		(0.07)	(0.07)		(0.07)	(0.07)		(0.07)	(0.07)	
Institutional distance															
$Region_{ij}$	0.0879	0.0694	−21.05	0.0808	0.0557	−31.06	0.0575	0.0513	−10.78	0.0549	0.0413	−24.77	0.1009	0.0710	−26.63
	(0.28)	(0.25)		(0.27)	(0.23)		(0.23)	(0.22)		(0.23)	(0.20)		(0.30)	(0.26)	
$Country_{ij}$	0.7357	0.6465	−12.12	0.5271	0.4517	−14.30	0.4138	0.3282	−20.69	0.6874	0.5810	−15.48	0.3571	0.2759	−22.74
	(0.44)	(0.48)		(0.50)	(0.50)		(0.50)	(0.47)		(0.46)	(0.50)		(0.48)	(0.45)	
Social proximity															
$Socialprox_{ij}$	—	0.6439	—	—	0.5842	—	—	0.4462	—	—	0.6723	—	—	0.4828	—
		(0.48)			(0.49)			(0.50)			(0.47)			(0.50)	
Organizational proximity															
$Educprox_{ij}$	6.1203	5.9715	−2.43	6.4282	6.2213	−3.22	4.2184	3.8615	−8.46	6.0781	5.9531	−2.06	6.4664	6.3732	−1.44
	(1.61)	(1.69)		1.32	(1.51)		(2.21)	(2.16)		(1.64)	(1.70)		(1.09)	(1.21)	
$Staffdist_{ij}$	1973.42	1933.14	−2.04	1998.80	1981.83	−0.85	1883.60	1822.17	−3.26	2033.37	2012.00	−1.05	1531.05	1569.63	2.52
	(1661.11)	(1646.25)		(1622.16)	(1625.04)		(1767.60)	(1675.70)		(1723.08)	(1721.78)		(1199.39)	(1318.44)	
Economic distance															
$GDPdist_{ij}$	0.0059	0.0061	3.39	0.0056	0.0058	3.57	0.0059	0.0057	−3.39	0.0062	0.0065	4.84	0.0064	0.0065	1.56
	(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)	

	All Countries			ES			GR			IT			PT		
	01–05	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**
$R\&Ddist_{ij}$	0.1081	0.1079	−0.19	0.0921	0.0945	2.61	0.1030	0.0999	−3.01	0.1367	0.1373	0.44	0.0738	0.0768	4.07
	(0.09)	(0.09)		(0.08)	(0.08)		(0.08)	(0.08)		(0.09)	(0.09)		(0.07)	(0.07)	
$Convergence_{ij}$	0.3776	0.4004	6.04	0.3945	0.4094	3.78	0.4353	0.4740	8.89	0.3498	0.3734	6.75	0.5210	0.4970	−4.61
	(0.48)	(0.049)		(0.49)	0.49		(0.50)	(0.50)		(0.48)	(0.48)		(0.50)	(0.50)	
N°. Obs.	1763	2738	55.30	829	1419	71.17	87	195	124.14	1075	1599	48.74	238	493	107.14

Source: ISI Web of Science, EUMIDA and Eurostat. Own elaboration.

Table 3. Change in average distance of collaborations per country. Chemistry and chemical (Mean and standard deviation).

	All Countries			ES			GR			IT			PT		
	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%
Geographical distance															
$Geodist_{ij}$	7.1094	7.5915	6.78	9.1112	9.4409	3.62	13.8785	13.7744	−0.75	7.5194	8.2657	9.92	10.5613	9.9425	−5.86
	(6.78)	(6.87)		(6.88)	(7.04)		(11.86)	(11.30)		(6.65)	(6.68)		(10.31)	(8.88)	
Cognitive distance															
$Cogndist_{ij}$	0.0826	0.0905	9.56	0.0856	0.0924	7.94	0.1051	0.1079	2.66	0.0716	0.0788	10.06	0.1357	0.1285	−5.31
	(0.08)	(0.09)		(0.07)	(0.08)		(0.10)	(0.10)		(0.08)	(0.08)		(0.11)	(0.10)	
Institutional distance															
$Region_{ij}$	0.0817	0.0705	−13.71	0.0796	0.0585	−26.51	0.0769	0.0643	−16.38	0.0500	0.0448	−10.40	0.0803	0.0628	−21.79
	(0.27)	(0.26)		(0.27)	(0.23)		(0.27)	(0.25)		(0.21)	(0.21)		(0.27)	(0.24)	
$Country_{ij}$	0.7427	0.6786	−8.63	0.5361	0.4712	−12.11	0.3846	0.3684	−4.21	0.7116	0.6441	−9.49	0.3092	0.2601	−15.88
	(0.44)	(0.47)		(0.50)	(0.50)		(0.49)	(0.48)		(0.45)	(0.48)		(0.46)	(0.44)	
Social proximity															
$Socialprox_{ij}$	—	0.5828	—	—	0.4901	—	—	0.3275	—	—	0.6159	—	—	0.4081	—
		(0.49)			(0.50)			(0.47)			(0.49)			(0.50)	
Organizational proximity															
$Educprox_{ij}$	6.2551	6.1164	−2.22	6.5286	6.4119	−1.79	4.2596	4.1813	−1.84	6.2348	6.0814	−2.46	6.6426	6.5381	−1.57
	(1.43)	(1.53)		(1.27)	(1.30)		(2.16)	(2.07)		(1.32)	(1.50)		(0.95)	(1.07)	
$Staffdist_{ij}$	2023.76	1913.63	−5.44	2067.18	1964.94	−4.95	1708.24	1679.84	−1.66	2106.67	2008.19	−4.67	1673.21	1573.48	−5.96
	(1712.73)	(1634.33)		(1686.91)	(1635.40)		(1572.05)	(1585.78)		(1781.08)	(1696.95)		(1451.14)	(1325.04)	
Economic distance															
$GDPdist_{ij}$	0.0059	0.0060	1.69	0.0057	0.0059	3.51	0.0059	0.0058	−1.69	0.0065	0.0065	0	0.0063	0.0064	1.59
	(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)	

http://dx.doi.org/10.5772/intechopen.77370

	All Countries			ES			GR			IT			PT		
	01–05	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**
$R\&Ddist_{ij}$	0.1076	0.1066	−0.93	0.0882	0.0896	1.59	0.1026	0.1047	2.05	0.1341	0.1360	1.42	0.0864	0.0821	−4.98
	(0.09)	(0.09)		(0.07)	(0.07)		(0.09)	(0.09)		(0.09)	(0.10)		(0.08)	(0.07)	
$Convergence_{ij}$	0.3773	0.3991	5.78	0.3881	0.4226	8.89	0.5146	0.5146	0	0.3616	0.3703	2.41	0.5100	0.5269	3.31
	(0.48)	(0.49)		(0.49)	(0.49)		(0.50)	(0.50)		(0.48)	(0.48)		(0.50)	(0.50)	
N°. Obs.	1811	2483	37.11	804	1242	54.48	104	171	64.42	1120	1450	29.46	249	446	79.12

Source: ISI Web of Science, EUMIDA and Eurostat. Own elaboration.

Table 4. Change in average distance of collaborations per country. Life sciences (Mean and standard deviation).

	All Countries			ES			GR			IT			PT		
	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%	01–05	06–10	%
Geographical distance															
$Geodist_{ij}$	7.3162	8.0470	9.99	9.5294	10.2536	7.60	14.1949	15.3879	8.40	8.0791	8.9452	10.72	9.4753	9.8693	4.16
	(6.49)	(7.28)		(6.90)	(7.61)		(10.52)	(11.25)		(6.36)	(7.01)		(8.74)	(9.02)	
Cognitive distance															
$Cogndist_{ij}$	0.1065	0.1047	−1.69	0.0888	0.0877	−1.24	0.0860	0.0938	9.07	0.1193	0.1178	−1.26	0.0745	0.0732	−1.74
	(0.93)	(0.09)		(0.07)	(0.07)		(0.07)	(0.07)		(0.11)	(0.10)		(0.06)	(0.06)	
Institutional distance															
$Region_{ij}$	0.0875	0.0820	−6.29	0.0769	0.0651	−15.34	0.0729	0.0571	−21.67	0.0495	0.0480	−3.03	0.1256	0.1022	−18.63
	(0.28)	(0.27)		(0.27)	(0.25)		(0.26)	(0.23)		(0.22)	(0.21)		(0.33)	(0.30)	
$Country_{ij}$	0.7035	0.6534	−7.12	0.4842	0.4433	−8.45	0.2813	0.3029	7.68	0.6430	0.5886	−8.46	0.3478	0.2972	−14.55
	(0.46)	(0.48)		(0.50)	(0.50)		(0.45)	(0.46)		(0.48)	(0.49)		(0.48)	(0.46)	
Social proximity															
$Socialprox_{ij}$	—	0.6168	—	—	0.5409	—	—	0.3543	—	—	0.6554	—	—	0.4149	—
		(0.48)			(0.50)			(0.48)			(0.48)			(0.49)	
Organizational proximity															
$Educprox_{ij}$	5.8197	5.7896	−0.52	6.0315	6.0474	0.26	4.1770	4.2229	1.10	5.7806	5.7815	0.02	6.2512	6.1920	−0.85
	(1.83)	(1.82)		(1.65)	(1.61)		(2.05)	(2.02)		(1.90)	(1.88)		(1.28)	(1.33)	
$Staffdist_{ij}$	2042.42	2007.24	−1.72	2107.02	2075.72	−1.49	1856.37	1772.30	−4.53	2126.58	2074.59	−2.44	1529.81	1584.70	3.59
	(1683.19)	(1665.54)		(1663.22)	(1643.10)		(1682.12)	(1572.16)		(1763.80)	(1725.96)		(1139.38)	(1387.78)	
Economic distance															
$GDPdist_{ij}$	0.0059	0.0061	3.39	0.0058	0.0059	1.72	0.0062	0.0063	1.61	0.0063	0.0065	3.17	0.0062	0.0067	8.06
	(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.00)		(0.00)	(0.01)	

http://dx.doi.org/10.5772/intechopen.77370

	All Countries			ES			GR			IT			PT		
	01–05	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**	**01–05**	**06–10**	**%**
$R\&Ddist_{ij}$	0.1095	0.1066	−2.65	0.0931	0.0939	0.86	0.0962	0.1076	11.85	0.1350	0.1312	−2.81	0.0786	0.0724	−7.89
	(0.09)	(0.09)		(0.08)	(0.08)		(0.7)	(0.09)		(0.10)	(0.09)		(0.07)	(0.07)	
$Convergence_{ij}$	0.3469	0.3734	7.64	0.3506	0.3727	6.30	0.4681	0.5202	11.13	0.3282	0.3468	5.67	0.4493	0.5046	12.31
	(0.48)	(0.48)		(0.48)	(0.48)		(0.50)	(0.50)		(0.47)	(0.48)		(0.50)	(0.50)	
N°. Obs.	1703	2158	26.72	793	1076	35.69	96	175	82.29	1112	1332	19.78	207	323	56.03

Source: ISI Web of Science, EUMIDA and Eurostat. Own elaboration.

Table 5. Change in average distance of collaborations per country. Physics & astronomy (Mean and standard deviation).

Next, we check if these results hold for *Chemistry & Chemical*, *Life Sciences* and *Physics & Astronomy* when analyzed separately (**Tables 3–5**). As shown by **Table 3**, geographical distance in *Chemistry & Chemical* increases, with a growth rate ranging from 8.68% in Spain to 18.73% in Italy. Specialization distance also rises, from 2.91% in Spain to 7.32% in Italy. Institutional proximity, that is, collaboration among universities located in the same region/nation decreases over time in peripheral countries. Organizational distance, as measured by differences in staff decreases over time, with the exception of Portugal (where it increases by 2.52%). However, it also comes that there is a trend toward collaboration between universities with different educational profiles, suggesting that organizational distance is increasing throughout the period of analysis. Generally, economic distance in terms of the difference in GDP and R&D expenditures among partners for each country also increases along time, excepting in Greece where it decreases. Our results also show a trend toward collaboration among convergence regions (average growth of 6.04%), except in Portugal.

Table 4 displays the evolution of distance dimensions for *Life Sciences*. Geographical distance shows contradictory results by countries, with an increase in average collaboration distance in Spain (3.62%) and Italy (9.92%), but a decrease in average distance in Greece and Portugal (0.75 and 5.86%, respectively), specialization distance also increases in Spain (7.94%), Greece (2.66%) and Italy (10.06%). Again, Portugal shows a decrease in average distance (5.31%). Variables capturing institutional proximity show that there is a decrease in intra-regional collaboration and national collaboration, in favor to interregional and international collaboration. In contrast, organizational distance is decreasing over time, with an average decrease of 2.22% in education proximity, 5.44% differences in size. This is universities tend to collaborate more over time with other universities with similar institutional characteristics. Economic distance yields different results for each indicator. GDP distance increases in collaboration pairs in Spain and Portugal but a decrease in Greece. Focusing on R&D distance, there is an increase in Spain (1.59%), Greece (2.05%), and Italy (1.42%) and a decrease in Portugal (4.98%). There is also a trend toward collaboration between convergence regions in Spain, Italy, and Portugal, while Greece remains equal.

Table 5 shows that geographical distance in collaboration in *Physics & Astronomy* increases in peripheral countries, ranging from an increase of 4.16% in Portugal to 10.72% in Italy. Specialization distance decreases by 1.69% in peripheral countries, being Greece an exception (with an increase of 9.07%). Generally speaking, institutional proximity strongly decreases over time, with an increasing share of international and interregional collaboration. The results show that universities tend to collaborate more and more with other universities with similar educational profiles, with the exception of Portugal where it slightly decreases by 0.85%. Distance in size decreases by 1.72% in the whole sample, with the exception of Portugal, where it increases by 3.59%. Economic distance in GDP also increases along time, from 1.61% in Greece to 8.06% in Portugal. Distance in R&D shows different growth rates across countries. It increases in Spain and Greece and decreases in Italy and Portugal. Our results also show a strong trend (7.64%) toward collaboration among convergence regions in all peripheral countries.

5. Conclusions

The objective of this chapter was to analyze patterns of SC along different notions of proximity in the period 2001–2010. For this purpose, we use data on 152,140 collaborations in publications in Science and Engineering (excluding social sciences) indexed in the Science Citation Index (SCI) provided by the ISI Web of Science (WoS) and co-authored among academics from different universities. Our analysis includes 175 public universities from peripheral countries in Southern Europe: Spain, Greece, Italy, and Portugal. The methodology relies on a descriptive analysis of collaborations in 12 scientific fields in which publications in science and engineering can be classified. In addition, we also provide descriptives for *Chemistry & Chemical*, *Life Sciences* and *Physics & Astronomy*, which are among the disciplines with the highest rate of collaboration over publications.

Our results for the whole sample and also for each country and discipline show that there is a clear trend toward collaboration along the greater geographical distance in peripheral countries. This result is in line with the finding obtained by Hoekman et al. [18] for 33 European Countries. There is also a trend toward increasing collaboration across cognitive and institutional distances. We cannot obtain clear conclusions for the evolution of organizational distance since we obtain controversial results for each of the indicators that measure this notion. Besides, our data reveals a trend toward collaboration among convergence regions, an increase in collaboration across larger economic distance in terms of GDP differences, but the opposite result is obtained in terms of R&D differences.

From a policy viewpoint, we can make some contributions. First, despite we find some heterogeneity in the results by scientific fields and countries, general patterns described in this chapter suggest a decrease in the importance of distance as a barrier to scientific collaboration in peripheral countries. Therefore, this evidence for peripheral countries suggests that there has been an advance in the construction of a European Research Area, as pursued by the EU policy. However, differences across countries and disciplines in the evolution of distance in collaborations suggest the convenience of elaborating tailor-made EU research policies adapted to their specific needs[3]. For example, for the model for all disciplines (**Table 2**), it is clear that although Portugal is collaborating across larger geographical distance (2.74%), it is lagging behind the rest of countries in our sample (Spain 7.57%, Greece 6.59%, and Italy 9.53%). Then, Portugal might benefit from policies oriented toward promoting the creation and diffusion of knowledge in collaboration across universities located at a distance. By doing so, it could catch up with the rest of peripheral countries. A similar analysis for the evolution of the rest of proximity notions could serve as a guide to elaborate EU policies for peripheral countries.

This study has four main limitations. First, we cannot provide evidence on trends in social distance since we did not have data on previous collaborations for the period 2001–2005. Second, we formatted our data as a cross-sectional series and measured variables at a unique time reference for the two periods, so we are not able to provide yearly statistics

[3]As pointed out by Hoekman et al. [18] it may be that each scientific discipline has different requirements due to their research topics or needed infrastructures.

on collaboration and different notions of proximity. Third, we do not control for scientific quality of universities that may be a factor affecting scientific collaboration patterns (see Hoekman et al. [18]). Fourth, our results must be taken with caution because we do not consider all countries in EU, but only peripheral countries in Southern Europe: Spain, Greece, Italy, and Portugal. Thus, future research may aim at providing evidence on collaboration across all EU countries, which may serve to extract policy implications on a wider framework.

Author details

Ana Fernández, Esther Ferrándiz* and M. Dolores León

*Address all correspondence to: esther.ferrandiz@uca.es

Department of Economics, Facultad de CC. EE. y Empresariales de Cádiz, University of Cadiz, Spain

References

[1] Gazni A, Sugimoto CR, Didegah F. Mapping world scientific collaboration: Authors, institutions, and countries. Journal of the American Society for Information Science and Technology. 2012;**63**(2):323-335

[2] Waltman L, Tijssen RJW, Eck NJV. Globalisation of science in kilometres. Journal of Informetrics. 2011;**5**(4):574-582

[3] Franceschet M, Costantini A. The effect of scholar collaboration on impact and quality of academic papers. Journal of Informetrics. 2010;**4**(4):540-553

[4] Katz J. Geographical proximity and scientific collaboration. Scientometrics. 1994;**31**(1):31-43

[5] Sonnenwald DH. Scientific collaboration. Annual Review of Information Science and Technology. 2007;**41**(1):643-681

[6] Foray D. Economics of Knowledge. MIT Press; 2004

[7] Romer P. Endogenous technological change. Journal of Political Economy. 1990;**98**(5): 71-102

[8] European Commission. Social Policy Agenda, Communication from the European Commission, COM (2000) 379 Final. Brussels; 2000a

[9] European Commission. Towards an European Research Area, Communication from the European Commission, COM (2000) 6 Final. Brussels: EC; 2000b

[10] Acosta M, Coronado D, Ferrándiz E, León MD. Factors affecting inter-regional academic scientific collaboration within Europe: The role of economic distance. Scientometrics. 2011;**87**(1):63-74

[11] Fernández A, Ferrándiz E, León MD. Proximity dimensions and scientific collaboration among academic institutions in Europe: The closer, the better. Scientometrics. 2016;**106**: 1073-1092

[12] Carrincazeaux C, Lung Y, Vicente J. The scientific trajectory of the French School of Proximity: Interaction- and institution-based approaches to regional innovation systems. European Planning Studies. 2008;**16**(5):617-628

[13] Rallet A, Torre A. Is geographical proximity necessary in the innovation networks in the era of global economy? GeoJournal. 1999;**49**(4):373-380

[14] Torre A, Gilly JP. On the analytical dimension of proximity dynamics. Regional Studies. 2000;**34**(2):169-180

[15] Boschma R. Proximity and innovation: A critical assessment. Regional Studies. 2005;**39**(1): 61-74

[16] Hwang K. International collaboration in multilayered center-periphery in the globalization of science and technology. Science Technology Human Values. 2008;**33**(1):101-133

[17] Hoekman J, Frenken K, van Oort F. The geography of collaborative knowledge production in Europe. Annals of Regional Science. 2009;**43**:721-738

[18] Hoekman J, Frenken K, Tijssen RJW. Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. Research Policy. 2010;**39**(5):662-673

[19] Cohen WM, Levinthal DA. Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly. 1990;**35**(1):128-152

[20] Nooteboom B, van Haverbeke W, Duysters G, Gilsing V, Van den Oord A. Optimal cognitive distance and absorptive capacity. Research Policy. 2007;**36**(7):1016-1034

[21] Gilsing V, Nooteboom B, Vanhaverbeke W, Duysters G, van den Oord A. Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. Research Policy. 2008;**37**(10):1717-1731

[22] Boschma R, Frenken K. Some notes on institutions in evolutionary economic geography. Economic Geography. 2009;**85**(2):151-158

[23] Hennemann S, Rybski D, Liefner I. The myth of global science collaboration-collaboration patterns in epistemic communities. Journal of Informetrics. 2012;**6**(2):217-225

[24] Hoekman J, Scherngell T, Frenken K, Tijssen R. Acquisition of European research funds and its effect on international scientific collaboration. Journal of Economic Geography. 2012;**13**:23-52

[25] Hong W, Su Y-S. The effect of institutional proximity in non-local university–industry collaborations: An analysis based on Chinese patent data. Research Policy. 2013;**42**(2): 454-464

[26] Petruzzelli AM. The impact of technological relatedness, prior ties, and geographical distance on university–industry collaborations: A joint-patent analysis. Technovation. 2011;**31**(7):309-319

[27] Balland PA. Proximity and the evolution of collaboration networks: Evidence from research and development projects within the global navigation satellite system (GNSS) industry. Regional Studies. 2011;**46**(6):741-756

[28] Cummings JN, Kiesler S. Coordination costs and project outcomes in multi-university collaborations. Research Policy. 2007;**36**(10):1620-1634

[29] Mowery D, Sampat B. Universities in National Innovation Systems. In: Fagerberg J, Mowery D, Nelson R, editors. The Oxford Handbook of Innovation. Oxford: Oxford University Press; 2004. pp. 209-239

[30] Acosta M, Coronado D, Ferrándiz E, León MD. Regional scientific production and specialization in Europe: The role of HERD. European Planning Studies. 2014;**22**:949-974

[31] Paci R, Usai S. Knowledge flows across European regions. The Annals of Regional Science. 2009;**43**:669-690

Chapter 4

Mapping a Research Field: Analyzing the Research Fronts in an Emerging Discipline

Gerardo Tibaná-Herrera,
María Teresa Fernández-Bajón and
Félix de Moya-Anegón

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76731

Abstract

The mapping overlay technique described in the scientific literature to analyze scientific domains must be complemented with procedures to identify and analyze the research fronts included in the cognitive structure of the represented domain. One possibility is the use of wordcloud maps to visually represent the cognitive structure of a discipline in any thematic domain, taking advantage of its capacity for abstraction and impact on the audience to stimulate new research processes. The case described in this chapter proposes an analysis of an emerging scientific discipline by using this combination of techniques (superposition and wordcloud) to explore its possibilities and limitations.

Keywords: bibliometric, mapping overlay, wordcloud, emerging field, e-learning, SCImago Journal & Country Rank

1. Introduction

The world scientific production analysis contributes, among many other things, to define the knowledge areas and subject categories that structure the generation of knowledge. Each classification system of scientific production defines its own areas and categories, which are mostly accepted by the scientific community that consults and feeds them. In this way, Scopus[1] classifies the works into 5 thematic clusters (life sciences, physical sciences, health sciences, social sciences and humanities), 27 knowledge areas and more than 300 subject categories.

[1] http://www.scopus.com

Web of Science[2] does it in 3 knowledge areas (sciences, social sciences and arts and humanities) and 250 thematic categories.

Scopus currently has more than 70 million records and a defined group of metadata[3] that are rigorously linked to each publication to describe its academic, social and geopolitical context. These two characteristics, having large volumes of structured information, are the inputs for the application of visualization techniques that generate new representations of knowledge, thus becoming powerful tools for science analysis.

The bibliometric data are very valuable to identify the scientific publications with the greatest impact in a given discipline, (i.e., Information Systems [1], Renewable Energy, Sustainability and Environment [2], to recognize different scientific fields and understand their internal dynamics and cognitive structure [3], either as an already consolidated research field or as an emerging discipline.

There are multiple methods and tools to visualize bibliometric information. For example, the distance-based, the graph-based or the time-based [4]. Mapping and clustering are also used to analyze the research fields of a scientific domain and the relationship between research fields and the evolution of the domain over time. As a tool, VOSViewer[4] assures the comprehensive visualization of nodes labels on the map. These maps, called science maps, help to locate research results to explore collaborations and publication trends, to observe the evolution of a certain subject or discipline and for benchmarking activities between regions, countries, institutions, authors and disciplines [5]. However, the visualization must have the capacity to handle large amounts of data at a small and large scale. This reduces the visual search time, providing a better understanding of a complex data set. It also reveals relationships that otherwise would not be noticed, allowing a data set to be viewed simultaneously from several perspectives, aiding the formulation of hypotheses and being an effective source of communication [6].

Through the overlay of science maps, the research bodies can be located visually within the sciences, analyzing the scientific development of properly established disciplines, trends or emerging research topics that do not fit into traditional subject categories. This is achieved thanks to the existence or construction of a stable corpus on which another smaller body can be overlaid [7], producing intuitive comparisons, of greater interpretation and with the potential to be used in scientific analysis.

In its essence, science maps are matrices of similarity measures, calculated from the correlation between items of information present in the structure of scientific communication. In other words, they show the disciplinary structure of the sciences in terms of publications. The stable or base map is constructed with bibliographic data from a database that has a definite categorization of the sciences. The analysis made from the overlap will be conditioned by the size of the data selected for it.

---

[2] http://clarivate.com/products/journal-citation-reports/

[3] https://www.elsevier.com/solutions/scopus/content

[4] http://www.vosviewer.com

In the words of Guzmán, "we can say that the analysis of information with science maps, supported by metric information studies, allows graphically representing the relationships between documents published by specific disciplines or scientific fields. These show the subareas of research in which the discipline has been focused over the years in order to identify, analyze and visualize the intellectual structure, as well as the temporal evolution in which the analyzed disciplines are being developed." [8].

Based on the above, science maps contribute to the identification of emerging disciplines by categorizing the publications that constitute their scientific communication channel [9].

However, a very select group of specialists usually carries out the analysis of these research products, since the results obtained are not easy enough to understand for most of the scientific community that is interested in knowing in detail the paths and trends that their discipline is taking. Faced with this need, other visualization techniques, such as wordclouds [10], infographics [11] and dashboards [12] have been positioned in virtual media as an alternative for the research results to achieve greater diffusion beyond the borders of scientific communication channels [13]. Wordclouds are used mostly to visualize a data set collected from surveys or forms. Among its advantages are: (a) its ability to abstract towards the essential, identifying and grouping existing patterns in writing [10], (b) they help to provide a general sense of the text (the same visceral response does not occur when looking at a text page) through the analysis of sentiment [14], (c) they provide a quick response on possible topics of interest and research for their community [15], (d) the visual representation of data generates impact among the audience, stimulating more questions than answers, and (e) they allow to share the results of the research in a way that does not require a deep understanding of the technicalities. Its link with the bibliometric analysis can be established considering the keywords field as the set of data collected from users (researchers) in a form (submit manuscript).

This study combines the use of the mapping overlay technique with the visualization of terms in wordclouds to represent the research fronts of a subject, in this case, the e-learning emerging discipline. The aim is to determine if this technique combination produces more intuitive, dynamic and easily accessible results for researchers and non-researchers.

2. Mapping a research field

To perform the research field mapping, we must first establish a body of documents to perform the bibliometric analysis, ensuring access to the bibliometric data of this set of publications. To analyze the e-learning case, we started with the methodology and findings of Tibaná-Herrera and others [9] for the subject categorization.

Secondly, the subject research fronts are identified, which determine the consolidation of the different tendencies over time that have contributed to the development and growth of the subject in scientific communications [16]. We propose the use of wordclouds composed of keywords [17], to visualize the research fronts of the field due to its representation capacity and rapid appropriation of the community to which it is presented.

2.1. Establishment of the body of documents

We start from the base that every research field has a set of scientific communications that contribute to the development of the subject. To identify these communications and analyze them, the subsequent steps can be followed:

Step 1. Definition of descriptors. It is about knowing all those terms present in the primary scientific literature with which the subject has been described. As expected, we start from a core term, which is generally the same as the research field. With this term, all the publications whose title, summary and keywords include the core term are identified in a comprehensible database.

E-learning case

- Core term: e-learning
- Data source: SCOPUS, database that indexes mostly journals and conference proceedings [18].

The search results should be refined according to the desired coverage degree in the analysis and the access availability of the bibliometric data.

E-learning case

- Publication type: Journal and Conference Proceeding
- Document type: Article, conference paper and review
- Analyzed timespan: 2012–2014. It corresponds to a period in which there is a stable worldwide production in e-learning, since in the previous period it was in constant growth and in the following period there was a significant decrease in production [19].
- Language: English.

The set of publications obtained can be used in its entirety or from a statistically representative sample.

E-learning case

- Results: 9291
- Representative sample: 2000 (21.6%)

Then, a bibliometric analysis based on keywords co-occurrence is carried out, aimed to determining the primary descriptors that are mostly present in the publications, their relationships and relevance, by means of the Visualization of Similarities (VoS) technique [20]. Additionally, they include secondary possible descriptors that reflect the same meaning, fruit of the linguistic similarities and/or acronyms or abbreviations that are used in the natural language. For example, when including the keywords of an article you can choose to use the *e-learning* or *elearning* descriptor [21].

E-learning case

- Keywords: 4521
- Primary descriptors: 51. E-learning, LMS, b-learning, online learning, Moodle, m-learning, ICT, learning objects, technology acceptance model, e-learning platform, adaptive learning, e-assessment, web-based learning, virtual learning environments, adult learning, informal learning, instructional design, SCORM, augmented reality, educational technology, intelligent tutoring systems, remote laboratory, simulation, learning analytics, learning environments, e-learning 2.0, teaching and learning, interactive learning environments, educational data mining, gamification, learning design, social learning, lifelong learning, metadata, MOOC, virtual classroom, labview, learning methods, personal learning environments, adaptive e-learning systems, computer-based learning, information literacy, virtual learning, Blackboard, continuing education, game-based learning, interactive learning, personalized learning, recommender systems, virtual laboratories, virtual reality.
- Secondary descriptors: 13. elearning, electronic learning, Learning management system, blearning, blended learning, mlearning, mobile learning, Information and communications technologies, eassessment, electronic assessment, VLE, Massive Open Online Courses and PLE.

Step 2. Correspondence of publications and descriptors. In a matrix containing all the indexed scientific publications and the primary and secondary descriptors identified, the number of articles published by the Conference Proceeding or the Journal with that descriptor in the title, abstract and keywords fields is recorded at each crossing. It is very important to use the same selection criteria described in the previous step to ensure information integrity. Then, the primary and secondary descriptors related to the same term are added, assuming that the sum reflects unique publications related to each other by the descriptors.

E-learning case

- Journals and conference proceedings included in the matrix: 12.923

Step 3. Percentage of participation in the subject (PP). It is the percentage of articles in the publication that are related to the subject during the timespan established in the initial criteria, this is done by taking the maximum number of articles per descriptor, bearing in mind that an article may be related to more than one descriptor.

E-learning case

Correspondence matrix description (**Figure 1**):

- 3.680 journals and conference proceedings do not have any publication related with any of the 64 descriptors.
- 7.801 journals and conference proceedings have a PP lower than 5%.

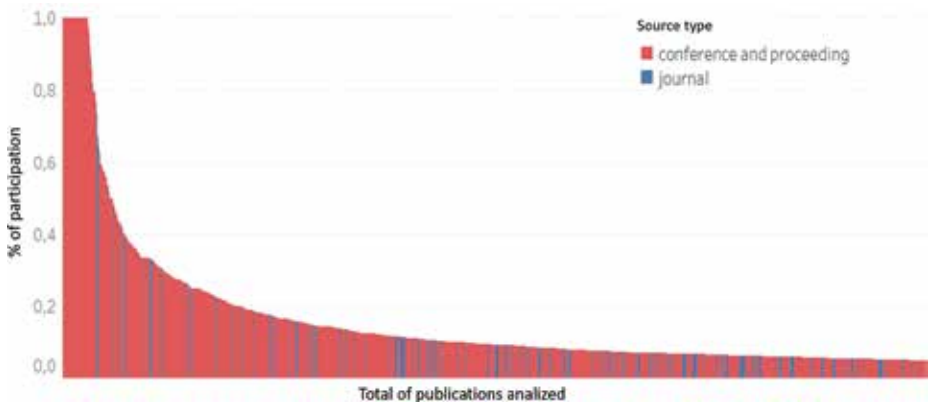


Figure 1. Percentage of participation (PP) of the term in journals and conferences. Source: [9].

Step 4. Cut-off point for the inclusion of publications in the analysis. You must determine the cut-off point over the PP from which the publications for the categorization of the thematic will be included. Other studies have classified publications among "pure", "hybrid" and "unrelated" publications in a given subject [1] and on the determination of the core set of publications [21]. However, we believe that this value should be established through the combination between the maximum allowed error of the subject relation of the publication and the average PP of the total set of publications. The higher the cut-off point, the greater the precision in the selection of journals will be. Although, this precision means a reduced volume, and if not, a low cut-off point increases the error in the selection and its volume. Once the cut-off point is established, all publications that exceed this threshold are considered as the basic set of analysis of the emerging subject category.

E-learning case

- The set of publications must maintain an average PP higher than 50%, for which the cut-off point per publication was established at 25% (coinciding with the classification of pure and hybrid publications [1]).
- The cut-off point included 11 publications that were excluded because they defined other areas of knowledge in their scope.
- 82 journals and 137 conference proceedings that meet the criteria of the methodology were identified.

Step 5. Publication set analysis. The set of selected publications is analyzed under a bibliometric approach (a) to determine if it represents the existence of a scientific community that communicates its knowledge through these channels and (b) to recognize it as an emerging and distinctive scientific discipline that can be defined as a transversal thematic category [5]. For this, the mapping overlay technique [7] can be used, which facilitates the exploration of the knowledge bases of an emerging discipline and its evolutionary dynamics. This technique requires a base map on which to overlay a local map (thematic) and thus make comparisons. This overlap allows placing the discipline in the general topology of scientific knowledge and identifying whether a cluster effect occurs, which should be considered as evidence of the existence of a specific disciplinary field from the point of view of scientific communication guidelines followed by the researchers.

The relation degree of publications is established by the normalized value produced by the combination of citations, co-cites and coupling [22, 23]. In addition, this analysis can be enriched with the distribution by clusters that visualization tools perform, such as VOSViewer [24].

E-learning case

- The base map is a global map of science that includes the total number of publications indexed in SCOPUS, made up of 7 clusters, which in a clockwise and broad sense can be named as follows: Social Sciences (red), Psychology (light cyan), Medicine (green), Health Sciences (purple), Life Sciences (yellow), Physical Sciences (dark cyan) and Engineering and Computer Science (blue) (**Figure 2**).
- The composite indicator was arranged by SCImago Journal & Country Rank[5].
- The local map that is overlaid on the global map of science is the set of 219 publications selected in the previous step (**Figure 3**).
- There is a cluster effect that shows a high cohesion among publications, which is sufficient evidence, in terms of scientific communication, that e-learning is a distinctive

[5] http://www.scimagojr.com

scientific discipline, since there is a network of relationships and interactions that are established between the authors and scientists who share thought structures, cooperation patterns, language and forms of communication.

- The publications distribution shows a main group in Social Sciences and other small groups in Computer Science and Psychology.

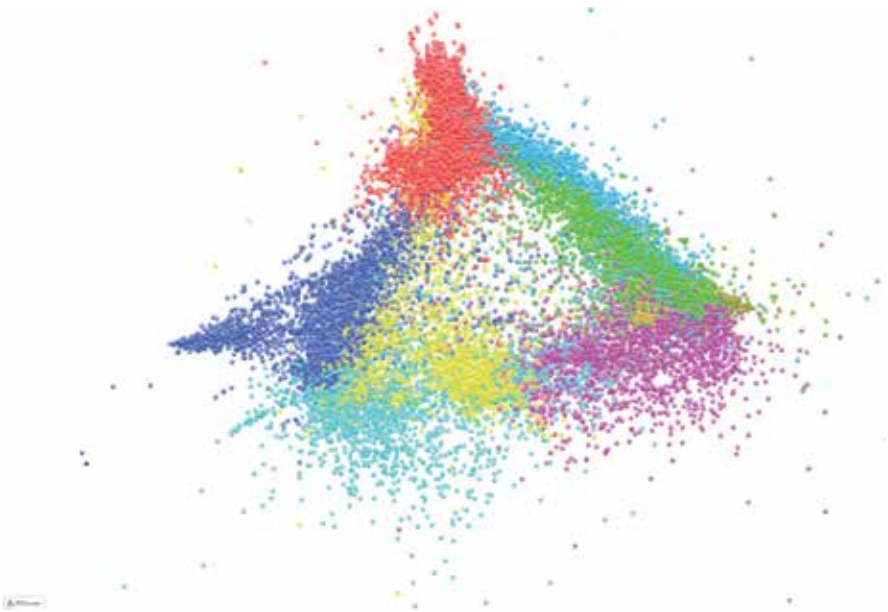


Figure 2. Global map of science based on SCOPUS and SCImago Journal & Country Rank using VOSViewer with its density map setting (Source: [9]).

2.2. Identification of research fronts

To identify the research fronts through the visualization of keywords in a wordcloud, it is necessary to identify the body of publications on which the analysis is going to be carried out (previous section). Then, all the keywords of the publications are extracted, keeping the same filters defined in the previous stages, with the confidence of finding a set of structured and well-defined terms. This technique provides value when the data has a treatment that ensures a correct interpretation. This is done through two tasks, being the first to refine the set of terms (which can be in the order of thousands) to obtain those that are mostly different and that can be visually represented without loss of information. The refinement process may include a minimum threshold of articles published by a journal or conference report to ensure that there is a volume and regularity guaranteed in the conceptual development of the thematic. It can also be refined by defining the number of terms to be displayed in the wordcloud.

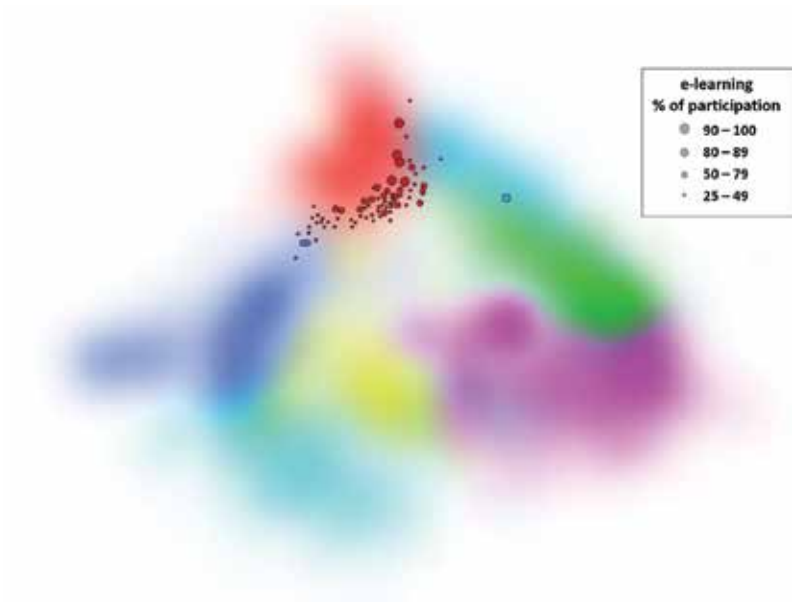


Figure 3. Distribution of publications related to the thematic, using the mapping overlay technique with VOSViewer in its density map configuration. The color of the publication indicates the area of knowledge in which it is superimposed and its size corresponds to the percentage of participation The size of the selected publications has been modified for visual purposes (Source: [9]).

E-learning case

- Publication type: Journal and Conference Proceeding
- Document type: Article, conference paper and review
- Analyzed timespan: 2012–2014.
- Language: English.
- Minimum number of papers published by Journal/Conference Proceeding: 100
- Number of terms to display: 100

The second task is to configure the variables that determine the form of the wordcloud, among which are:

1. Keep each term with its own length. You can fall into the error of disaggregating terms, for example, the term *Information and Communication Technologies* should remain as one and not separate it into 3 or 4 parts.
2. Don't include terms in the visualization that correspond to the same name of the scientific field analyzed, places, dates, proper names, names of organizations and all others that don't contribute to the identification of research fronts.

3. Define simple shapes to represent the cloud. Today there are multiple wordcloud creation tools. Most of them allow to use a defined image for the cloud layout. It is recommended to use images without internal content, only frame, so that the words can be distributed inside without obstacles.
4. Select a Sans Serif font. The wordclouds are presented more frequently in digital media, in which a clean, non-blurred reading is sought, to avoid visual fatigue
5. Define an intention for the color usage. The visual representation should be as enriched as possible. Therefore, the color defined for each term must show its own characteristic. A good intention of color is to establish clusters of terms [23] that determine the main research fronts.

E-learning case

Examples of wordclouds.

Option 1:

Option 2:

Option 3:

Finally, by means of a rapid visual analysis of the generated wordcloud, the research fronts of the scientific field can be identified in a differentiated way.

E-learning case

- Based on the results shown in **Figures 4–6**, two significant clusters can be identified (**Table 1**):
- The most outstanding research fronts of e-learning are those that analyze the design and construction of interactive learning environments and teaching and learning strategies in the virtual modality

A limitation of wordclouds, that can affect the reader's interpretation, is the term length that can capture a quick attention being located in a central place of the visualization without having significant weight. However, this visualization technique is a powerful tool to abstract relevant information from large volumes of information, in addition, it can be used to observe the main trends of other bibliometric data. For example, journals and congresses with the greatest influence in the discipline or the institutions and countries that contribute the most to the discipline productivity.

http://dx.doi.org/10.5772/intechopen.76731

Figure 4. Wordcloud of e-learning worldwide, based on data from SCImago Journal & Country Rank in a positive diagonal format (Source: Self-made).

Figure 5. Wordcloud of e-learning worldwide, based on data from SCImago Journal & Country Rank in positive and negative diagonal format (Source: Self-made).

3. Conclusions

This study proved that bibliometric analysis combined with visualization techniques provides sufficient elements to map an emerging discipline, in this case study, e-learning.

Figure 6. Wordcloud of e-learning worldwide, based on data from SCImago Journal & Country Rank in horizontal format (Source: Self-made).

CLUSTER 1 (Red color)	
KEYWORD	OCCURRENCES
Interactive learning environments	54
Teaching/Learning strategies	50
CLUSTER 2 (Orange color)	
KEYWORD	OCCURRENCES
Collaborative learning	38
Pedagogical issues	28
Social media	26
Application in subject areas	26
Computer-mediated communication	25
Mobile learning	24
Improving classroom teaching	22
Media in education	20
Elementary education	20

Table 1. Main fronts of e-learning research worldwide, with the occurrence values in the wordcloud obtained from SCImago Journal & Country Rank (Source: self-made).

The mapping overlay technique allows visualizing the existing cohesion between the scientific communications generated by the community of researchers in the subject, determining the knowledge areas in which the research activity is developed and establishing the base set of

publications for other bibliometric analyzes. Through this technique it was determined that e-learning has its scientific development mainly in the social sciences.

The visualization of the main keywords present in the set of publications of a discipline through wordclouds, allows to clearly identify the research fronts of this subject, by grouping the research topics and showing their relative weight in the scientific development of the discipline. In the case study, two main research fronts were identified in e-learning, interactive learning environments and teaching and learning strategies.

Acknowledgements

To SCImago Research Group for providing citation data for publications.

Conflict of interest

Not applicable.

Notes/Thanks/Other declarations

The data related to this research were obtained, on the one hand, from the access to SCOPUS and on the other, provided by SCImago Research Group. These are protected by licensing and copyright respectively.

Author details

Gerardo Tibaná-Herrera[1,2*], María Teresa Fernández-Bajón[1] and Félix de Moya-Anegón[3]

*Address all correspondence to: gtibana@gmail.com

1 Complutense University of Madrid, Madrid, Spain

2 SCImago Research Group, Bogotá, Colombia

3 SCImago Research Group, Madrid, Spain

References

[1] Chan HC, Guness V, Kim HW. A method for identifying journals in a discipline: An application to information systems. Information and Management [Internet]. 2015;**52**(2): 239-246. DOI: 10.1016/j.im.2014.11.003

[2] Romo Fernández LM, Guerrero Bote VP, Moya AF. Análisis de la producción científica española en energías renovables, sostenibilidad y medio ambiente (Scopus, 2003-2009) en el contexto mundial. Investigación Bibliotecológica Archivonomía, Bibliotecología e Información. 2013;**27**(60):125-151. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0187358X13725462

[3] Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. Journal of Informetrics [Internet]. 2011;**5**(1):146-166. DOI: 10.1016/j.joi.2010.10.002

[4] van Eck NJ, Waltman L. Visualizing bibliometric networks [Internet]. Measuring Scholarly Impact. 2014. pp. 285-320. Available from: http://link.springer.com/10.1007/978-3-319-10377-8_13

[5] Leydesdorff L, de Moya-Anegón F, Guerrero-Bote VP. Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of Scopus data (1996–2012). Journal of the Association for Information Science and Technology [Internet]. 2015 May; **66**(5):1001-1016. DOI: 10.1002/asi.23243

[6] Börner K, Chen C, Boyack KW. Visualizing knowledge domains. Annual Review of Information Science and Technology [Internet]. 2005;**37**(1):179-255. DOI: 10.1002/aris.1440370106

[7] Rafols I, Porter AL, Leydesdorff L. Science overlay maps: A new tool for research policy and library management. Journal of the Association for Information Science and Technology [Internet]. 2010 Sep;**61**(9):1871-1887. DOI: 10.1002/asi.21368

[8] Guzmán Sánchez MV, Trujillo Cancino JL. Los mapas bibliométricos o mapas de la ciencia: una herramienta útil para desarrollar estudios métricos de información. Bible University [Internet]. 2013;**16**(2):95-108. Available from: http://revistas.unam.mx/index.php/rbu/article/view/43851

[9] Tibaná-Herrera G, Fernández-Bajón MT, de Moya-Anegón F. Categorization of an emerging discipline in the world publication system (SCOPUS): E-learning. 2017 Oct 16 [cited 2018 Feb 1]. Available from: http://arxiv.org/abs/1710.05723

[10] Flamary R, Anguera X, Oliver N. Spoken WordCloud: Clustering recurrent patterns in speech. In: 2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI) [Internet]. IEEE; 2011. pp. 133-138. Available from: http://ieeexplore.ieee.org/document/5972534/

[11] Alcíbar M. Information visualisation as a resource for popularising the technical-biomedical aspects of the last Ebola virus epidemic: The case of the Spanish reference press. Public Understanding of Science [Internet]. Apr 10, 2018;**27**(3):365-381. Available from: http://journals.sagepub.com/doi/10.1177/0963662517702047

[12] Verbert K, Duval E, Klerkx J, Govaerts S, Santos JL. Learning analytics dashboard applications. American Behavioral Scientist [Internet]. 2013 Oct 28;**57**(10):1500-1509. Available from: http://journals.sagepub.com/doi/10.1177/0002764213479363

[13] Dinsmore A, Allen L, Dolby K. Alternative perspectives on impact: The potential of ALMs and Altmetrics to inform funders about research impact. PLoS Biology [Internet]. 2014; **12**(11):e1002003. DOI: 10.1371/journal.pbio.1002003

[14] Bashri MFA, Kusumaningrum R. Sentiment analysis using latent Dirichlet allocation and topic polarity wordcloud visualization. In: 2017 5th International Conference on Information and Communication Technology (ICoIC7) [Internet]. IEEE; 2017. pp. 1-5. Available from: http://ieeexplore.ieee.org/document/8074651/

[15] Jo Y, Kim E, Shin Y. Graphical keyword service for research papers with text-mining method. In: ICCDA '17 Proceedings of the International Conference on Compute and Data Analysis [Internet]. 2017. pp. 185-190. Available from: https://dl.acm.org/citation.cfm?id=3093242&dl=ACM&coll=DL

[16] Pinto M. Viewing and exploring the subject area of information literacy assessment in higher education (2000–2011). Scientometrics [Internet]. 2015 Jan 15;**102**(1):227-245. Available from: http://link.springer.com/10.1007/s11192-014-1440-2

[17] Cantos-Mateos G, Zulueta M-Á, Vargas-Quesada B, Chinchilla-Rodríguez Z. Estudio evolutivo de la investigación española con células madre. Visualización e identificación de las principales líneas de investigación. El Profesional de la Información [Internet]. 2014; **23**(3):259-271. Available from: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2014.may.06

[18] Leydesdorff L, De Moya-Anegón F, Guerrero-Bote VP. Journal maps on the basis of scopus data: A comparison with the journal citation reports of the ISI. Journal of the American Society for Information Science and Technology. 2010;**61**(2):352-369

[19] Tibaná-Herrera G, Fernández-Bajón MT, de Moya-Anegón F. Global Analysis of the E-Learning Scientific Domain: A Declining Category? Scientometrics [Internet]. 2017 Dec 5. Available from: http://link.springer.com/10.1007/s11192-017-2592-7

[20] Waltman L, van Eck NJ, Noyons ECM. A unified approach to mapping and clustering of bibliometric networks. Journal of Informetrics [Internet]. 2010;**4**(4):629-635. DOI: 10.1016/j.joi.2010.07.002

[21] Chiang JK, Kuo C-W, Yang Y-H. A bibliometric study of E-learning literature on SSCI database. In: International Conference on Technologies for E-learning and Digital Entertainment [Internet]. 2010. pp. 145-55. Available from: http://link.springer.com/10.1007/978-3-642-14533-9_15

[22] Madhugiri VS, Ambekar S, Strom SF, Nanda A. A technique to identify core journals for neurosurgery using citation scatter analysis and the Bradford distribution across neurosurgery journals. Journal of Neurosurgery [Internet]. 2013;**119**(5):1274-1287. http://thejns.org/doi/10.3171/2013.8.JNS122379 or https://doi.org/10.3171/2013.8.JNS122379

[23] Hassan-Montero Y, Guerrero-Bote V, De-Moya-Anegón F. Graphical interface of the SCImago journal and country rank: An interactive approach to accessing bibliometric information. El Profesional de la Informacion [Internet]. 2014;**23**(3):272-278. Available from:

http://www.scopus.com/inward/record.url?eid=2-s2.0-84903581415&partnerID=40&md5=cd75fe9502aa257c52bf0b9334e9ba07

[24] van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics [Internet]. 2010 Aug 31;**84**(2):523-538. DOI: 10.1007/s11192-009-0146-3

Chapter 5

Collaboration and Citation Analysis Within Social Sciences: A Comparative Analysis Between Two Fields

Alexander Maz-Machado and
Noelia Jiménez-Fanjul

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76732

Abstract

The present study focuses on a collaboration of a citation analysis of the JCR journals of the categories *Demography* and *Urban Studies* indexed in *Social Science Citation Index* from the period 2000–2016. A total of 64 journals were covered (26 for Demography and 38 for Urban Studies). We found that the percentages of multi-authored documents in both categories are very similar; moreover, the citation distribution is shown to be increasing in both but behaves slightly different in the two samples analysed. It seems to be a relation between the number of citations a document received and the number of authors. Regarding international collaboration, both categories present a similar type of network with densities of the kind of social science networks. Anglo-Saxon countries are the most prolific ones and the biggest collaborators in both networks. *Urban Studies* shows a relative importance to countries of emerging economies since it indexed more journals in the sample with a wider regional scope.

Keywords: bibliometrics, collaboration, citation, scientific production, social sciences

1. Introduction

Enquiries about science point to the existence of valid indicators to measure the level of scientific activity and scientific accomplishments from various perspectives: scientific fields, authors, institutions, faculties, departments, research groups and countries [1, 2]. The results of such studies are complemented with another set of indicators and are used at different governmental and organisational levels in, among other things, allocating economic and human resources [3]. It is increasingly evident given the development and consolidation of research

evaluation systems in almost every country. The situation represents a crucial shift in the nature of the behaviour of institutions and organisations that develop research programmes and projects [4, 5].

Van Raan [6] includes, as one of the objectives of bibliometric analysis, the ability to establish a set of standardised indicators that facilitate the evaluation of scientific production. The characteristics and indicators that are obtained from bibliometric studies are useful for planning, developing and organising the resources and services of the institutions in charge of the administration [7, 8].

Bibliometric studies are enormously relevant to the identification and characterisation of the scientific profile of countries, institutions for research and scientific fields themselves [9]. This statement is based on how they facilitate, among other things, the detection of research patterns or research strengths for each of the agents participating in the scientific process. Furthermore, evaluations with a basis on bibliometric indicators for citation have become commonplace in national processes for the evaluation of research at a university, faculty and even departmental levels [10].

1.1. What is scientific collaboration?

Scientific work is no longer an individual task having researchers work in isolation but a collaborative endeavour, instead. In this manner, collaboration is present in all the fields of knowledge and takes a wide range of forms. Scientific co-authorship is thought of as a reaction to the process of professionalisation of research, in terms of publication [11]. Katz and Martin [12] state that it can happen between individuals, groups, departments, institutions, sectors, regions or countries.

Many are the reasons that lead researchers to collaborate, from which the following stand:

1. Professionals seek opportunities to collaborate in order to increase their visibility within their field; it can be assumed that it applies to all fields of knowledge, since sciences generally share a common reward structure [13].
2. To gain access to equipment, resources or materials that may facilitate or improve research [12].
3. To improve the composition of research groups with a view to increase the chances of gaining financial support in open calls.
4. To know and share new methodological techniques.
5. To increase efficacy and efficiency, as well as quality of research [14].
6. To establish research networks with a greater social and scientific salience.
7. The chances of researching about interdisciplinary matters that touch on different areas of knowledge, due to which experts from each of them are necessary.
8. To interact with institutions of equal or higher prestige or to support the development of others of a less established research tradition.

9. To increase the scientific productivity of either research groups or their members.
10. To work with colleagues who share the same interests, ideas, theoretical frameworks or problems.
11. To increase citation and, hence, the impact and visibility of scientific production [15].

Occasionally, professionals who seek to add something new to their field may find that the reward is greater in doing so through the search of diverse ideas and remote collaborators than in collaborating with others from their own laboratory [16]. The increase in international collaboration in research may be regarded as a consequence of the mentioned rationales for establishing new links within science.

When remote collaborators have different points of view and experiences, they can be more easily prone to questioning—or perhaps complementing—the perspectives and capacities of the other participants [16]. For this reason, it is likely that these collaborations result in research studies of a more innovative kind and promote progress within the field of research itself. Nonetheless, collaboration between over-specialised scientists is in some cases necessary to tackle certain problems that are highly specific within a particular field of knowledge [17].

Glänzel [18] points out that the relation between collaboration and scientific productivity is a very important aspect of research. This has led to bibliometric analysis becoming highly recursive in the literature on informational sciences or social studies about science. There have been attempts to find collaboration patterns in countries or regions for a specific scientific field; for instance, clinical medicine in Taiwan [19] and epidemiology in Bulgaria [20]. Similarly, collaboration patterns at the global level of sciences have been studied in Eastern Europe [21, 22] and, in Spain, the production in Science Citation Index (SCI), Social Sciences Citation Index (SSCI) and Arts & Humanities [23, 24]. The field of Library Information Science itself (LIS) has been subject to various collaboration analyses [25–30].

Many of the studies reveal that collaboration raises not only participants' productivity but also the impact of their research [15]. However, Katz and Hicks [31] assert that the impact of an article in terms of citation is partially related with the number of participant authors, institutions and countries. In a study carried out by Narin and Whitlow [32] for the European Union, it was found that articles in which several institutions participated were more cited than those in which only one does. Likewise, articles are more cited when collaborators are foreign as compared with those that are signed by local or national collaborators.

Another aspect that attracts the attention of research on collaboration is the types of collaboration in terms of regions, determining if it is local, national or international [25, 33].

To measure collaboration, various indicators have been established, among which we highlight the following:

a) Collaboration Index (CI) defined by Lawani [34]: $IC = \frac{\Sigma_{j=1}^{A} j f_j}{N}$

b) Degree of Collaboration (DC) [35]: $DC = 1 - \frac{f_1}{N}$

c) Collaborative Coefficient (CC) [36]: $CC = 1 - \frac{\sum_{j=1}^{A}\left(\frac{1}{j}\right) f_j}{N}$

fj = number of documents with j authors in collection *K*.

N = total number of documents in *K*.

A = total number of authors in collection *K*.

Collaborative research studies generally focus on a particular field in relation with itself or to a country or region. When studies in Social Sciences seek to compare collaboration indicators, it is usually done among subdisciplines within the same scientific field.

In this study, we aim to compare the collaboration between two different scientific fields of the Journal Citation Report (JCR), Social Sciences edition [37] with differences in the volume of scientific production indexed in the Web of Science (WOS) in the period 2000–2016.

2. Materials and methods

The 2016 JCR® Social Sciences Edition [37] was retrieved on June 1, 2017, to find out the name and number of the journals within the categories of *Demography* and *Urban Studies*. For the former, 26 journals were found, and 38 for the latter.

The time interval covered in this study is from 2000 to 2016. The procedure to obtain the data consisted in analysing the information contained in the SSCI, for which all the records were searched using the parameters: *Publication Name* [name of each journal in the chosen category] and *Year Published* [2000–2016]. In order to extract information only from citable documents, these were filtered once again by their categorisation as *Article* or *Review* (from now on, we are to refer them as documents). The category of *Demography* produced 11,361 documents whereas *Urban Studies* produced 24,010. Out of those documents, those in which the author was anonymous, or the author field was blank, were discarded. Lastly, 11,361 entries were considered for *Demography* and 23,998 for *Urban Studies*, all of which constitute the sample of this study.

All the information was uploaded to an *ad hoc* Microsoft® Access® 2016 relational database (version 1801) for the treatment and normalisation of data, as well as to produce the different graphs. The data were collected by year and collaboration was analysed into two levels. The first level was authorship, looking at collaboration in relation with the number of signatory authors; the number of authors in each document was full-counted, calculating a particular Collaboration Index (CI) and Degree of Collaboration (DC). The second level was established in relation with international collaboration, identifying the countries of each of the authors' institutions.

With a view to count the authors of each document, we opted for the complete counting system, as suggested by Cronin and Overfeld [38], attributing full authorship to each co-author,

considering them equally. The same procedure was applied in the case of countries. The documents were grouped according to collaboration by country, as has been done in other similar studies [39]. Given that documents can be signed by authors from different countries, the sum of the percentages is greater than 100%.

To analyse, treat and visualise collaborative networks, we have used the Pajek software [40].

3. Results and discussion

The category *Urban Studies* presents 28 indexed journals in the 2016 JCR [37], 12 more than *Demography*; this is 31.57% more. In the period between 2000 and 2016, the documents indexed within the category *Demography* mounted to less than half of those in *Urban Studies*, more precisely, only 47.36%. During those years, a total of 35,359 documents were indexed, considering both categories, in the SSCI.

3.1. Collaboration in the category demography

Concurrently with the increase of the production of documents along the period between 2000 and 2016, there was also an increase in the number of authors per article and, with it, collaboration in the category *Demography* (**Figure 1**). There is a correlation of 0.992 with a significance of 0.01, between the number of published documents and the number of documents with multiple authorships. Early in the set period of time, the difference between single and multiple authorship documents was of only 8.8%. Despite continuous ups and downs, such difference increased slowly by up to 20% in 2009. In 2010, the difference increased to 42% and remained ever since within an interval of a minimum of 30.3% to a maximum of 51.7%.

All this multiple authorship has an impact on collaboration indexes. In this line, the DC increased gradually from 0.52 in year 2000 to a top 0.67 in 2014 and 2015. Likewise, the CI ranges between an initial 1.87 and a maximum of 2.34 in 2014 (**Table 1**). The overall values for the time interval between years 2000 and 2016 are DC = 0.605 and CI = 2.14.

Figure 2 shows that the 70.4% of the documents from the *Demography* category are signed by one or two authors. A total of 39.5% of the papers have only one author, while articles with four or less authors only represented 13.07%.

The production of documents within the category *Demography* between 2000 and 2016 received a total of 147,024 citations. The average citation is of 12.9 cites per document (*SD* = 32.54), notwithstanding that 1840 received no citation at all, which represents 16.25 of total production.

Analysing citation in relation with author collaboration, it can be seen that multi-author documents receive 63.98% of the total citations while single-author documents receive 36.02%. In differentiating documents according to the number of authors, the highest citation is received by the documents signed by a single author, followed by those signed by two and three

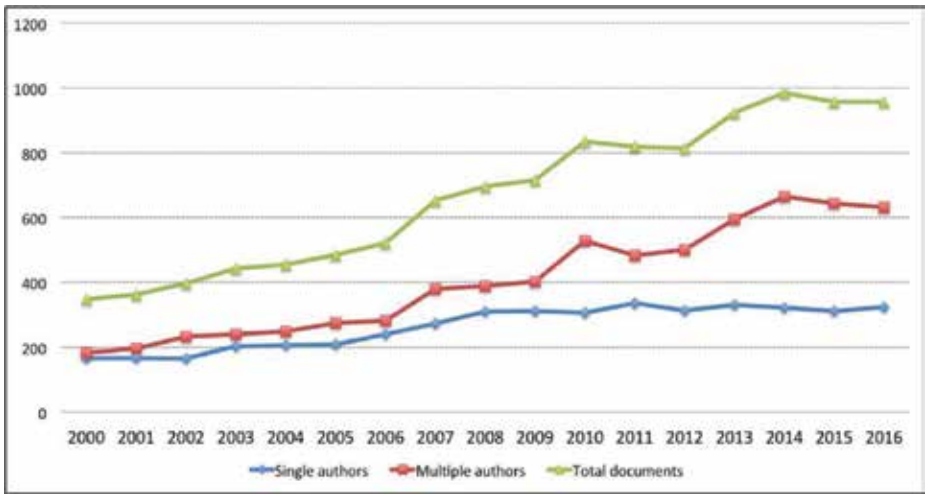


Figure 1. Diachronic type of authorship in the category demography.

Year	DC	CI
2000	0.52	1.87
2001	0.54	1.90
2002	0.58	2.03
2003	0.54	2.01
2004	0.55	2.03
2005	0.57	2.09
2006	0.54	1.96
2007	0.58	2.02
2008	0.56	2.01
2009	0.56	2.05
2010	0.63	2.16
2011	0.59	2.11
2012	0.62	2.17
2013	0.64	2.28
2014	0.67	2.32
2015	0.67	2.34
2016	0.52	1.87

Table 1. Degree of collaboration and collaboration index in the category demography.

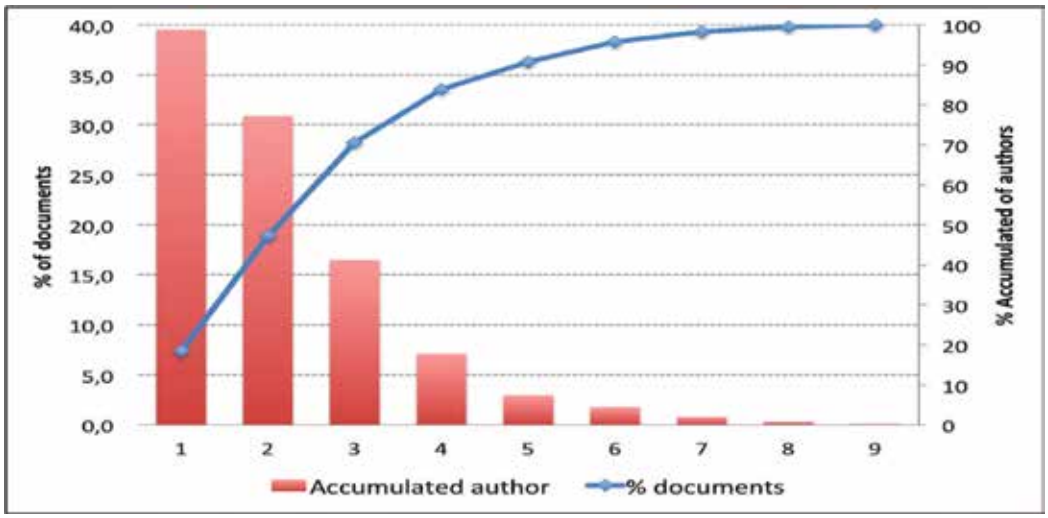


Figure 2. Co-authored distribution in demography (2000–2016).

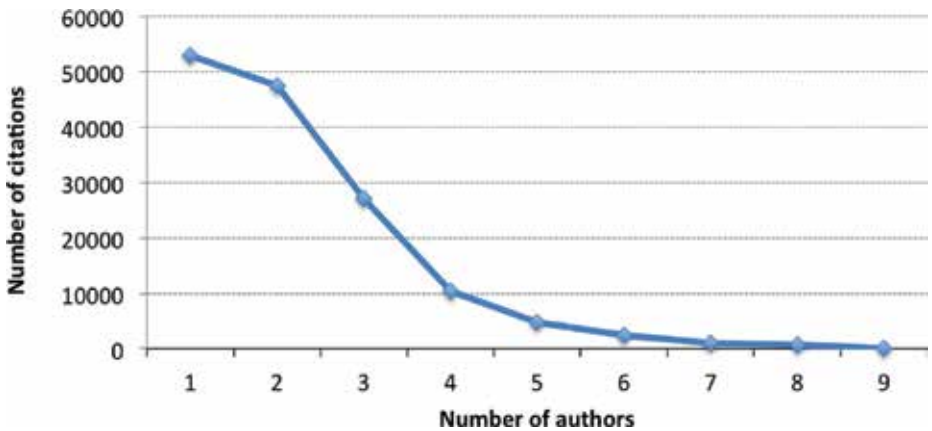


Figure 3. Citation by number of authors in demography 2000–2016.

authors, and decreases as the number of signatory authors increases (**Figure 3**). There is a moderated correlation between citation and number of signatory authors, with a Pearson's correlation coefficient of 0.709 (p = .001) for the category. In the same line, citation and DC present a correlation coefficient of 0.542 (p = .025). Eight articles received more than 300 citations; one article received 806 citations.

Regarding international collaboration, only 10,479 documents (out of 11,361) presented affiliation information. The documents of the category *Demography* were written by authors affiliated to institutions of 147 different countries. Most of the documents (77.6%) in the sample are written by authors from the same country regardless if they are written by multiple authors

or not. **Figure 4** shows a tendency in the increase of the international collaboration between authors which provide much better visibility and further citation to the work [13, 15].

There are only 11 countries (*Barbados, Bolivia, Cape Verde, Hong Kong, Malta, Oman, Solomon Islands, Syria, Trinidad and Tobago, Yemen, Yugoslavia*) that do not collaborate with other countries in the sample. The country with most co-authorship with other countries in the world is the *USA*, relating with 104 countries. A total of 50.34% of the countries (74) have relationships with a maximum of four other countries.

France, Germany, England and the *USA* are the only four countries that co-write articles with more than 50 other countries.

The network depicted in **Figure 5** shows a general view over the country network for Demography considering all the period. Every vertex represents a country; the volume of a vertex is proportional to the number of documents written by authors of the country. The lines between vertices show that the linked countries co-write documents and the colour of the lines are proportional to the number of documents shared. International collaboration networks tend to be very dense. The density of the network in **Figure 5** is 0.06737933 which indicated that the network is dense for social sciences. The average degree of the countries is 9.9048, which means that each of the 147 countries in the network shares documents with almost 10 other countries.

There are 728 collaborations detected, most of them being anecdotal; 48.08% of these collaborations appear only once, which means that these two countries only co-write one document in the whole period. The most prolific relationships among countries are found to be between *England* and the *USA, Canada* and the *USA, Germany* and the *USA* with more than 100 documents shared by each.

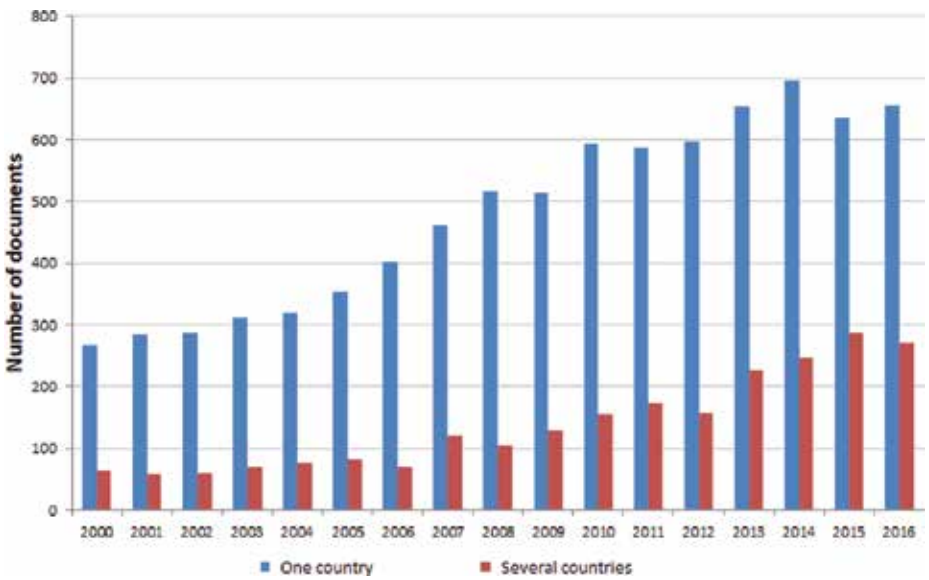


Figure 4. Diachronic international collaboration in demography 2000–2016.

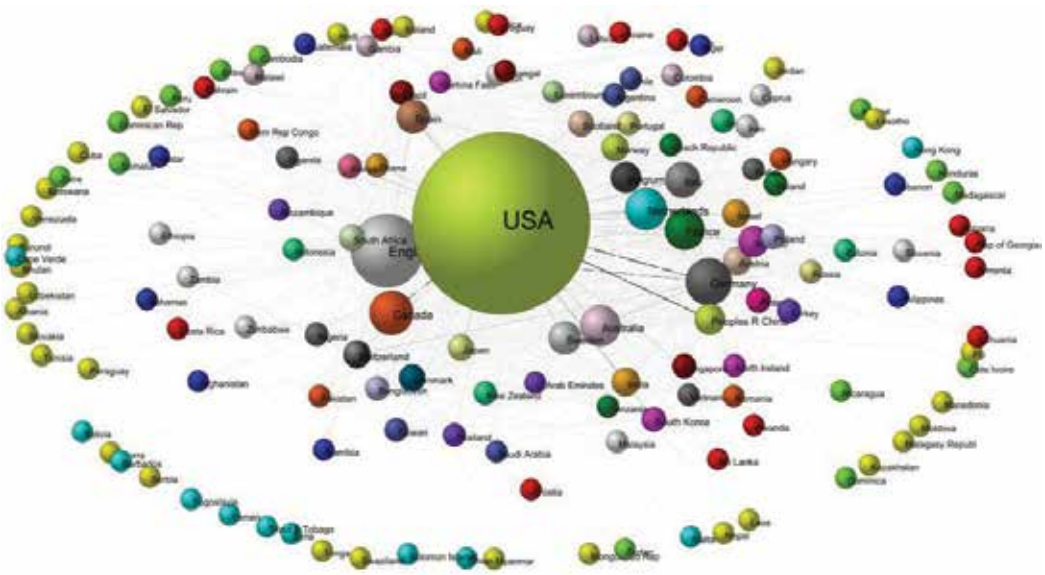


Figure 5. General network for international collaboration in demography 2000–2016.

3.2. Collaboration in the category urban studies

In this category, the documents ranged from 1015 in the year 2000 to 2170 in 2016, so that production has doubled since the beginning of the period studied (**Figure 6**). Initially, the percentage of documents signed by only one author (54.58%) was slightly higher than the one for multi-authored documents (45.42%). These values have varied along the years, with the proportion being reversed in 2016, reaching 68.89% for multi-authorship and 31.11% for single authorship. Since 2005, the number of multi-authored documents prevails, showing a continued growth. The average of authors is 2.06 authors per document ($SD = 1.27$). There is a correlation between the total production of the documents and those of multi-authorship with a positive significance (.926, $p < .01$).

Indicators suggest that this collaboration has increased in the period. The DC increased from 0.45 to 0.69, while the CI varied from 1.67 in 2000 to 2.43 in 2016 (**Table 2**). Globally for the interval analysed, the value of DC = 0.813 and CI = 2.07.

Figure 7 shows that the 89.12% of the documents from the *Demography* category are signed by one, two or three authors. A total of 40.89% of the papers have only one author, while articles with four or less authors only represented 10.88%.

Between 2000 and 2016, the *Urban Studies* category received 377,473 citations. The average is that every document in the sample has been cited 15.6 times ($SD = 29.27$). A total of 11.14% of the documents have never been cited.

The multi-authored documents received 61.8% of the citations, while those written by a single author received 38.2%. According to the number of authors, the highest citation is received by papers signed by a single author, followed by those of two and three authors, all of whom

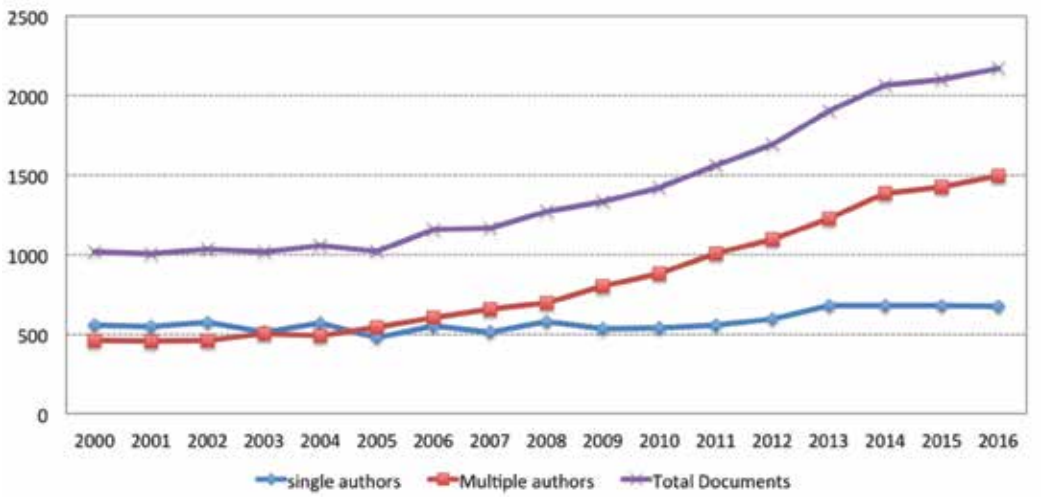


Figure 6. Diachronic type of authorship in the category urban studies.

Year	DC	CI
2000	0.45	1.67
2001	0.45	1.68
2002	0.44	1.69
2003	0.50	1.76
2004	0.46	1.75
2005	0.53	1.86
2006	0.52	1.86
2007	0.56	1.97
2008	0.55	1.91
2009	0.60	2.04
2010	0.62	2.09
2011	0.65	2.16
2012	0.65	2.19
2013	0.64	2.21
2014	0.67	2.31
2015	0.68	2.40
2016	0.69	2.43

Table 2. Degree of collaboration and collaboration index in the category urban studies.

received 88.31% of the citations (**Figure 8**). Data present a high correlation between citation and number of signatory authors, with a Pearson's correlation coefficient of 0.892 (p = .00) for the category *Urban Studies*. Citation-DC correlation coefficient was 0.878 (p = .00) which is an evidence of strong correlation between both variables. It is evident that for documents signed by more than four authors there is a decrease in the number of citations received. Seven articles received more than 500 citations.

The most cited document has 2004 citations and is signed by 2 authors.

For the international collaboration, only 23,577 registers were considered for being the only ones that incorporate information about authors' affiliation. The authors were affiliated to institutions of 133 countries.

The documents of the category *Urban Studies* were mostly written by authors affiliated to the same country, in fact only 16.12% of the documents were written in international collaboration. **Figure 9** shows these results analysing the international collaboration along the period. We can see an increase in this collaboration since 2010, resulting in this tendency being slightly lower than the one found for *Demography* sample. It is remarkable that the category *Urban Studies* involved less countries than *Demography* which led to a less collaboration among countries.

Only 9 countries out of 133 contributed with documents without international collaboration (*Algeria, Azerbaijan, Bolivia, Hong Kong, Morocco, Sudan, Tunisia, Uruguay, Yugoslavia*) that do

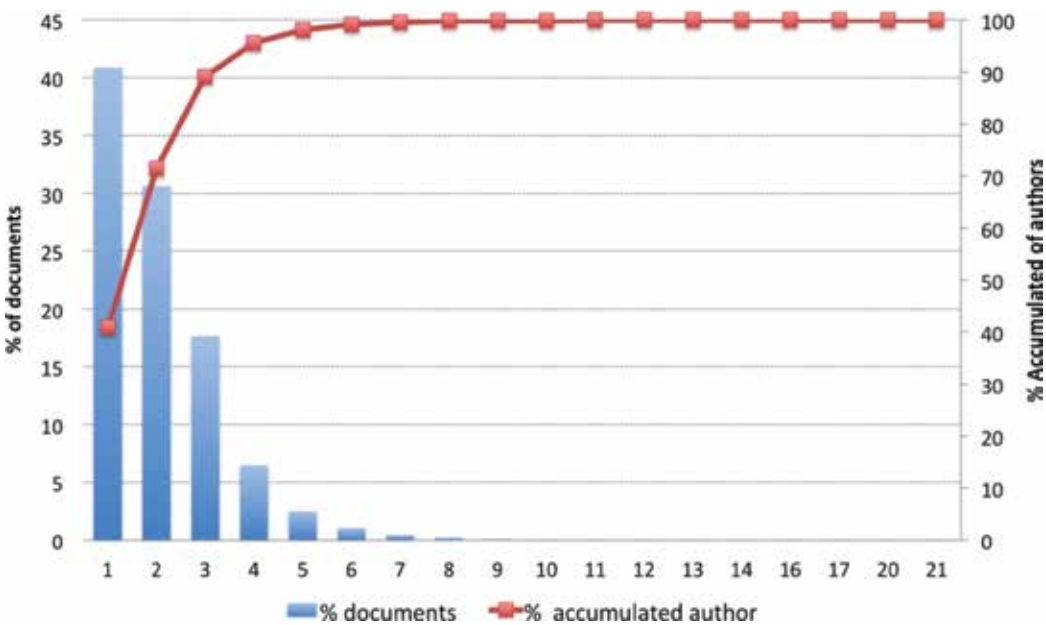


Figure 7. Co-authored distribution in urban studies (2000–2016).

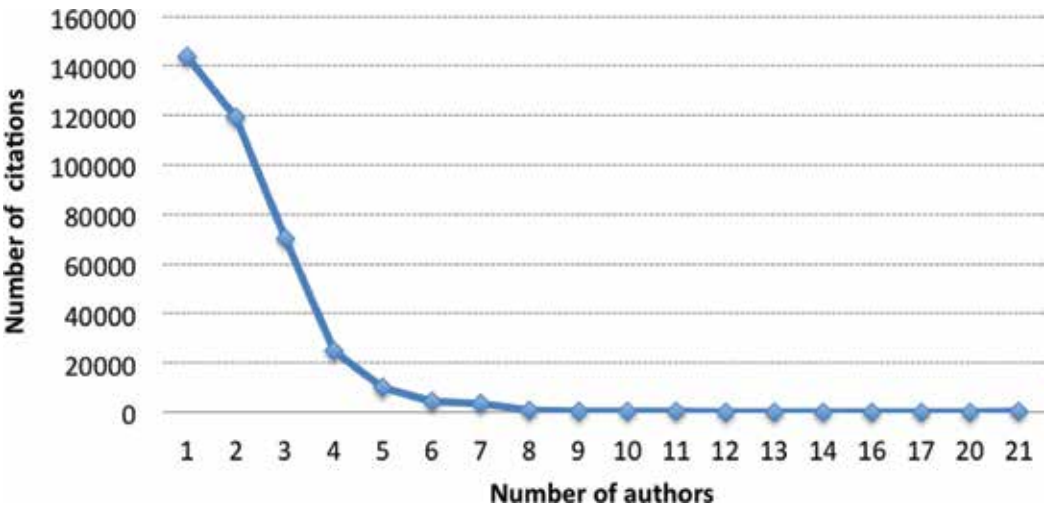


Figure 8. Citation by number of authors in urban studies 2000–2016.

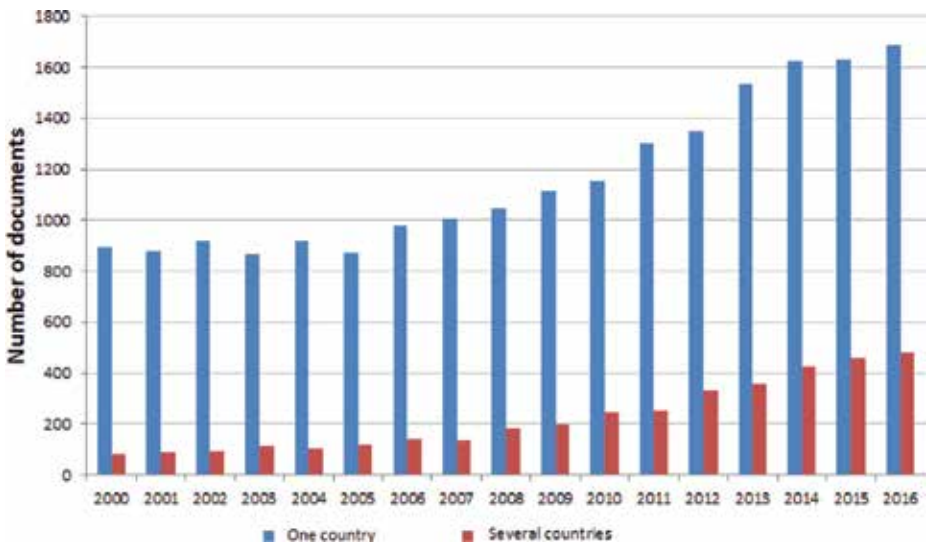


Figure 9. Diachronic international collaboration in urban studies 2000–2016.

not collaborate with other countries in the sample. The country with most co-authorship with other countries in the world is the *USA*, collaborating with 81 countries. A total of 49.62% of the countries (66) have relationships with a maximum of 5 other countries.

France, Canada, the Netherlands, England and the *USA* are the countries collaborating with more than 50 other countries in the category.

The network depicted in **Figure 10** shows a general view over the country network for *Urban Studies* considering all the period. The density of the network is 0.09808612, higher than the one found for *Demography* which also indicated that the network is dense for social sciences. The average degree of the countries is 12.9473.

There are 861 collaborations detected, most of them being anecdotal; 44.83% of these collaborations appear only once which means that these 2 countries only co-write 1 document in the whole period. The most prolific relationships among countries are found to be between *People's Republic of China* and the *USA, Canada* and the *USA* and *England* and the *USA* with more than 100 documents shared by each.

3.3. Comparison between the categories demography and urban studies

Comparing the number of journals indexed in JCR for the two categories analysed, it can be seen that *Demography* accounts for 68.4% of the number of journals for *Urban Studies* and its production only represents 47.36% of the second. In both categories, the percentages of multi-authored documents have very similar values with minor differences around 1% (**Table 3**).

Throughout 2000 and 2016, the citation in *Urban Studies* has been increasing with an exponential behaviour ($R^2 = 0.9712$) as well as the number of multi-authored articles ($R^2 = 0.8214$). However, the category *Demography* behaves differently, the increase in citations has a logarithmic behaviour ($R^2 = 0577$) and the number of articles written in collaboration represents a linear model ($R^2 = 0.9557$) (**Figure 11**). This relationship between the citations received and the number of multi-authored documents in the two categories (*Urban studies: Pearson's coeff.* 0.892, $p = .00$; *Demography: Pearson's coeff.* = 0.709, $p = .001$) is in agreement with that found in other studies in which it has been shown that co-authorship has a tangible effect on the impact of the citations [41, 42].

It is remarkable that, for *Demography*, DC values have always been higher than 0.5; in addition, DC values are similar to those obtained for some other research fields of social sciences such as *basic psychology* between 1926 and 2005 [43]. There is a linear dependency between

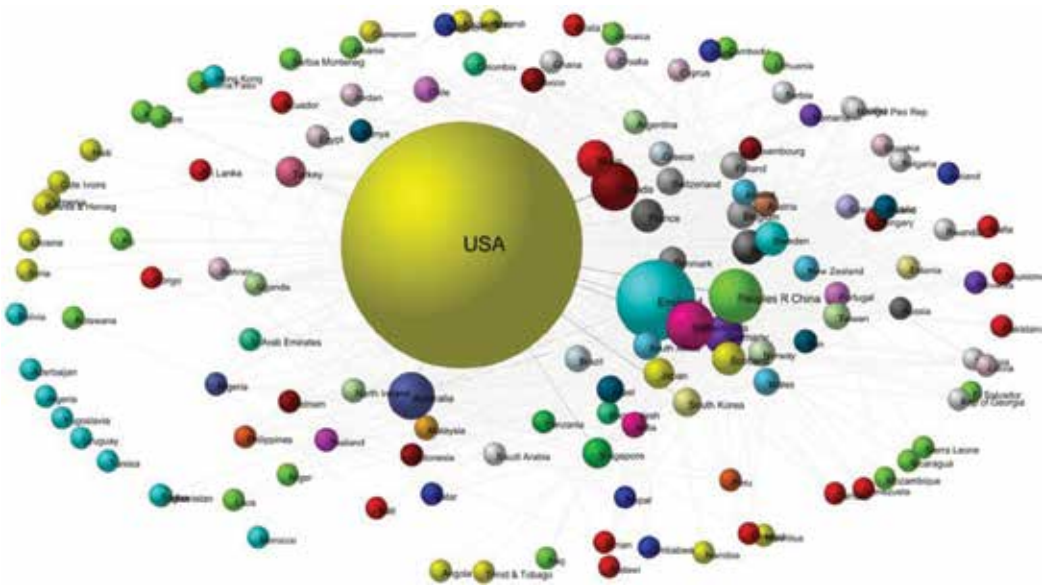


Figure 10. General network for international collaboration in urban studies 2000–2016.

Category	Demography	%	Urban Studies	%	Total
Journals	26	40.62	38	59.38	64
Documents	11,361	32.13	23,988	67.87	35,349
Multi-authored documents	6869	32.62	14,188	67.38	21,057
Single-authored documents	4492	31.40	9810	68.60	14,302
Authors per document	2.15		2.06		
DC	0.605		0.591		
CI	0.591		2068		
Citations	147,024	28.03	377,473	71.97	524,497
Citations/paper	12.94	87.25	15.73	14.83	14.83

Table 3. Demography versus urban studies multi-authorship (2000–2016).

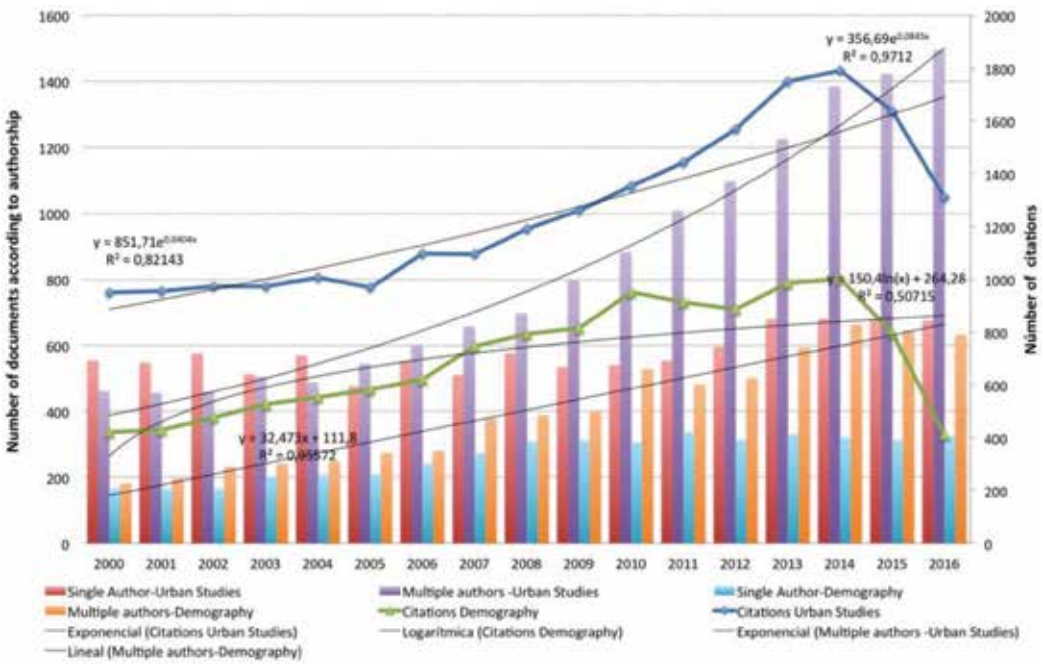


Figure 11. Citations and number of documents according to authorship.

DC values and citations which is found to be high for *Urban studies* category (*Pearson's coeff.* = 0.878, $p = .00$) and moderated for *Demography* category (*Pearson's coeff.* = 0.542, $p = .025$).

Focusing on the international collaboration, it is shown that both categories have similarities such as a high percentage of documents assigned to a single country. Moreover almost the half of the collaboration produced are no kept across time and frequently end in sporadic

connection between countries. The category *Urban Studies* present less countries despite the fact that it involved more documents than *Demography*.

Analysing DC values and citation in relation with international collaboration, it is found that the linear dependency between them is higher when international collaboration is involved, being the correlation coefficients 0.922 (p = .00) for *Urban studies* and 0.933 (p = .00) for *Demography*.

The USA is the most prolific country in both categories, whereas minority countries or countries of emerging economies are residual in *Demography* but relative important in the category *Urban Studies*. This could be explained by the fact that *Urban Studies* it indexed more journal in the sample with a wider regional scope.

The ranking of the most productive and collaborating countries is clearly dominated by English countries in both categories.

4. Conclusions

It has been verified that, in the period 2000–2016, there is a predominance of documents written in multi-authorship in the categories *Demography* and *Urban Studies*. Likewise, the number of documents in collaboration has been increasing proportionally to the total production. The highest values in the collaboration indicators, DC, CI have been reached in the most recent years, showing a tendency to continue increasing. This increase in the number of citations in relation to the increase in the number of authors per article shows a similar pattern to those found for other branches of knowledge closer to the hard sciences.

Despite these results, the international collaboration is not so high, compared to author collaboration, which means that a great portion of the multi-authored documents are written by authors affiliated to institutions of the same country.

The analysis of the scientific production of these two scientific categories in social sciences, *Urban studies* and *Demography*, has confirmed the findings of previous studies [44, 45] stating international collaboration in science is growing rapidly. This international collaboration has a correlation with the increase in the citation of multi-authored publications. The internationalisation of science in these two categories is largely due to the collaboration of researchers from the USA, England and Canada.

Author details

Alexander Maz-Machado* and Noelia Jiménez-Fanjul

*Address all correspondence to: ma1mamaa@uco.es

University of Cordoba, Córdoba, Spain

References

[1] Moed HF. Citation Analysis in Research Evaluation. Dordrecht, The Netherlands: Springer; 2005

[2] Vinkler P. The Evaluation of Research by Scientometric Indicators. Cambridge: Chandos Publishing; 2010

[3] Beyers J, Eilising R, Maloney W. Researching interest group politics in Europe and elsewhere: Much we study, Little we know? West European Politics. 2008;**31**(6):1103-1128

[4] Whitley R. Changing governance of the public sciences. In: Whitley R, Gläser J, editors. The Changing Governance of the Sciences the Advent of Research Evaluation Systems. Dordrecht: Springer; 2007. pp. 3-12

[5] Gläser J. The social orders of research evaluation systems. In: Whitley R, Gläser J, editors. The Changing Governance of the Sciences the Advent of Research Evaluation Systems. Dordrecht: Springer; 2007. pp. 245-266

[6] van Raan A. Measuring science. In: Moed HF, Glänzel W, Schmoch U, editors. Handbook of Quantitative Science and Technology Research. Dordrecht: Kuwer Academic Publishers; 2004. pp. 19-50

[7] Gupta DK. Scientometric study of biochemical literature of Nigeria, 1970-1984: Application of Lotkas's law and the 80720-rule. Scientometrics. 1989;**15**(3-4):171-119

[8] Schmoch U, Schubert T. When and how to use bibliometrics as a screening tool for research performance. Science and Public Policy. 2009;**36**(10):753-762

[9] Miguel S, Moya-Anegon F, Herrero-Solana V. Aproximación metodológica para la identifcación del perfil y patrones de colaboración de dominios científicos universitarios. Revista Española de Documentación. 2006;**28**(1):34-53

[10] Leydesdorft L. Caveats for the use of citation indicators in research and journal evaluations. Journal of the American Society for Information Science and Technology. 2008;**59**(2):278-287

[11] Morrison PS, Dobbie G, McDonald FJ. Research collaboration among university scientists. Higher Education Research and Development. 2003;**22**(3):275-296

[12] Katz JS, Martin BR. What is research collaboration? Research Policy. 1997;**26**(1):1-18

[13] Whitley R. The Intellectual and Social Organisation of the Sciences. Oxford, England: Oxford University Press; 1984

[14] Adams JD, Black GC, Clemmons JR, Stephan P. Scientific teams and institutional collaboration: Evidence from U.S. universities, 1981-1999. Research Policy. 2005;**34**:259-285

[15] Lee S, Bozeman B. The impact of research collaboration on scientific productivity. Social Studies of Science. 2005;**35**(5):673-702

[16] Wagner CS. Six case studies of international collaboration in science. Scientometrics. 2005;**62**(1):3-26

[17] Bordons M, Gómez I. Collaboration networks in science. In: Cronin B, Atkins HB, editors. The web of knowledge. A Festschrift in Honor of Eugene Garfield. Medford, NJ: Asis; 2000. pp. 233-250

[18] Glänzel W. Modelling and measuring multilateral co-authorship in international scientific collaboration. Part I: Development of a new model using a series expansion approach. Scientometrics. 1997;**40**(3):593-604

[19] Chen TJ, Chen YC, Hwang SJ, Chou LF. International collaboration of clinical medicine research in Taiwan, 1900-2004: A bibliometric analysis. Journal of the Chinese Medical Association. 2007;**70**(3):110-116

[20] Kundra R, Tomov D. Collaboration patterns in indian and Bulgatian epidemiology of neoplasms in Medline for 1966-1999. Scientometrics. 2001;**52**(3):519-523

[21] Glänzel W. International collaboration: Will it be keeping alive east European research? Journal of Intelligent Information Systems. 2006;**7**(1):247-254

[22] Winterhager M. International collaboration of three east European countries with Germany in the sciences, 1980-1989. Scientometrics. 1992;**25**(2):219-228

[23] Bordons M, Gómez I. La actividad científica española a través de indicadores bibliométricos en el período 1990-93. Revista General de Información y Documentación. 1997;**7**(2):69-86

[24] Bordons M, González-Albo B, Díaz-Faes AA. Colaboración científica e impacto de la investigación. In: González-Alcaide G, Gómez J, Agulló V, editors. La colaboración científica: una aproximación multidisciplinar. Valencia: Nau llibres; 2013. pp. 169-181

[25] Ardanuy J. Scientific collaboration in library and information science viewed through the web of knowledge: The Spanish case. Scientometrics. 2102;**90**(3):877-890

[26] Chaudhry AS. Collaboration in LIS education in Southeast Asia. New Library Word. 2007;**10**(1/2):23-31

[27] Hart R. Funded and non-funded research: Characteristics of authorship and patterns of collaboration in the 1986 library and information science literature. Library and Information Science Research. 1990;**12**(1):71-86

[28] Sin SCJ. Longitudinal trends in internationalisation, collaboration types, and citation impact: A bibliometric analysis of seven LIS journals (1980-2008). Journal of Library and Information Studies. 2011;**9**(1):27-49

[29] Sugimoto CR. Collaboration in information and library science doctoral education. Library and Information Science Research. 2011;**33**(1):3-11

[30] Maz-Machado A, Jiménez-Fanjul N, Madrid MJ. Collaboration in the Iberoamerican journals in the category Information Science & Library Science in WOS. Library Philosophy and Practice (e-journal). 2015. Paper 1270

[31] Katz JS, Hicks D. How much is a collaboration worth? A calibrated bibliometric model. Scientometrics. 1997;**40**(3):541-554

[32] Narin F, Whitlow ES. Measurement of Scientific Co-Operation and Co-Authorship in CEC-Related Areas of Sciences. Luxemburg: Office for Official Publications in the European Communities; 1990

[33] Wang L, Thijs B, Glänzel W. Characteristics of international collaboration in sport sciences publications and its influence on citation impact. Scientometrics. 2015;**105**(2):843-862

[34] Lawani SM. Quality, Collaboration and Citations in Cancer Research: A Bibliometric Study [thesis]. USA: Florida State University; 1981

[35] Subramanyam K. Bibliometric studies of research collaboration: A review. Journal of Information Science. 1983;**6**(1):33-38

[36] Ajiferuke I, Burrel Q, Tague J. Collaborative coefficient: A single measure of the degree of collaboration in research. Scientometrics. 1988;**14**(5-6):421-433

[37] Clarivate Analytics. 2016 Journal Citation Reports® Social Sciences Edition [Internet]. 2017. Available from: https://www.recursoscientificos.fecyt.es/factor/getJCR.php [Accessed: 01-07-2017]

[38] Cronin B, Overfelt K. Citation-based auditing of academic perfomance. Journal of the American Society for Information Science. 1994;**45**(2):61-71

[39] Maz-Machado A, Jiménez-Fanjul N, Villarraga M. La producción colombiana SciELO: Un análisis bibliométrico. Revista Interamericana de bibliotecología. 2016;**39**:15-26

[40] Batagelj V, Mrvar A. Pajek (Version 4.08) [Computer program]. 1996/2016. Available from: http://mrvar.fdv.uni-lj.si/pajek/

[41] Franceschet M, Costantini A. The effect of scholar collaboration on impact and quality of academic papers. Journal of Informetrics. 2010;**4**(4):540-553

[42] Leimu R, Koricheva J. Does scientific collaboration increase the impact of ecological articles? Bioscience. 2005;**55**(5):438-443

[43] Zafrunnisha N, Pullareddy V. Authorship and degree of collaboration in Psichology. Annals of Library and Information Studies. 2009;**56**:255-261

[44] Leydesdorft L, Wagner CS. International collaboration in science and the formation of a core group. Journal of Informetrics. 2008;**2**(4):317-325

[45] Cimini G, Zaccaria A, Gabrielli A. Investigating the interplay between fundamentals of national research systems: Performance, investments and international collaborations. Journal of Informetrics. 2016;**10**(1):200-211

Chapter 6

A Scientometric Study on Graphene and Related Graphene-Based Materials in Medicine

Nicola Bernabò, Rosa Ciccarelli,
Alessandra Ordinelli,
Juliana Sofia Somoes Machado, Mauro Mattioli and
Barbara Barboni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77288

Abstract

Here we carried out a scientometric analysis of scientific literature published referred to the use of graphene and graphene-based materials. We found that in the last 15 years, more than 1200 issues have been produced, with an *H*-index of 67 cited 2647 times. The countries that have a larger production, in terms of number of issues published, are China, the United States, South Korea, India, and Iran, and the most relevant subject categories in which they are indexed are materials science, chemistry, science and technology, physics, and engineering, while the biological and medical specialties seem to be actually not deeply involved.

Keywords: graphene, graphene-based materials, graphene oxide, nanotubes, biomaterials, medicine, bioengineering, scientometrics

1. Introduction

Graphene consists of a single layer of carbon atoms packed into a honeycomb lattice. Its particular atomic organization of the carbon atoms affords graphene a set of very unique characteristics that justify the attention researcher of all fields have given it. The more standing out properties are a high mechanical strength, thermal and electrical conductibility, high surface-to-mas ratio, and relative transparency [1]. Many studies use graphene oxide (GO) or reduced

graphene oxide (rGO) instead of pristine graphene, because the oxidized forms are easier to process and can be dispersed in water while at the same time maintaining most of graphene's properties.

Graphene oxide has shown great potential enhancing differentiation and proliferation of human stem cells in vitro, which tend to adhere to graphene plates. In particular, it favors differentiation of human neuronal stem cells (hNSC) toward neurons rather than glia cells [2]. Combined with its inherent flexibility and strength, the possibility of creating a 3D structure that mimics the original organ, graphene appears to be a great scaffold for stem cell-based therapy [3].

Furthermore, a lot of research has come forward regarding the use of graphene in biosensors. Compared to previously used materials, graphene shows increased resistance and sensitivity. Also, being biocompatible it can be worn, allowing for the possibility of a permanently used sensor. Additionally, graphene can be bound to a wide range of molecules and proteins that allow for better selectivity [4].

Another field to which graphene's ability to be bound to specific molecules has been applied is drug carrying and delivery. In particular, it has been successfully used for specific anti-cancer drug delivery [5]. It presents novel perspective in combining site detection and drug delivery. Peptides bound to the GO plates allow for detection by specific cell types, minimizing uptake by other healthy cells [6].

Graphene's use in the medical field raises a lot of questions regarding its safety and toxicity. In this regard, there are many conflicting studies and opinions. It appears that the matter of toxicity varies greatly depending on the physicochemical characteristics of the administrated graphene, also on the form of administration, and the model, varying between different species and cell types. The characteristics of graphene like concentration, dimensions (lateral and number of layers), surface structure functional groups, and protein corona influence its toxicity in biological systems. Despite its relevance to the effect, some toxicological studies do not give a proper characterization of the form of graphene used. Though most agree on the interaction of graphene with the cellular membrane, the question of its uptake is more controversial [7]. For example, the studies of Yue et al. on the viability of six different cell lines when treated with GO of varying dimensions show that only two phagocytic cell lines were able to internalize both nano- and micro-sized GO sheets. Furthermore, there was no difference in the viability of any of the six cell line studies when the concentration was lower than 20 μg/mL. On the other hand, inhalation of GO particles may lead to an accumulation in the pulmonary surfactant and initiate an inflammatory process [8].

Interestingly, although GO does not show to be absorbed through the gastrointestinal tract, a low dose of GO can cause more damage to the gastrointestinal surface being drank as a suspension than a high dose of GO [9]. Most toxic effects seem to surge from the use of high doses of GO and the sequential aggregation and formation of conglomerates than can block small blood vessels and result in dyspnea [10]. However, recent publications detect no pathological effects in mice exposed to low dosages of GO and functionalized graphene when administrated by intravenous injection [11].

While studying toxicity it is very important to analyze the effect on the reproductive system and development because this can lead to more lasting effects. Graphene plaques seem unable to penetrate the blood-testis barrier in mice, and therefore sperm function and male reproductive activity show no alteration even for high doses of graphene [12]. In the female, there are no alterations if GO is administered before mating or during early gestation, and the female can give birth to healthy litters. However, if administered during late gestation, it leads to abortion and even death of the pregnant mice for high dose [13]. Injection of chicken eggs leads to reduced vascularization of the heart [14]. Despite showing no obvious malformation or mortality in zebrafish embryo, GO aggregates were retained in many organelles leading to hypoxia and ROS generation in these areas [15].

Even though graphene toxicity has drawn a lot of attention from scientists, there is a remarkable lack of understanding of the mechanisms underlying this effect. The use of different models and forms of graphene seem to lead to very dissimilar conclusions. There is a clear need for more systematic and in-depth studies, before graphene can be brought to its full potential use [7].

In this context, it is evident that, in one hand, graphene and graphene-related materials are even more used in medicine and bioengineering; on the other one, the information about their safety, their toxicity, and about the way of their possible interaction with living being (and human body and fluids) are still incomplete.

For this reason, here, we carried out a scientometric study on this very interesting topic, with the aim to study the scientific literature and to identify the most relevant topic and the countries that are more involved in this research activity.

2. Materials and methods

2.1. Data collection and dataset

We accessed the data from Web of Science repository (https://apps.webofknowledge.com/) in December 2017–January 2018. The data have been filtered using the Advanced Search tool with the following syntax:

$$TS = (topic\ 1)\ AND\ TS = (topic\ 2) \tag{1}$$

where *TS* is the topic; *AND* is the Boolean operator.

In our queries, we used as topic 1 "graphene" or "graphene oxide" or "graphene-related material," combined with the following keywords as topic "medicine," "biomaterials," "scaffold," "regenerative medicine," or "bioengineering." Then all the data sets obtained were merged with the "Combine Sets" tool. As a result, we obtained a dataset in .txt format containing a list of 1208 articles with their attributes. All the following analyses have been carried out on this data set:

- **Number of citable issues:** are considered **e**xclusively articles, reviews, and conference papers.
- **Number of cites per documents:** it is the number of citation of documents published in specific years.
- **H index:** a topic/journal/author has index h if h of its N_p papers has at least h citations each, and the other ($N_p - h$) papers have no more than h citations each.

2.2. Temporal and geospatial analysis

The data were processed for temporal and geospatial analysis by Sci² Tool (Sci² Team). We generated temporal visualization of burst detection analysis of ISI keywords used in the papers. The geolocation of author collaboration was realized using Citespace (http://cluster.cis.drexel.edu/~cchen/citespace/) and Google Earth (https://www.google.it/intl/it/earth/).

3. Results and discussion

Overall, we found 1248 issues characterized by the bibliometric parameters shown in **Table 1**.

The number of issues published per year is described in **Figure 1**. As it is evident in the period 2009–2011, the number of papers published per year was very low (<10/year); then it increased with a linear trend, to reach about 350 issues published in 2017.

The time trend of citations (sum of cited per year) has a different pattern, described by more than linear pattern, as reported in **Figure 2**.

Interestingly, the distribution of cites per year, as shown in **Figure 3**, in keeping with the Bedford's, follows a power law, with a negative exponent.

In addition it has been possible to compute the main parameters of cites/year distribution (see **Table 2**).

To explore the temporal pattern of the most important themes studied, we analyzed the burst in citations referred to specific keywords (see **Table 3** for the list of citation bursts identified).

Parameter	Value
H-index	67
Average citation per item	17.65
Sum of time cited	2647
Citing articles	14,055

Table 1. Bibliometric parameters referred to the studied dataset.

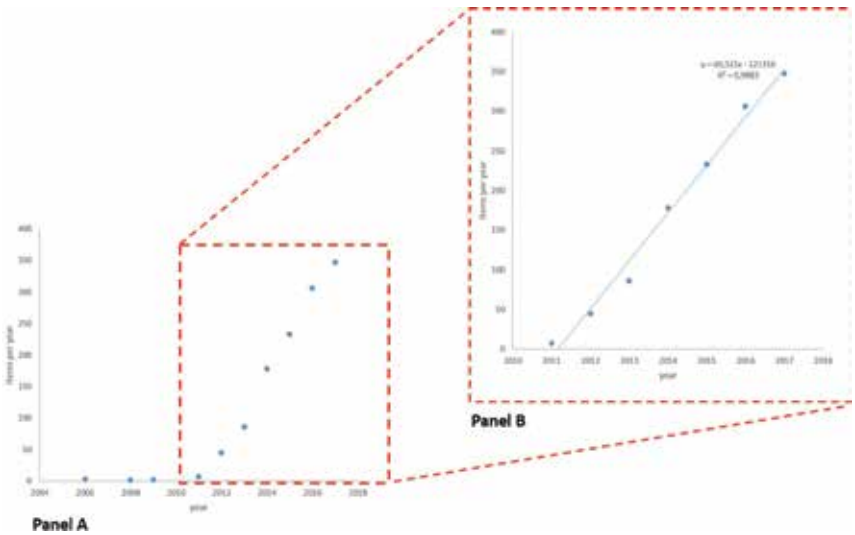


Figure 1. Graph showing the time trend of issues published per year.

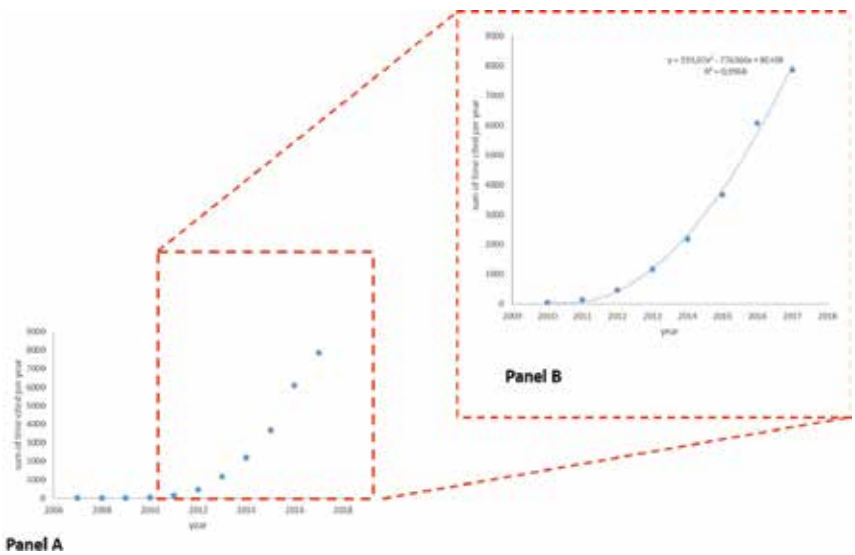


Figure 2. Graph showing the time trend of sum of cited per year.

Interestingly, we investigated the number of issues published by each country, thus estimating the contribution of different countries in research on graphene application in medicine (**Table 4**). These data demonstrate that graphene and graphene-based material are used in a wide variety of application in biomedicine such as cell and stem cell culture, translational medicine, bioengineering, toxicology, and development, thus confirming that these materials are becoming to represent a reality in life sciences.

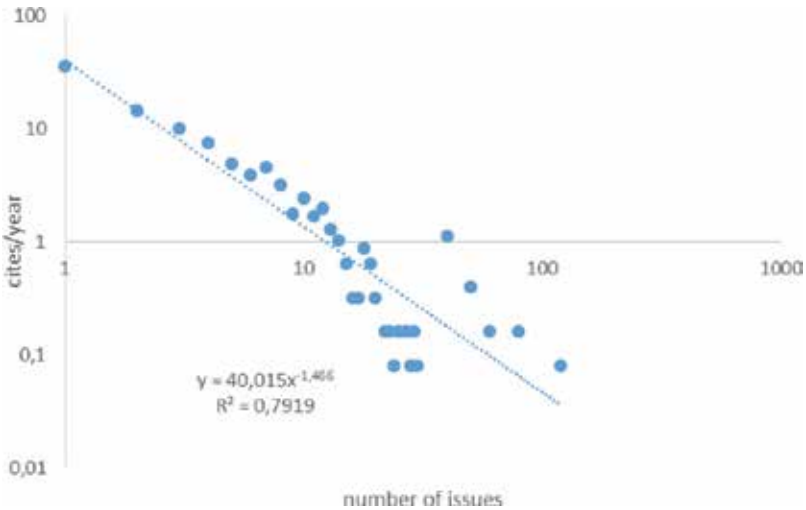


Figure 3. Graph showing the distribution of cites/year.

Parameter	Value
Max	117
95° percentile	17.96825
75° percentile	6
Median	2.25
25° percentile	0.666667
5° percentile	0
Min	0

Table 2. Citation parameters.

As it is evident, the most of issues have been published in China, with a total number of issues that accounts for about a third of worldwide production, followed by the United States and South Korea, India, and Iran. This datum is very interesting, because it demonstrates that Asiatic countries are the most important contributor, at least quantitative point of view, to this such important field of research.

To better explore the context to which the research is referred, we assessed the subject categories, as reported in **Table 5**.

As it is evident from the analysis of **Table 5**, the most of issues are indexed in nonbiological fields (materials science, chemistry, science and technology, physics, and engineering) rather than in biological fields. This seems to indicate that, to date, the research is led and defined by hard science scientist and, possibly, the contribution of researched belonging to biological and medical areas could be markedly increased in next years.

Term	Span	Weight	Begin	End
Accelerated-differentiation	2	30.562	2013	2014
Erk1–2	2	46.299	2012	2013
Evidenced-by	2	33.865	2015	2016
Fe3o4-go	3	36.129	2011	2013
Fe3o4-go-nanocomposites	3	27.411	2011	2013
Fe3o4-nanoparticles	2	36.777	2010	2011
Film-is	2	28.366	2012	2013
Films-of	5	31.674	2010	2014
Films-were	3	35.185	2011	2013
Films-with	3	31.612	2011	2013
Functional-theory	5	26.907	2006	2010
G-and	2	39.214	2011	2012
G-and-go	2	41.282	2011	2012
Genotoxicity-of	2	27.537	2012	2013
Graphene-content	2	43.921	2014	2015
Graphene-films	4	5.692	2009	2012
Graphene-nanocomposites	4	33.876	2011	2014
Graphene-nanoflakes	4	32.128	2011	2014
Graphene-nanostructures	2	36.785	2013	2014
Added-to	5	35.578	2009	2013
Graphene-sheets	8	130.541	2006	2013
Graphene-using	2	27.779	2013	2014
Graphite-oxide	3	43.642	2011	2013
Growth-of	3	48.996	2013	2015
Hectorite-clay	2	29.144	2013	2014
Adhesive-performance	2	38.487	2011	2012
Human-neural	4	41.777	2011	2014
Human-neural-stem	4	40.028	2011	2014
Human-neural-stem-cells	4	33.938	2011	2014
Adsorption-on	8	2.73	2006	2013
Indicates-that	2	26.756	2014	2015
Induction-of	2	32.246	2012	2013
Interaction-between	4	27.177	2010	2013
Investigated-using	3	44.764	2012	2014

Term	Span	Weight	Begin	End
Ag-nanoparticles	3	9.15	2010	2012
Mammalian-cells	2	32.251	2010	2011
Medical-research	2	92.302	2013	2014
Metabolic-activity	2	32.637	2013	2014
Mineralization-of	2	30.056	2014	2015
Modified-electrode	3	39.563	2010	2012
Molecular-dynamics	8	38.142	2006	2013
Monitoring-of	2	36.127	2012	2013
Multi-walled	2	34.922	2012	2013
Neural-stem	3	56.255	2011	2013
Neural-stem-cell	3	27.096	2011	2013
Neural-stem-cells	4	33.218	2011	2014
Nitrogen-doped	2	34.021	2014	2015
Oxidation-of	3	2.979	2010	2012
Antibacterial-activity	3	39.684	2010	2012
Peak-current	3	38.873	2010	2012
Porous-scaffolds	2	38.569	2015	2016
Prepared-via	3	29.212	2013	2015
Properties-of-graphene	4	34.088	2010	2013
Protein-corona	2	36.322	2014	2015
Rgo-ppy	4	28.174	2016	
Schwann-cells	2	38.247	2015	2016
Sheets-in	2	55.564	2013	2014
Sheets-in-the	2	31.928	2013	2014
Sheets-on	2	36.322	2014	2015
Similar-to-1	2	3.542	2013	2014
Size-dependent	2	29.383	2012	2013
Stabilizing-agent	2	32.246	2012	2013
Stem-cell-differentiation	4	26.513	2010	2013
Studied-by	4	28.884	2012	2015
Surface-chemistry	2	27.123	2013	2014
Time-dependent	2	26.413	2013	2014
Traditional-Chinese	2	40.822	2010	2011
Translational-medical	2	97.161	2013	2014

Term	Span	Weight	Begin	End
Translational-medical-research	2	92.302	2013	2014
Transmission-electron-microscope	2	27.348	2012	2013
Van-der	8	39.866	2006	2013
Van-der-waals	8	39.866	2006	2013
Walled-carbon-nanotubes	8	32.082	2006	2013
Water-molecules	2	52.214	2012	2013
Water-soluble	3	48.753	2012	2014
wt-wt	2	29.572	2012	2013
×-10	2	31.433	2010	2011
Beta-tcp	2	37.851	2015	2016
Bioactivity-of	3	32.757	2012	2014
2015-elsevier-b	2	76.506	2015	2016
bmp-2	2	122.945	2013	2014
Bone-cells	4	29.536	2011	2014
Bone-cement	2	29.572	2012	2013
Cancer-cells-and	2	43.062	2013	2014
Cancer-stem	2	59.119	2014	2015
Cancer-stem-cells	2	55.153	2014	2015
Carbon-nanotubes	5	83.764	2006	2010
Cell-differentiation	3	28.996	2012	2014
Cell-membranes	2	26.423	2014	2015
Cell-to	3	42.168	2011	2013
Cells-on-the	2	29.197	2013	2014
Cellular-uptake	2	27.088	2014	2015
Chemical-inducers	2	27.537	2012	2013
Chitosan-and	2	30.566	2010	2011
Chitosan-composite	2	27.779	2013	2014
Chitosan-film	3	27.411	2011	2013
Collagen-scaffolds	2	43.921	2014	2015
3d-rgo	4	39.778	2016	
3d-rgo-ppy	4	26.517	2016	
Composite-film	4	28.754	2011	2014
Composite-films	4	63.612	2011	2014
Concentration-of	2	59.312	2011	2012

Term	Span	Weight	Begin	End
Conductivity-of	2	29.085	2014	2015
Cultured-on-the	2	30.056	2014	2015
Cytotoxicity-of	3	48.985	2012	2014
Cytotoxicity-of-the	3	26.639	2012	2014
5-×-10	2	26.522	2010	2011
Delivery-of	2	61.189	2013	2014
Density-functional	5	26.907	2006	2010
Density-functional-theory	5	26.907	2006	2010
Der-waals	8	39.866	2006	2013
Differentiation-of-human	3	27.305	2011	2013
Doped-graphene	2	28.541	2013	2014
Embryonic-stem	3	37.831	2012	2014
Embryonic-stem-cells	3	31.442	2012	2014
Energy	4	27.199	2006	2009

Table 3. Citation bursts.

Number of issues	% on total issues published	Countries/territories
558	34.0	China
230	14.0	The United States
154	9.4	South Korea
75	4.6	India
69	4.2	Iran
41	2.5	Spain
39	2.4	Singapore
39	2.4	The United Kingdom
37	2.3	Australia
37	2.3	England
35	2.1	Taiwan
34	2.1	Italy
30	1.8	Japan
28	1.7	Germany
22	1.3	Canada
20	1.2	Saudi Arabia

Number of issues	% on total issues published	Countries/territories
18	1.1	Brazil
14	0.9	Poland
12	0.7	Romania
11	0.7	Denmark
11	0.7	Sweden
10	0.6	France
10	0.6	Russia
10	0.6	Turkey
9	0.5	Malaysia
9	0.5	Portugal
8	0.5	Egypt
7	0.4	Argentina
7	0.4	Czech Republic
7	0.4	Finland
6	0.4	Belgium
6	0.4	Switzerland
5	0.3	Israel
5	0.3	Thailand
4	0.2	Greece
4	0.2	Mexico
4	0.2	The Netherlands
4	0.2	Serbia
3	0.2	Ireland
3	0.2	Morocco
2	0.1	Pakistan
2	0.1	Scotland
2	0.1	Vietnam

Table 4. Number of issues per country.

The same trend could be identified looking on the WC, i.e., the classification system adopted by Web of Science (see **Figures 4** and **5**).

From these data, we could infer that we are seeing a first phase of the use of graphene and graphene-based materials, in which the studies on basic issues (synthesis, chemical characterization, description of chemical and physical properties) rather than the application in biology

Issues	Subject category
705	Materials science
583	Chemistry
423	Science and technology, other topics
222	Physics
161	Engineering
66	Electrochemistry
62	Polymer science
56	Biophysics
43	Biochemistry and molecular biology
35	Biotechnology and applied microbiology
33	Pharmacology and pharmacy
26	Environmental sciences and ecology
20	Energy and fuels
17	Instruments and instrumentation
17	Toxicology
14	Optics
12	Cell biology
9	Metallurgy and metallurgical engineering
8	Research and experimental medicine
4	Computer science
3	Crystallography
3	Dentistry, oral surgery, and medicine
3	Mechanics
3	Microscopy
3	Oncology
2	Food science and technology
2	Life sciences and biomedicine, other topics
2	Public, environmental, and occupational health
2	Spectroscopy
2	Water resources
1	Acoustics
1	Education and educational research
1	Endocrinology and metabolism
1	General and internal medicine
1	Genetics and heredity

Issues	Subject category
1	Hematology
1	Immunology
1	Information science and library science
1	Mathematical and computational biology
1	Medical informatics
1	Microbiology
1	Neurosciences and neurology
1	Nutrition and dietetics
1	Ophthalmology
1	Pathology
1	Physiology
1	Plant sciences
1	Radiology, nuclear medicine, and medical imaging
1	Telecommunications
1	Transplantation

Table 5. List of subject categories.

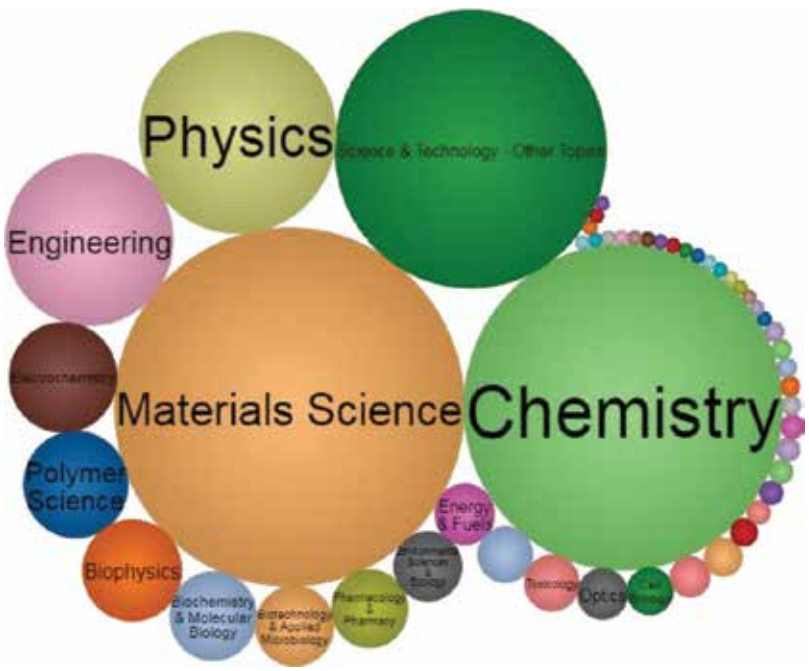


Figure 4. Classification of subject categories (the diameter is proportion to the number of issues).

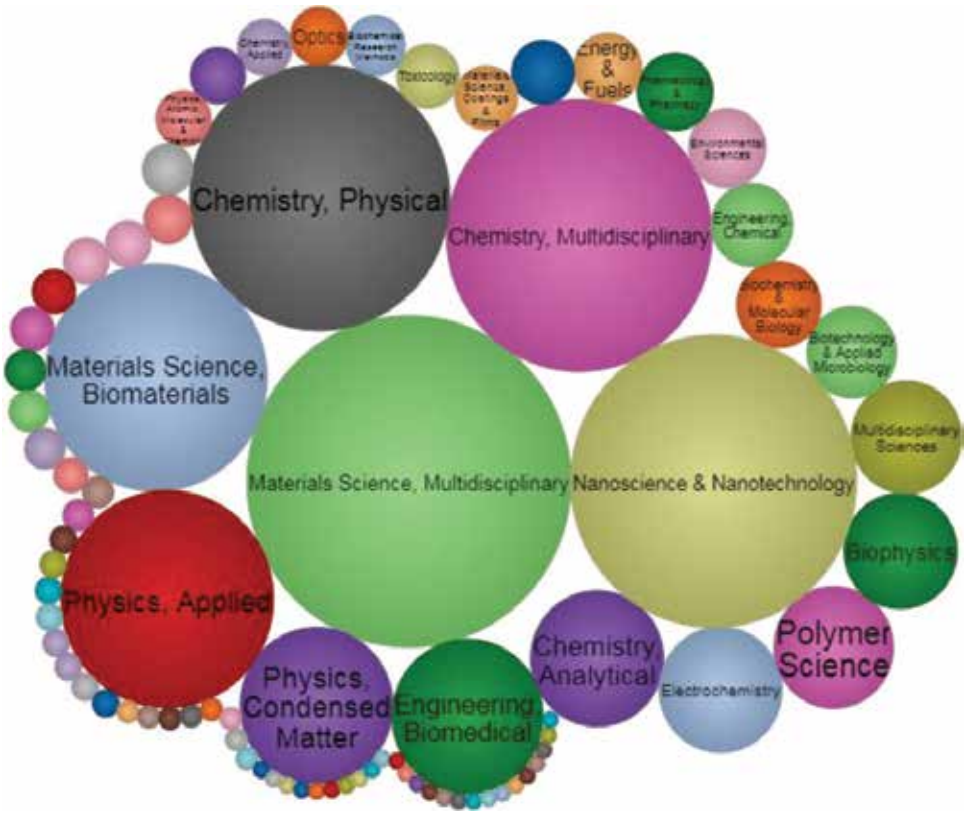


Figure 5. Classification of WoS categories (the diameter is proportion to the number of issues).

and medicine are predominating. Likely, it is possible to hypothesize that in next years, the contribution of life scientists and researchers and clinicians involved in medical field could acquire higher importance.

4. Conclusion

The use of graphene and graphene-based materials in biomedicine and bioengineering is an emergent technology that promises a wide variety of application in human health, diagnostics, and therapeutics. Here, for the first time, we carried out a scientometric analysis on this topic, finding as a result that the number of published issues and of their citations is quickly and markedly increasing, as proof of the intense activity in this field. The countries that display a more active production (in quantitative term) are from Asia (China, South Korea, India, and Iran) and from North America (the USA). The issues published are mainly referred to hard sciences (materials science, chemistry, science and technology, physics, and engineering) rather than biology or medicine. Despite that these materials are used in a wide variety of biomedical and bioengineering applications (from cell culture to stem cell

differentiation, from the realization of scaffolds to toxicological studies), the research activity on these issues seems still in an early stage, characterized by the physical and chemical characterization of materials, rather than the massive application in biomedicine and bioengineering.

Acknowledgements

Juliana Sofia Simoes Machado is granted by Rep-Eat-H2020-MSCA-COFUND-2015 No. 713714.

Conflict of interest

The authors declare that they have no competing interests.

Author details

Nicola Bernabò[1]*, Rosa Ciccarelli[1], Alessandra Ordinelli[1], Juliana Sofia Somoes Machado[1], Mauro Mattioli[1,2] and Barbara Barboni[1]

*Address all correspondence to: nbernabo@unite.it

1 Faculty of Bioscience and Technology for Food, Agriculture and Environment, University of Teramo, Italy

2 Istituto Zooprofilattico Sperimentale "G. Caporale", Teramo, Italy

References

[1] Mao HY, Laurent S, Chen W, Akhavan O, Imani M, Ashkarran AA, Mahmoudi M. Graphene: Promises, facts, opportunities, and challenges in nanomedicine. Chemical Reviews. 2013;**113**(5):3407-3424. DOI: 10.1021/cr300335p

[2] Park SY, Park J, Sim SH, Sung MG, Kim KS, Hong BH, Hong S. Enhanced differentiation of human neural stem cells into neurons on graphene. Advanced Materials. 2011;**23**(36):H263-H267. DOI: 10.1002/adma.201101503

[3] Li N, Zhang Q, Gao S, Song Q, Huang R, Wang L, Liu L, Dai J, Tang M, Cheng G. Three-dimensional graphene foam as a biocompatible and conductive scaffold for neural stem cells. Scientific Reports. 2013;**3**:1604. DOI: 10.1038/srep01604

[4] Ping J, Zhou Y, Wu Y, Papper V, Boujday S, Marks RS, Steele TW. Recent advances in aptasensors based on graphene and graphene-like nanomaterials. Biosensors & Bioelectronics. 2015;**64**:373-385. DOI: 10.1016/j.bios.2014.08.090

[5] Liu Z, Robinson JT, Sun X, Dai H. PEGylated Nanographene oxide for delivery of water-insoluble Cancer drugs. Journal of the American Chemical Society. 2008;**130**(33):10876-10877. DOI: 10.1021/ja803688x

[6] Park YH, Park SY, In I. Direct noncovalent conjugation of folic acid on reduced graphene oxide as anticancer drug carrier. Journal of Industrial and Engineering Chemistry. 2015;**30**:190-196. DOI: 10.1016/j.jiec.2015.05.021

[7] Ou L, Song B, Liang H, Liu J, Feng X, Deng B, Sun T, Shao L. Toxicity of graphene-family nanoparticles: A general review of the origins and mechanisms. Particle and Fibre Toxicology. 2016;**13**(1):57. DOI: 10.1186/s12989-016-0168-y

[8] Hu Q, Jiao B, Shi X, Valle RP, Zuo YY, Hu G. Effects of graphene oxide nanosheets on the ultrastructure and biophysical properties of the pulmonary surfactant film. Nanoscale. 2015;**7**(43):18025-18029. DOI: 10.1039/c5nr05401j

[9] Mao L, Hu M, Pan B, Xie Y, Petersen EJ. Biodistribution and toxicity of radio-labeled few layer graphene in mice after intratracheal instillation. Particle and Fibre Toxicology. 2016;**13**:7. DOI: 10.1186/s12989-016-0120-1

[10] Singh SK, Singh MK, Kulkarni PP, Sonkar VK, Grácio JJA, Dash D. Amine-modified Graphene: Thrombo-protective safer alternative to Graphene oxide for biomedical applications. ACS Nano. 2012;**6**(3):2731-2740. DOI: 10.1021/nn300172t

[11] Mu Q, Su G, Li L, Gilbertson BO, Yu LH, Zhang Q, Sun YP, Yan B. Size-dependent cell uptake of protein-coated graphene oxide nanosheets. ACS Applied Materials & Interfaces. 2012;**4**(4):2259-2266. DOI: 10.1021/am300253c

[12] Liang S, Xu S, Zhang D, He J, Chu M. Reproductive toxicity of nanoscale graphene oxide in male mice. Nanotoxicology. 2015;**9**(1):92-105. DOI: 10.3109/17435390.2014.893380

[13] Fu C, Liu T, Li L, Liu H, Liang Q, Meng X. Effects of graphene oxide on the development of offspring mice in lactation period. Biomaterials. 2015;**40**:23-31. DOI: 10.1016/j.biomaterials.2014.11.014

[14] Sawosz E, Jaworski S, Kutwin M, Hotowy A, Wierzbicki M, Grodzik M, Chwalibog A. Toxicity of pristine graphene in experiments in a chicken embryo model. International Journal of Nanomedicine. 2014;**9**:3913-3922. DOI: 10.2147/ijn.s65633

[15] Chen Y, Hu X, Sun J, Zhou Q. Specific nanotoxicity of graphene oxide during zebrafish embryogenesis. Nanotoxicology. 2016;**10**(1):42-52. DOI: 10.3109/17435390.2015.1005032

Chapter 7

Technology Roadmapping of Emerging Technologies: Scientometrics and Time Series Approach

Iñaki Bildosola, Rosamaría Río-Bélver,
Gaizka Garechana and Enara Zarrabeitia

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76675

Abstract

The present work is framed within tech mining and technology forecasting fields. It proposes an approach which combines a set of quantitative methods to completely describe an emerging technology, based on science, technology & innovation data. These methods are scientometrics, with which a customized and clean database is generated; hierarchical clustering to generate the ontology of the technology; principal component analysis, which is used to identify the main sub-technologies; time series analysis to quantitatively analyze the evolution of the technology, as well as future development; and technology roadmapping to integrate all the generated information in a single visual element. The results can be regarded as inputs for competitive technical intelligence activities, as they provide information about the past evolution of the technology, as well as potential future fields of application. The practical application of the approach, to BD technology, yields outcomes that allow conclusions to be drawn, such as how competitive intelligence, query processing and internet of things sub-technologies have been dominating the basic technology during the initial evolution, and how competitive intelligence and data communications systems will do so in the short-term future.

Keywords: technology roadmapping, technology forecasting, time series analysis, emerging technologies, scientometrics, big data

1. Introduction

This work aims to contribute to the fields of tech mining and technology forecasting (TF), based on science, technology & innovation (ST&I) data, from a quantitative methodological point of view. Tech mining aims to generate Competitive Technical Intelligence (CTI) using bibliometric and text mining (TM) software for analyses of ST&I information resources [1].

Meanwhile, TF can be generically defined as a prediction of the future characteristics of useful machines, procedures, or techniques [2]. The interrelation of both fields is proved by the fact that TF studies in companies are often called CTI [3].

Both activities (CTI and TF) are crucial for current enterprises, since they address organizational and cultural barriers to adopt and harness the potential of strategic emerging technologies. In fact, literature suggests that this is even more important for SMEs, since they are slow adopters of technology, often purchasing long after release and regularly dealing with technology handed down from other companies [4]. If a company, especially medium or small, does not succeed in the early adoption of an emerging technology, it can be irremediably surpassed by those competitors who did know how to adopt it correctly. Additionally, the TF field also includes more social and diffuse measurements. For example, governments use national foresight studies to assess the course and impact of technological change for the purposes of effecting public policy [3], and some studies are also used as an awareness-raising tool, alerting industrialists to opportunities emerging in S&T or alerting researchers to the social or commercial significance and potential of their work [5].

Within this framework, the importance of correctly structuring the ST&I information for a consistent analysis of a given technology should be underscored, as it facilitates the elicitation of meaningful implications by reducing the dimensions of original data and eliminating noise that normally exists in multivariate data [6]. Accordingly, any attempt to understand the main characteristics of a technology and to discover its future evolution based on ST&I data should go through three phases: the application of scientometrics in order to structure and prepare the data related to it; the use of TM techniques, making it possible to go beyond processing the content of the data and transforming it into information; exploit the generated information to forecast the future evolution of the technology by means of TF techniques.

Based on the above, the present work proposes an approach which makes use of tech mining and TF techniques for describing an emerging technology in full. Its application to a specific field or technology brings out information that can be regarded as inputs for CTI activities. It provides the structure of the technology, the dominating subfields throughout its evolution and the potential dominating concepts of short-term future. Besides, all the information is condensed and structured in a technology roadmap (TRM), which allows a complete depiction of the technology in a single visual item.

The work is divided as follows. Section two introduces the background of the work, paying attention to similar efforts that can be found in literature. Section three describes the proposed approach, going into the detail of the techniques on which is structured and their combination. Section four is used to apply the approach to a specific technology: big data (BD). Finally, in section five the applicability and validity of the approach is discussed and the future lines of work are described.

2. Background

The interconnection among CTI, TF and TRM activities is identified by means of the abundance of reference literature. In the 90s, Porter et al. proposed a method, called technology

opportunities analysis (TOA), which used ST&I data and bibliometrics with the purpose of identifying and assessing the implications of emerging scientific areas and new research technologies [7]. Following this path, Lee and Jeong used bibliometric data, co-word analysis, to generate a strategic diagram to be used for the analysis of the development trends of a specific technology domain [8]. Similarly, Lee et al. proposed a new TRM methodology to increase roadmapping effectiveness to support effective decision-making in new product and technology planning processes. The data source was patents and the method was founded on keyword-based product–technology maps, from which objective and quantitative information can be derived [9].

Latest efforts in this field are focused on the integration of more complex statistical methods and (semi)automatization proposals. In this regard, works can be found such as that proposed by Zhang et al. [10], in which a TRM composing method is described where data inputs are raw science textual data sources. The method seeks to identify macro-trends for R&D decision makers and is primarily based on a clustering-based topic identification model, a multiple science data sources integration model, and a semi-automated fuzzy set-based TRM composing model with expert aid. With similar goals, Joung and Kim propose technical keyword-based analysis of patents to monitor emerging technologies [11]. The approach includes the automatic selection of keywords and the identification of the relatedness among them. This task is based on the analysis of a technical keyword-context matrix, which is obtained by means of text-mining tools and techniques.

However, when it comes to introduce a consistent forecasting method based on ST&I data, there is a lack of time series analysis (TSA) methods. In terms of statistical methods, the most common approach for forecasting the future evolution of a technology based on bibliometric data is growth curve analysis (see [12] for further discussion). When it comes to combine scientometrics and TF, the inclusion of specific time series models is hardly encountered within the reference literature (see for example [13, 14]). What is more, the time series commonly take the frequency of generic items, such as patents or articles, as indicators without going down to a lower level, such as keywords, which provide richer information about the technology or field that is being analyzed. This kind of strategy is roughly chosen by Park and Jun [15] within the patent analysis field. Here, time series regression and clustering techniques are combined to construct a technological trend model of identified clusters, and that furthermore, these clusters are described by means of top keywords.

The following section describes the proposed approach, which is based on the combination of methods and techniques discussed here, in an attempt to identify an optimal combination of the most representative ones.

3. Research approach

As previously stated, the present approach combines a set of methods which belong to tech mining and technology forecasting fields. Namely:

- Scientometrics: to retrieve scientific publications related to an emerging technology and structure a customized database of the corresponding records.

- Text mining: to structure and clean the text of the records and to generate time series based on the analysis of the content.
- Hierarchical clustering: to uncover the sub-technology-based structure of the technology.
- Principal component analysis (PCA): to identify the fields of greatest research activity within the technology.
- Time series modeling and forecasting: to specify appropriate models for obtained time series and to obtain forecasts of the short-term development of the research activity related to the technology.
- Technology roadmapping: to merge all the information in a single visual item.

All the methods are interrelated, in the sense that the results of the application for some represent the input for others. All the methods described below are repeated twice in the full application of the approach. The first round analyzes the research related to the basic technology of the field that is being studied; whereas the second round is focused on the applications of it. This fact impacts directly on the first task, the retrieval of research publications. The data sources for this task are multidisciplinary online databases, whose online search tools are used to perform the query and set the required Boolean conditions. Thus, making use of a scientometrics approach, when it comes to retrieve data related to basic technology, terms such as 'based on…', 'application of…', 'using…' etc., have to be avoided; and only those research areas that are directly related to the technology should be included in the query. Conversely, when it comes to the applications, those terms are not restricted in the query and the research fields should be those in which the technology is presented as an application to improve features such as performance or efficiency. The objective fields of those publications are the title, abstract, publication date and keywords.

The data set is then processed by means of TM in order to clean and structure it. Those records which lack title, abstract, publication date or keywords are removed. Natural language processing (NLP) is applied to titles and abstracts to obtain meaningful words and phrases, and these terms are combined with the keywords in order to obtain a single list of significant terms, sorted by frequency of appearance. This list is subsequently treated with fuzzy logic to group all those terms which have equivalent meanings but are not written in exactly the same way into a single term. This task falls within the text summarization field and is largely used when it comes to condense large text data (see [16] for more discussion).

The obtained terms are the base to identify the structure of the technology research. They represent the hot topics and, by means of clustering techniques, the relationships between them can be identified. Thus, the application of a hierarchical clustering method to this data will provide the vertical structure of the technology in which the main fields of research, as well as the most important subfields, can be identified.

Once a static picture of the technology is obtained, it is time to analyze the dynamics, i.e. the evolution. First of all, main sub-technologies have to be identified, as the evolution of the technology as a whole will be based on the evolution of its most important sub-technologies. To do so, PCA is applied to the list of terms generated in the previous step. PCA is a basic

http://dx.doi.org/10.5772/intechopen.76675

method within factor analysis, which is a statistical approach that can be used to analyze interrelationships among a large number of variables, and to explain these variables in terms of their common underlying dimensions (factors or components) [17]. In the present case, it yields a number of components which are characterized by means of a vector of terms. These terms are grouped within the same component because they appear frequently together within the publications, and PCA identifies this fact. Thus, these components can be treated as sub-technologies, and the terms included in them as the main topics of within those sub-technologies (see [18] for PCA applications in text mining).

The evolution of the sub-technologies is subsequently obtained by means of time series. The generation of these series starts by splitting the previously obtained list of significant terms into months. This task is made possible because publication date of all the records is available and to which record each term belongs is also known. Thus, this split produces a set of sub-lists, each corresponding to each month of the analyzed time-range. Then a counting process is applied to generate the time series of each sub-technology. For example, if the vector of terms corresponding to sub-technology_1 is composed for three terms (term_1, term_2 and term_3), and these terms occur 2, 4 and 3 times respectively in the list of terms of a specific month, the value of the time series for that point in time is the sum of those frequencies: 9. This value is called the frequency of related terms (FRT), and represents the y-axis of the time series. If this counting process is repeated for all the months of the sample, a time series representing the evolution of each sub-technology is generated. This task is of utmost importance, as the time series is used as proxy for the intensity and trend of the activity related to a specific sub-technology.

In order to perform a consistent analysis of the evolution and forecasting, the time series has to be modeled. There is a range of models within the TSA field, and depending on the nature of the series, the simplest possible model that fits the data correctly and fulfills the objectives properly should be selected. In the case of the present work, as an initial approach, a linear time trend model (LTTM) [19] has been selected to model the last 3 years of the series, with which the trend of the series is consistently identified.

Finally, all the information previously generated is integrated into a TRM. The x axis is the temporal axis, defined by the time-range of the analysis. Whereas the y axis has two main layers: technology and application, each being completed with the information from each round of application of the approach, as described in the first task. These two vertical layers are in turn divided into sub-layers, which are directly the components of the first row of the vertical structure, obtained by means of hierarchical clustering. Once we have the TRM structured, it is filled year by year with those top terms contained in the list that comes from the text summarization task. In addition, these terms are grouped within each sub-technology, based on the corresponding vector of terms. Finally, there is room for short-term future, which will be completed with those terms that represent ascending sub-technologies. Logically, the ascending, maintained or decreasing nature is directly obtained from the time series modeling.

All these items are therefore integrated into a single visual element, full of information, the TRM. By means of this, the application of the approach aims to provide a mechanism to help experts forecast S&T developments within a specific area; or raise awareness among

practitioners concerning the characteristics and future potential applications and developments of emerging technologies.

4. Results and discussion

In order to test the applicability of the approach, and to analyze the outcomes obtained from its application, the whole approach was applied to a cutting edge technology, big data (BD). The definition of BD has evolved rapidly since the term was coined, which has caused some confusion. Gartner, Inc. gave a nice definition: "Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (Gartner IT Glossary (n.d.). The appearance of such a concept was driven by several facts. Among other things, the decrease in storage costs, which dropped from $14,000,000 (1980) to approximately $50 nowadays ($ per terabyte); the number of nodes a company might have, which have gone from 1(1969) to 1 billion hosts; and bandwidth costs, which was approximately $1200 in 1998 to the current $5 ($per Mbps) [20]. Thus, it is accepted that BD technology falls within the fields of computer science and mathematics, although it has been developed and applied in a myriad of fields, as we will see in the results of the approach.

All the tasks were applied interlaced, and partial and final outcomes were obtained. First of all, scientific publications were retrieved from the Web of Science (WOS) and Scopus databases. In order to establish the data time-range, the authors took into account what is considered as the "starting point" of BD technology research, a special issue of Nature on Big Data, in which it is distinguished from information and data science [21]. However, in order to considerate only those years in which the amount of publications was enough to analyze it from a time series point of view, the time-range was established in the range 2012–2016. The conditions imposed for the retrieving of the articles were based on similar works, in which was concluded that combining title and author keywords turned out to be the most relevant indicator in identifying related research on Big Data [22]. Thus, the term "Big Data" had to appear within the title and keywords. In the case of basic technology publications, only those within computer science and mathematics fields were allowed and those publications that contain the following terms were excluded: overview, review, based on big data, big data based, using big data, and big data application. A total of 6425 records were imported (WOS: 2740, SCOPUS: 3685). With regard to retrieving publications related to the applications of the technology, which is analyzed separately, the aforementioned excluded terms were permitted (save 'review' and 'overview'), and the allowed fields were all but computer sciences and mathematics. In this case, a total of 6864 records were imported (WOS: 3272, SCOPUS: 3592).

All the records were imported and merged in VantagePoint software (www.thevantagepoint.com). All the duplications and those records which lacked title, abstract, publication date or keywords were removed, finally obtaining a cleaned database of 5334 records for basic technology and 5991 for applications. NLP was then applied to titles and abstracts with which a set of terms was obtained. This allowed those concepts discussed within these fields to be identified. These terms were combined with those belonging to the keywords field in order to obtain

http://dx.doi.org/10.5772/intechopen.76675

a complete set of descriptors. At the end of the task, a list of 20,5010 terms was obtained for basic technology and 29,573 terms for applications. These terms were processed by means of fuzzy matching/grouping equal terms in a single item; as a result the list was reduced to 18,434 and 26,905 respectively.

Once the lists were generated, hierarchical clustering was applied to obtain the structure of the technology. To carry out this task R software was used, as it offers various algorithms to perform this clustering process. For the present work, Agnes package [23] with Ward clustering method was selected, which has been used in a wide range of work related to term grouping. It should be noted that the clustering process needs a distance-matrix as an input, and to do so it is necessary to generate the co-occurrence matrix of the terms, which is available in VantagePoint. This matrix describes how often each term appears jointly with each of the rest of the terms, and this is the basis for the clustering task. That obtained is directly the ontology of BD technology, in which the vertical structure can be identified. This information can be found in **Figure 1** in the case of basic technology and **Figure 2** in the case of applications. Regarding the content of the ontologies, the main difference between the structures of both should be stressed. In the case of technology there are four clear main sub-fields, which represent the most important areas of research in BD: distributed systems, data mining, machine learning and privacy. Whereas in the case of application of BD, this first line is much more varied, and eight main subfields can be found: machine learning, business intelligence, cloud computing, distributed storage, internet of things, web-based big data and e-healthcare. This is justified by the fact that BD is applied in countless fields. The hierarchical clustering shows this feature by generating a first line of the ontology with multiple subfields. A further analysis provides a deeper insight of the structure, in which various levels and more specific fields of research can be identified.

The application of the approach follows with the identification of the main sub-technologies and their evolution, by means of PCA analysis. This task is carried out in VantagePoint, which contains PCA functionality. The list of terms was once again used as an input, however, in this

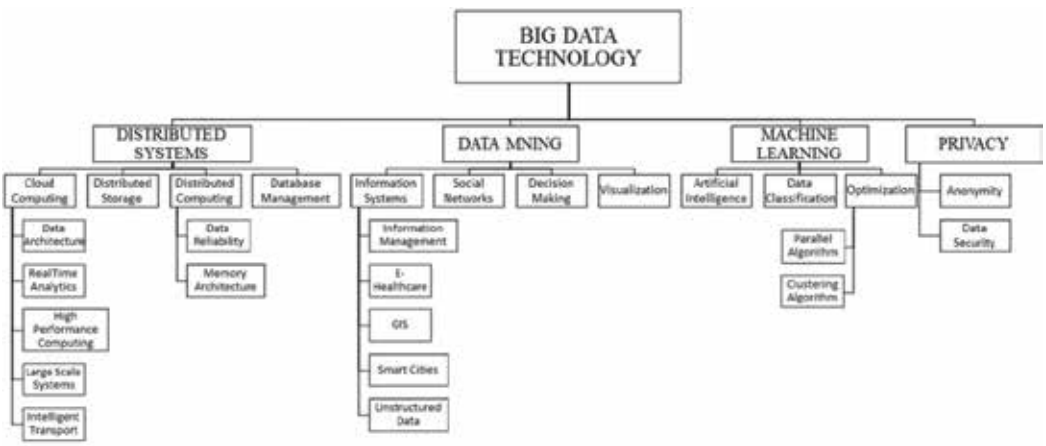


Figure 1. Big data technology ontology.

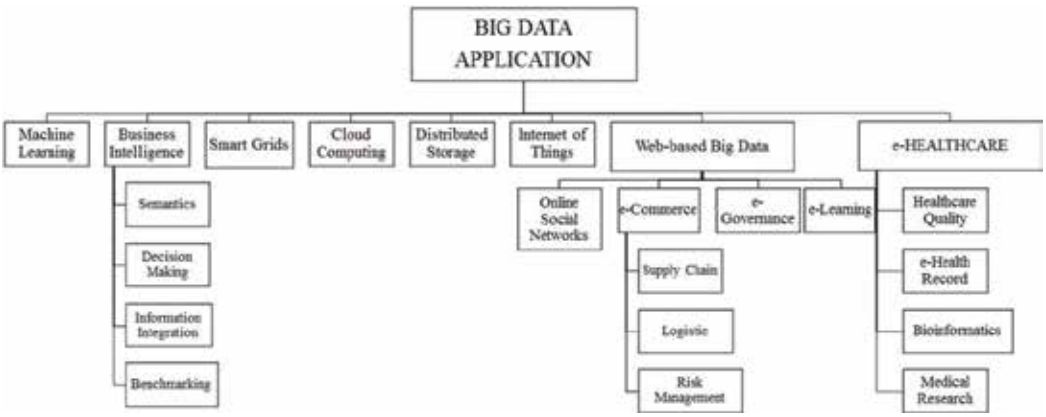


Figure 2. Big data application ontology.

Memory architecture	**Competitive intelligence**	**Learning Systems**	**Data privacy**	**Query processing**
Memory architecture Parallel architectures Program processors Parallel processing Data storage equipment Digital storage Computer hardware Network architecture Distributed storage Multiprocessing systems	Competitive intelligence Decision support system Business intelligent Decision support Decision making Management science Competition Information systems Competitive advantage Business Process	Learning systems Artificial intelligence Learning algorithms Machine learning Machine learning techniques Neural Network Deep learning Classification of information PCA Forecast	Data privacy Security of data Privacy Data security and privacy Privacy protection Privacy preserving Cryptography Privacy and security Mobile security Secure big data	Query processing Query language Query optimizer search engine Database System Computational linguistics Expert System Engines Information management Data integrity
Healthcare	**Data communication systems**	**Knowledge based systems**	**Internet of things**	**Data visualization**
Healthcare Medical computing Healthcare Hospitals Health Diagnosis Diseases Information science Medical images Data analytics	Data communication systems Data stream Stream big data Stream Computing Real time Data transfer Forestry Graphic methods Data handling	Knowledge based systems Knowledge base Semantic Web Ontology Semantic Natural language processing systems Information retrieval Extract information Knowledge extraction Knowledge management	Internet of things Internet Data reduction Data analysis Commerce Embedded systems Data acquisition Electronic commerce Cyber physical system Smart city	Data visualization Visualization Flow visualization Interactive visualization Big data visual Human computer interaction Visual analytics User interface Decision making Decision making process

Table 1. Big data basic technology top 10 components.

Internet of things	**Disaster prevention**	**Bioinformatics**	**Processing frameworks**	**Visual data**
Internet of things	Disaster	Bioinformatics	Processing	Visual data
Cyber physical systems	Disaster prevention	Biomedical	frameworks	Visuality
Embedded system	disaster management	engineering	Spark	Smart visual data
Industrial revolution	Emergency services	Biometrics	Map Reduce	Flow visualization
Network layers	Risk management	Alzheimer's disease	Computing	Three dimensional
Industry 4.0	Emergency management	Genetics	frameworks	computer graphics
Distributed computer	Online social network	Neuroimaging	Map Reduce	Information
systems	Risk perception	Genome	Hadoop	visualization
Ubiquitous computing	Social media	Biology	Open systems	Visual analytics
Manufacture	Data flow	Age	Information	Information system
Wireless		workflow	analysis	Big data visualization
telecommunication			Cluster	Data integrity
			computing	
			Open source	
			software	
Social big data	**Smart power grids**	**Machine learning**	**Energy efficiency**	**Traffic control**
Social network	Smart power grids	Machine learning	Energy efficient	Intelligent system
Natural language	Electric power	Artificial intelligence	Hardware	Traffic control
processing systems	distribution	Learning algorithms	Network	Intelligent transport
Online social network	Electric utilities	Natural language	architecture	system
Natural language	Electric power systems	processing	Energy	Traffic congestion
processing	Condition monitoring	Learning systems	conservation	Advanced technology
Machine learning	Electric power system	Online social	Computer	Motor transportation
Twitter	control	network	architecture	Vehicle
Sentiment analysis	Operation and	Classification of	Memory	Transportation
Recommender system	maintenance	information	architecture	Smart traffic control
Online learning	Data Processing Electric	Knowledge	System	Sustainable
Search engine	load forecasting	management	architecture	development
	Monitoring	Recommender	Energy	
	Electric power utilization	system	utilization	
		Forecast	Ecology	
			observatory	

Table 2. Big data application top 10 components.

case all the variables (terms) were grouped in components, and sorted by importance. Each component is represented by a vector of terms, which identifies the underlying topic. **Table 1** shows the main components of basic technology, interpreted as sub-technologies, and the top 10 terms for each. **Table 2** shows the same information in the case of applications. They are sorted by the explained variance, which means that the first contain more information about the complete original set of variables (terms). It should be noted that in order to keep as close as possible to the obtained quantitative results, the denomination of each component is always the corresponding first term, except in a few cases.

As shown, in the case of technology, even though the components were obtained from the content of publications directly related to basic technology research, topics which are actually applications of the technology can be identified. Once again, this is due to the characteristics of BD which, since the first research works, was already being applied to different fields. Thus, together with basic embryonic sub-technologies, such as memory architecture and data privacy, concepts like competitive intelligence or healthcare can be found, which are not strictly BD

foundational fields. As regards the components that belong to applications, logically these represent more specific fields, even though it might be another topic, the explained variance of each component is quite smaller than in the case of basic technology components. This means that the information is much more diversified, as expected when it comes to analyze the applications of a technology with the characteristics of BD. Lastly, it is worth mentioning the wealth of information contained in the vectors of each component. Consequently, by means of statistical techniques it is possible to identify such components, all of them with a high degree of homogeneity, and which show related and complementary concepts for different sub-technologies.

The utility of these components goes beyond their content, as a counting process to generate the corresponding time series - as previously described - can be applied. These series will provide complementary information, as they show both the intensity and the trend of each component, regarded as sub-technologies. As described in the approach's explanation, the y-axis values are measured in FRTs. Thus, those series with higher values represent those sub-technologies that have dominated the evolution of the technology in a given period of time. Additionally, the trends of the series provide meaningful information about how they have evolved throughout the analyzed period. Moreover, the trend for the last part of the series is valuable information allowing the future of the dominant and emerging sub-technologies to be forecast. However, whereas analysis of the FRT values can be done directly from the series, a consistent analysis of trends requires modeling, as this feature is not an observable component.

Figures 3 and **4** show the graphs of the top components (the complete set of values can be found in the Appendix). Note that the disparity in the range of values of the series prevents us from drawing all the graphs to the same scale. With regards to BD technology, the first analysis is centered on the levels of the series. In terms of absolute FRT values, attention should be paid to those components that have dominated the field throughout the years, which in this case are the sub-technologies of competitive intelligence, query processing and internet of things. The terms related to these have had a prominent presence, and therefore should be considered as key sub-technologies.

Additionally, which series started to present activity earlier in time can be analyzed. Thus, although all of them have a similar behavior, memory architecture and data visualization can be highlighted as those components that soon reached an important level of interest, within their range. These components can therefore be regarded as embryonic sub-technologies, since from the very beginning of the evolution of BD they started to have researchers and practitioners involved in their development. The same analysis for BD applications yields significant results. There is a clear dominant in terms of level values, social big data which, once activated, has values much higher than the rest. This indicates that it has attracted a lot of interest, directly related to its huge potential in a myriad of fields, ranging from marketing to customer relationship management (CRM). In terms of early starters, visual data is again one of those which started its activity earlier, together with processing frameworks. The latter, from the very beginning has been a field of interest, especially when it is approached from a benchmarking point of view, a fact confirmed by the data.

http://dx.doi.org/10.5772/intechopen.76675

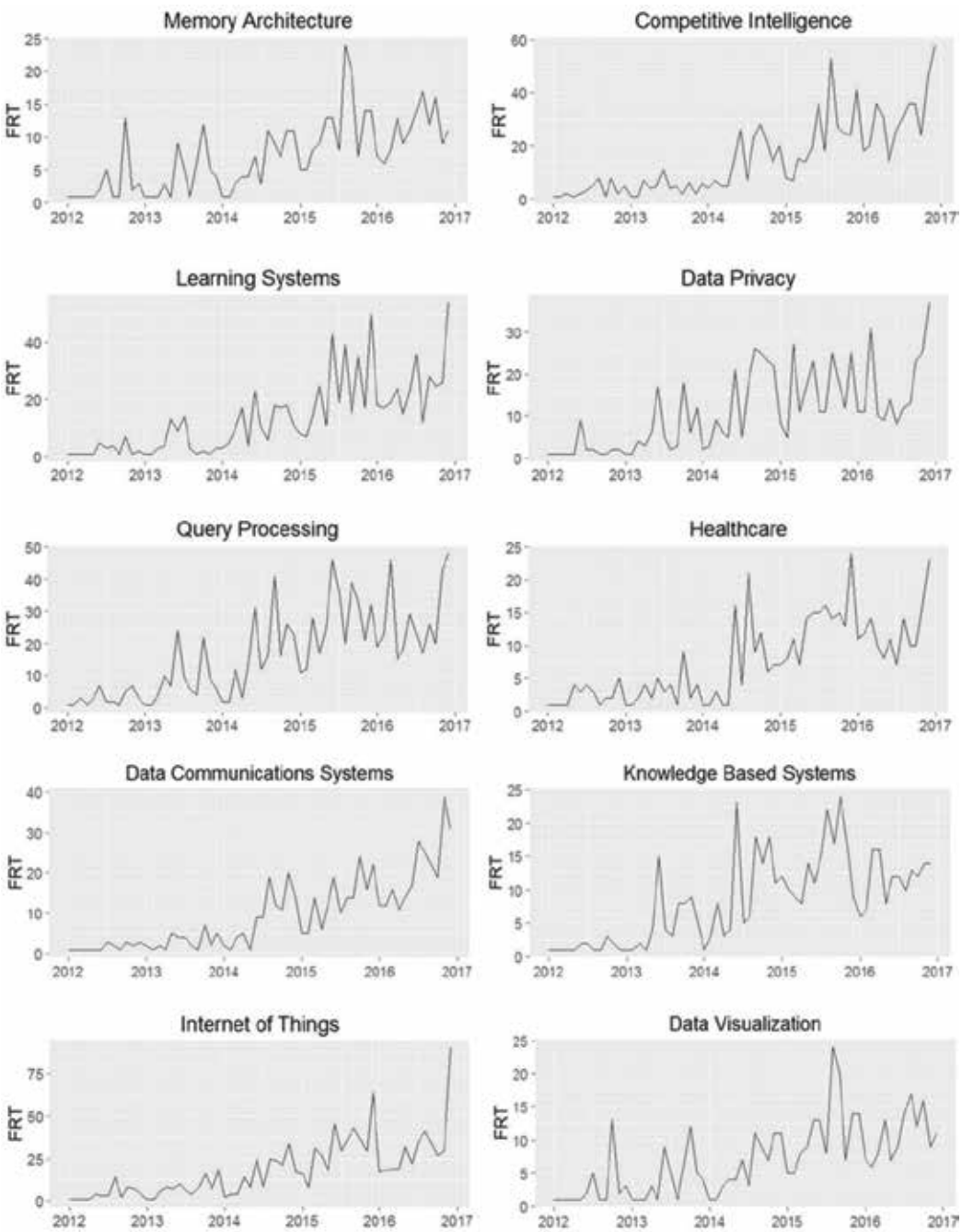


Figure 3. Time series graphs of big data basic technology top components.

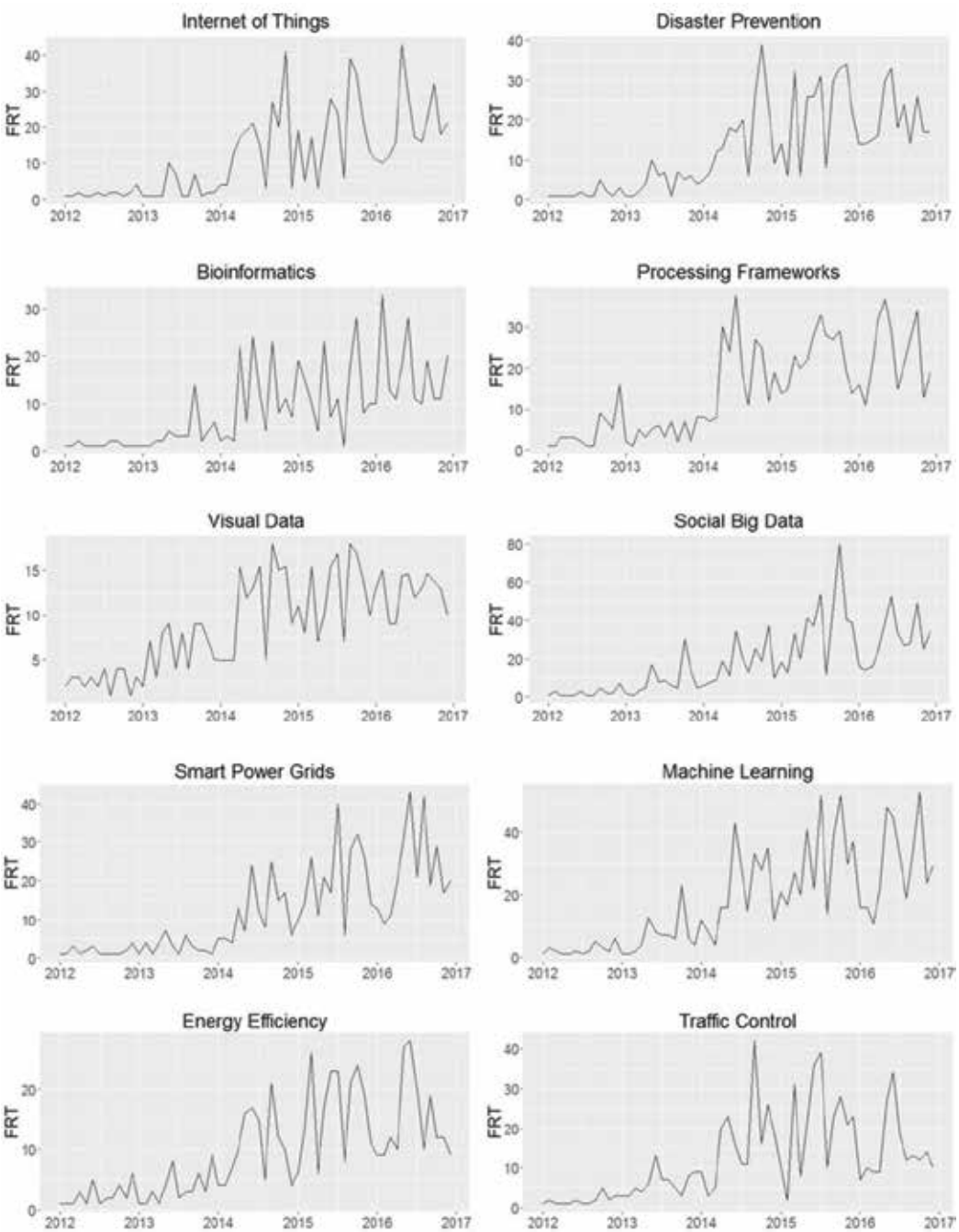


Figure 4. Time series graphs of big data applications top components.

The second part of the analysis is based on the modeling and trend identification of the series. As mentioned, the selected model was LTTM, and it was applied to the last 3 years of the series, since the goal was to identify the trend of the last phase of the evolution, in order to

Basic technology			**Applications**		
Sub-technology	R^2	Slope (p value)	Sub-technology	R^2	Slope (p value)
Memory architecture	0.35	0.032 (3.05e-04)	Internet of things	0.25	0.032 (1.09e-03)
Competitive intelligence	0.57	0.047 (1.08e-06)	Disaster prevention	0.10	0.019 (3.24e-02)
Learning Systems	0.40	0.042 (2.05e-05)	Bioinformatics	0.12	0.029 (2.23e-02)
Data privacy	0.16	0.028 (8.55e-03)	Processing frameworks	0.13	0.016 (1.94e-02)
Query processing	0.31	0.042 (2.26e-04)	Visual data	0.10	0.013 (3.71e-02)
Healthcare	0.37	0.052 (4.87e-05)	Social big data	0.27	0.031 (6.47e-04)
Data communication	0.52	0.059 (4.03e-07) 11	Smart power grids	0.31	0.034 (2.78e-04)
systems			Machine learning	0.17	0.024 (7.27e-03)
Knowledge based	0.19	0.029 (4.14e-03)	Energy efficiency	0.10	0.019 (3.17e-02)
systems Internet of	0.42	0.049 (1.03e-05)	Traffic control	0.23	0.028 (3.10e-04)
things Data Visualization	0.39	0.043 (3.07–05)			

Table 3. Parameter estimates and model validation of the main sub-technologies time series.

project it into the future. Thus, the model form is as follows: $\log(y_t) = a + bt + e_t$; where y_t represents the FRT value for a given month t = 1, 2, …, 36; a is the intercept of the model, which has no interpretation in the case of the present work; b represents the slope of the linear regression, which can be interpreted as the monthly percentage of growth of the series; and e_t represents the unexplained portion of the model, or term of error. The goodness of fit is given by the coefficient of determinations of the model (R^2), and the *p value* of the slope coefficient. If the series are observed it is clear that a linear model will not produce a good R^2 value, nevertheless, it is interesting that the *p value* of the slope coefficient is significant, since this is what is used as a proxy for the future projection. **Table 3** shows all the mentioned information for the complete set of time series.

As was expected, the R^2 values are not high enough to consider that the model is fitting the series tightly. The series present important variability and, logically, the linear model fails to follow it. However, trend identification by means of the slope value is statistically significant for all the cases at 5%. Based on these models, it is possible to analyze which sub-technologies are expected to raise more interest, and therefore develop further than others. Focusing on basic technology, the cases of data communication systems and healthcare should be noted, with a monthly percentage of increase of 5.9 and 5.2% respectively. The first is centered on issues arising from the management of communication of a huge quantity of data in the BD environment, and is apparently involving more people in its improvement. The second case, healthcare, has always been regarded as a promising field within BD technology, and the data show that it will gain importance in the short-term future. This is not the case for those that dominated the past years in terms of the series' absolute levels, memory architecture and data visualization, which with percentages of 3.5 and 3.9%, respectively have lost their dominance within the technology development.

In the case of applications, analysis of the values allows further conclusions to be drawn. Smart power grids (3.4%), internet of things (3.2%) and social big data (3.1%) are the ones with the highest trend values. All of them are growing faster than the rest of the sub-technologies and should be regarded as fields of great development. The case of social big data is even more remarkable, as it has also dominated the applications in terms of absolute

values, thus its great importance within BD applications is expected to increase. Once again, there are some sub-technologies that present lower increase values, such as energy efficiency, visual data and disaster prevention; all of them with a 10% value. Accordingly, these should be considered as fields that will gradually lose importance at the level of development and investment. In any case there is a general conclusion, which is the fact that the whole set of series present a positive trend value. This leads to a clear conclusion: BD as such is still increasing its importance among researchers and practitioners. It is still an emerging technology.

The final outcome of the approach is the TRM, in which all the previous partial results are integrated. What is more, the structuring and content of the TRM itself is conditioned by the partial results that have been obtained. The vertical structure is derived directly from the first level of the ontology in the case of the technology layer. This is not the case with the application layer, since the first line of its ontology had too many elements to sub-divide the layer based on them. Accordingly, the layer is presented without sub-divisions. The included terms are the most frequent terms, year by year, extracted from the list generated by means of the NLP task. It is required that terms exceed a certain level of frequency to be included in the TRM, and that is why more gaps appear during the initial years. In fact, it is from year 2014 when the TRM starts to be full of information, which coincides with the moment that the time series grew consistently. Furthermore, it is in the last years when the diversity of terms grows significantly, and consequently, the terms that describe more general concepts give way to others that represent more specific fields. The terms are grouped within the main sub-technologies identified above, and those terms that do not belong to any of these are placed loose. The vertical position of both the sub-technologies and loose terms, in the case of the technology layer, is based on the vertical structure of the TRM itself. Whereas for the application layer, as there is no such sub-division, placement is done by following the structure of the technology layer, as far as possible, to maintain a unified criterion throughout the TRM. Finally, the slope value of the models for each sub-technology is incorporated. The set of sub-technologies have been divided into five levels, from least to greatest slope, and have been painted accordingly with the following colors: gray; green; blue; orange; and red. Additionally, those with greater slopes have been extended further into the future, representing the probability of these being dominating fields in the short-term future. Thus, a third dimension has been added through the colors.

With regard to the content, the TRM provides a good summarization of the evolution of the technology characteristics. It can be seen how the first years show initial ideas that were developed within the different sub-technologies. For the technology layer, foundational terms such as distributed database systems in memory architecture and information management in competitive intelligence can be found. As time passes, more specific fields begin to appear, such as smart cities in internet of things and semantic web in knowledge based systems. Together with this, those topics within the fastest growing sub-technologies can be identified, which are candidates to have a strong presence in the short-term, such as business intelligence

in competitive intelligence, or diagnosis in healthcare. Similar behavior can be found in the application layer. Initially the TRM is filled with terms that refer to generalist fields, such as industry research in internet of things, MapReduce and Hadoop in processing frameworks or visual analytics in visual data. However, as you move forward in time, more specific ideas start dominating the roadmap, with examples such as industry 4.0 in internet of things and neuroimaging in bioinformatics. Finally, paying attention to emerging sub-technologies, attention should be paid to topics such as intelligent transport systems in traffic control, or sentiment analysis in social big data. All this information is presented in **Figures 5** and **6**, where the complete TRMs can be seen.

5. Conclusions and future work

The present work proposes an approach which makes use of tech mining and TF techniques for describing an emerging technology in full. The approach has been designed as a combination of quantitative methods through which various partial results are obtained, with which the technology analyzed is fully described. Within these methods, the main contribution is the idea of combining a more classical analysis based on scientometrics and common TM methods, such as clustering and text summarization; with less usual and more current methods such as PCA and especially TSA. Furthermore, technology roadmapping has been introduced to generate a final integrating element, in which all the information is aggregated. All this has permitted a fuller description of the technology, as well as a prospective exercise. To validate the applicability of the approach, it has been applied to BD technology, an emerging cutting edge technology. In that application, based on scientometrics analysis to generate a clean usable database, we have been able to apply the different methods with which the ontology of technology has been generated (hierarchical clustering method); and the main sub-technologies have been identified (PCA) (**Figures 5** and **6**).

Furthermore, a novel counting process has been presented to generate time series. These series have made it possible to understand the evolution of technology in detail. Additionally, they have been used to identify which sub-technologies have dominated the field throughout the years, and by means of a modeling process, which ones are expected to do so in the short-term future. It is at this point that it has been possible to identify that certain sub-technologies, such as memory architecture or energy efficiency, have shown limited growth in recent years, while others have accelerated their activity, with examples like competitive intelligence and smart power grids.

The results obtained come directly from the input data of the application: scientific publications. While more sophisticated results and deeper insights can be achieved on the analyzed technology, the aim has been to demonstrate that it is possible to generate such a powerful and information-filled element as the TRM by means of quantitative analysis of the data. In this sense, future lines of work should be directed towards the integration of more input data for the approach. In following with this, there are two elements that are being considered: patents

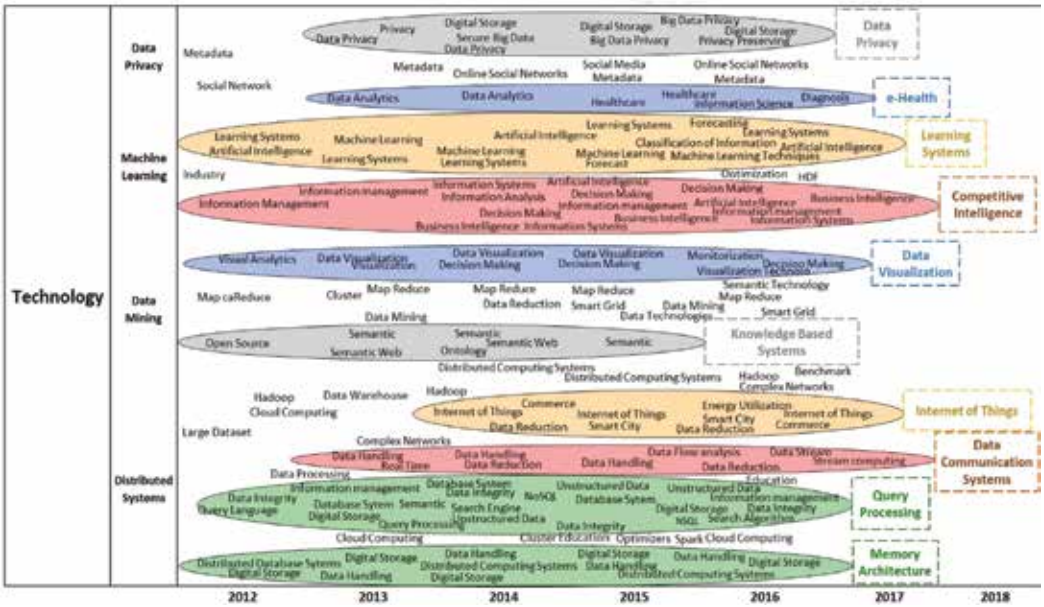


Figure 5. Technology roadmap of BD basic technology.

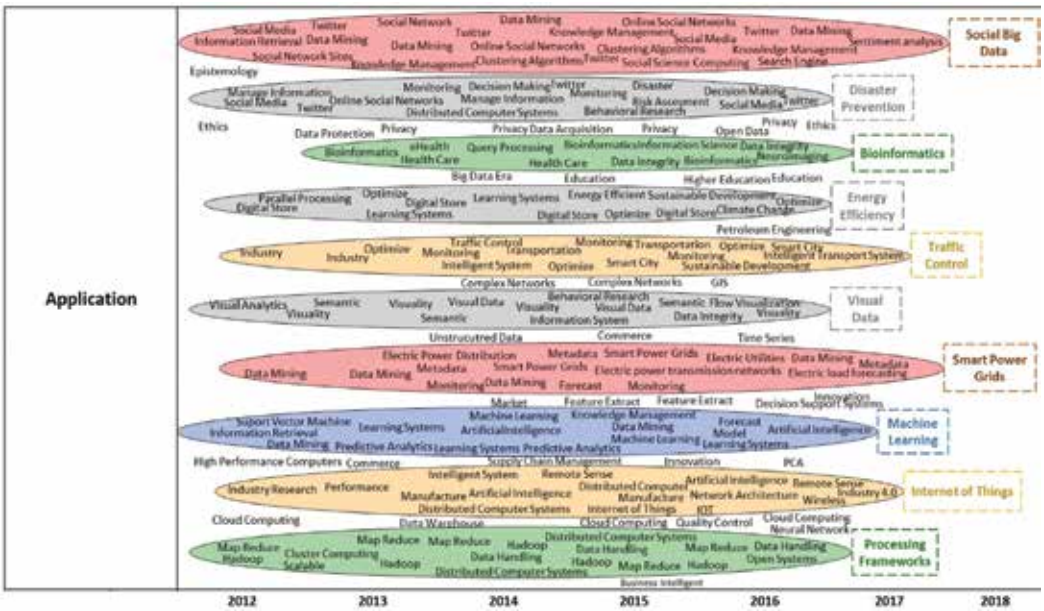


Figure 6. Technology roadmap of BD applications.

and web pages. The first will provide information about products or highly developed applications, while the webs will be used to analyze the technology at market level, based on web pages of enterprises that commercialize the technology. The same methods can be applied to these data and the results can be integrated by means of new layers in the TRM.

A. Appendix

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
01/2012	1	1	1	1	1	1	1	1	1	1
02/2012	1	1	1	1	1	1	1	1	1	1
03/2012	1	2	1	1	3	1	1	1	2	1
04/2012	1	1	1	1	1	1	1	1	1	1
05/2012	1	2	1	1	3	4	1	1	2	1
06/2012	2	3	5	9	7	3	1	2	1	2
07/2012	5	5	3	2	2	4	3	2	6	5
08/2012	1	8	4	2	2	3	2	1	1	1
09/2012	1	1	1	1	1	1	1	1	1	1
10/2012	13	8	7	1	5	2	3	3	4	13
11/2012	2	2	1	2	7	2	2	2	3	2
12/2012	3	5	2	2	3	5	3	1	4	3
01/2013	1	1	1	1	1	1	2	1	1	1
02/2013	1	1	1	1	1	1	1	1	1	1
03/2013	1	7	3	4	4	2	2	2	2	1
04/2013	3	4	4	3	10	4	1	1	3	3
05/2013	1	5	13	6	7	2	5	4	9	1
06/2013	9	11	9	17	24	5	4	15	6	9
07/2013	5	4	14	5	9	3	4	4	6	5
08/2013	1	5	3	2	6	4	2	3	10	1
09/2013	6	2	1	3	4	1	1	8	2	6
10/2013	12	6	2	18	22	9	7	8	6	12
11/2013	5	2	1	6	9	2	2	9	3	5

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
12/2013	4	6	3	12	6	4	5	5	12	4
01/2014	1	4	3	2	2	1	2	1	2	1
02/2014	1	7	5	3	2	1	1	3	1	1
03/2014	3	5	10	9	12	3	4	8	7	3
04/2014	4	5	17	6	3	1	5	3	7	4
05/2014	4	14	4	5	13	1	1	4	11	4
06/2014	7	26	23	21	31	16	9	23	13	7
07/2014	3	7	10	5	12	4	9	5	16	3
08/2014	11	23	6	18	16	21	19	6	14	11
09/2014	9	28	18	26	41	9	12	18	25	9
10/2014	7	22	17	25	16	12	11	14	15	7
11/2014	11	14	18	23	26	6	20	18	9	11
12/2014	11	20	11	22	23	7	14	5	14	11
01/2015	5	8	8	8	4	7	5	5	5	5
02/2015	1	7	7	5	5	3	5	2	4	1
03/2015	8	15	14	27	28	11	14	9	10	8
04/2015	9	14	25	11	17	7	6	3	13	9
05/2015	13	19	11	17	25	14	12	14	10	13
06/2015	13	36	43	23	46	15	19	11	26	13
07/2015	8	18	19	11	36	15	10	15	15	8
08/2015	24	53	39	11	20	16	14	22	39	24
09/2015	20	27	15	25	39	14	14	17	24	20
10/2015	7	25	35	18	33	15	24	24	30	7
11/2015	14	24	17	12	21	13	16	17	19	14
12/2015	14	41	50	25	32	24	22	9	29	14
01/2015	7	9	12	5	2	8	8	6	11	7
01/2016	6	12	11	5	7	12	8	7	10	6
02/2016	8	36	19	31	46	14	16	16	25	8
03/2016	13	31	24	10	9	10	11	16	24	13
04/2016	4	14	15	9	18	9	14	8	18	4

	Memory arch.	Competitive intelligence	Learning Systems	Data privacy	Query processing	Health care	Data comm. syst.	Knowledge based syst.	Internet of Things	Data visual.
05/2016	9	25	23	14	29	12	17	12	26	9
06/2016	14	30	36	8	23	13	28	12	18	14
07/2016	17	36	12	12	17	14	25	10	27	17
08/2016	12	36	28	13	26	10	22	13	31	12
09/2016	16	24	25	23	20	10	19	12	26	16
10/2016	9	46	26	25	43	16	39	14	35	9
11/2016	11	58	54	47	48	26	31	14	48	11

Table A1. Time series values of big data basic technology top components.

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
01/2012	1	1	1	1	1	1	1	1	1	1
02/2012	1	1	1	1	1	3	1	3	1	2
03/2012	2	1	2	3	2	1	3	2	1	1
04/2012	1	1	1	3	1	1	1	1	3	1
05/2012	1	1	1	3	1	1	2	1	1	1
06/2012	1	2	1	2	2	3	3	2	5	2
07/2012	1	1	1	1	1	1	1	1	1	1
08/2012	2	1	2	1	1	1	1	2	2	1
09/2012	3	5	2	9	2	5	1	5	2	2
10/2012	1	2	1	7	2	2	1	3	4	5
11/2012	2	1	1	5	1	2	2	2	2	2
12/2012	4	3	1	16	2	7	4	6	6	3
01/2013	1	1	1	2	1	2	1	1	1	1
02/2013	1	1	1	1	1	1	4	1	1	1
03/2013	1	2	2	5	2	3	1	2	3	3
04/2013	1	4	2	3	5	5	4	4	1	2
05/2013	9	10	4	5	2	17	7	13	4	4
06/2013	7	6	3	6	3	8	3	8	8	11

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
07/2013	1	7	3	3	4	9	1	7	2	5
08/2013	1	1	3	7	4	6	6	7	3	5
09/2013	8	7	14	2	1	5	3	6	3	3
10/2013	2	5	2	7	3	30	2	23	6	1
11/2013	2	6	4	2	10	14	2	6	3	6
12/2013	3	4	6	8	2	5	1	4	9	7
01/2014	6	5	2	8	1	6	5	12	4	9
02/2014	6	7	3	7	1	8	5	8	4	3
03/2014	10	12	2	8	3	9	4	4	7	5
04/2014	16	13	22	30	18	19	13	16	10	20
05/2014	19	18	6	24	11	11	7	16	16	23
06/2014	23	17	24	38	12	34	24	43	17	16
07/2014	19	20	12	18	18	20	12	28	15	11
08/2014	7	6	4	11	2	13	8	15	5	11
09/2014	27	27	23	27	17	25	42	33	21	42
10/2014	23	39	8	25	14	19	15	28	12	16
11/2014	41	23	11	12	18	37	17	35	10	26
12/2014	4	9	7	19	8	10	6	12	4	19
01/2015	24	14	19	8	10	18	10	21	6	10
02/2015	8	6	15	4	1	13	7	17	13	2
03/2015	24	32	10	23	16	33	26	27	33	31
04/2015	4	6	8	8	4	34	3	22	16	8
05/2015	17	26	23	22	9	41	21	41	17	21
06/2015	27	26	7	28	15	37	17	22	23	36
07/2015	24	36	11	33	14	54	40	52	23	39
08/2015	16	8	1	24	6	12	6	14	8	10
09/2015	39	30	18	27	17	47	28	39	21	23
10/2015	37	35	28	29	18	80	32	51	30	28
11/2015	24	34	8	19	14	40	26	30	19	21
12/2015	14	22	10	14	9	39	14	37	11	23

	Internet of Things	Disaster prevention	Bio-informatics	Processing framework	Visual data	Social big data	Smart Pow. grids	Machine learning	Energy efficiency	Traffic control
01/2016	15	14	10	10	3	16	7	16	10	7
02/2016	11	14	52	11	14	14	15	16	10	10
03/2016	11	15	13	21	8	16	19	11	12	9
04/2016	15	9	11	32	8	24	20	22	10	9
05/2016	42	30	18	37	14	39	31	48	27	27
06/2016	32	32	28	29	14	53	43	45	28	35
07/2016	19	18	11	15	11	33	21	32	20	18
08/2016	19	24	10	16	15	27	42	19	10	12
09/2016	23	14	19	27	15	28	19	35	19	13
10/2016	30	26	11	34	16	49	29	53	12	12
11/2016	18	12	12	13	12	25	17	24	12	14
12/2016	23	13	20	19	10	34	20	29	9	10

Table A2. Time series values of big data applications top components.

Author details

Iñaki Bildosola[1]*, Rosamaría Río-Bélver[2], Gaizka Garechana[1] and Enara Zarrabeitia[1]

*Address all correspondence to: inaki.bildosola@ehu.es

1 Industrial Management Department, University of the Basque Country (UPV/EHU), Faculty of Engineering in Bilbao, Bilbao, Spain

2 Industrial Management Department, University of the Basque Country (UPV/EHU), Engineering School of Vitoria-Gasteiz, Vitoria, Spain

References

[1] Zhang Y, Porter AL, Chiavetta D. Scientometrics for tech mining: An introduction. Scientometrics. 2017;**111**(3):1875-1878. DOI: 10.1007/s11192-017-2344-8

[2] Martino JP. Technological Forecasting for Decision Making. 3rd ed. McGraw-Hill, Inc.; 1993. p. 484

[3] Firat AK, Woon WL, Madnick S. Technological Forecasting–A Review. Composite Information Systems Laboratory (CISL). Cambridge, MA: Massachusetts Institute of Technology; 2008

[4] Beekhuyzen J, Hellens L, Siedle M. Cultural barriers in the adoption of emerging technologies. Las Vegas, Nevada USA: Proceedings of HCI International; 2005

[5] Coates V, Faroque M, Klavins R, Lapid K, Linstone HA, Pistorius C, Porter AL. On the future of technological forecasting. Technology Forecasting and Social Change. 2001;**67**(1): 1-17. DOI: 10.1016/S0040-1625(00)00122-0

[6] Engelsman EC, van Raan AF. A patent-based cartography of technology. Research Policy. 1994;**23**(1):1-26. DOI: 10.1016/0048-7333(94)90024-8

[7] Porter AL, Jin XY, Gilmour JE, Cunningham S. Technology opportunities analysis: Integrating technology monitoring, forecasting, and assessment with strategic planning. SRA journal. 1994;**26**(2):21-32

[8] Lee B, Jeong YI. Mapping Korea's national R&D domain of robot technology by using the co-word analysis. Scientometrics. 2008;**77**(1):3-19. DOI: 10.1007/s11192-007-1819-4

[9] Lee S, Lee S, Seol H, Park Y. Using patent information for designing new product and technology: Keyword based technology roadmapping. R&D Management. 2008;**38**(2): 169-188. DOI: 10.1111/j.1467-9310.2008.00509.x

[10] Zhang Y, Chen H, Zhang G, Zhu D, Lu J. Multiple Science Data-Oriented Technology Roadmapping Method. In: Management of Engineering and Technology (PICMET), Portland International Conference on. IEEE. 2015;2278-2287

[11] Joung J, Kim K. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change. 2017;**114**:281-292. DOI: 10.1016/j.techfore.2016.08.020

[12] Martino JP. A review of selected recent advances in technological forecasting. Technological Forecasting and Social Change. 2003;**70**(8):719-733. DOI: 10.1016/S0040-1625(02) 00375-X

[13] Jun S, Uhm D. Technology forecasting using frequency time series model: Bio-technology patent analysis. Journal of Modern Mathematics and Statistics. 2010;**4**(3):101-104. DOI: 10.3923/jmmstat.2010.101.104

[14] Chen H, Zhang G, Lu J. A time-series-based technology intelligence framework by trend prediction functionality. In: Systems, man, and cybernetics (SMC), 2013 IEEE International Conference on. IEEE. 2013:3477-3482

[15] Park SS, Jun S. New technology management using time series regression and clustering. International Journal of Software Engineering and Its Applications. 2012;**6**(2):155-160

[16] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence. 2009;**1**(1):60-76. DOI: 10.4304/jetwi.1.1.60-76

[17] Hair JF, Black WC, Babin BJ, Anderson RE, Tatham R. Multivariate Data Analysis. Prentice hall: Upper Saddle River, NJ; 1998

[18] Kongthon A. A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management. Atlanta, Georgia USA: Doctoral dissertation, Georgia Institute of Technology. 2008

[19] Maddala GS, Lahiri K. Introduction to Econometrics. Vol. 2. New York: Macmillan; 1992

[20] Storey VC, Song IY. Big data technologies and management: What conceptual modeling can do. Data & Knowledge Engineering. 2017;**108**:50-67. DOI: 10.1016/j.datak.2017.01.001

[21] Huang Y, Schuehle J, Porter AL. Youtie. A systematic method to create search strategies for emerging technologies based on the web of science: Illustrated for 'big data'. Scientometrics. 2015;**105**(3):2005-2022. DOI: 10.1007/s11192-015-1638-y

[22] Hu J, Zhang Y. Discovering the interdisciplinary nature of big data research through social network analysis and visualization. Scientometrics. 2017;**112**(1):91-109. DOI: 10.1007/s11192-017-2383-1

[23] Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Vol. 344. Hoboken, New Jersey USA: John Wiley & Sons; 2009. p. 342. DOI: 10.1002/9780470316801

Chapter 8

Altmetrics: State of the Art and a Look into the Future

Dirk Tunger, Marcel Clermont and Andreas Meier

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76874

Abstract

The development of alternative indicators (altmetrics) can be traced back to a discussion a few years ago where the central question was: does the focus on classical bibliometric indicators still adequately reflect the scientific and social significance of scientific work in the Internet age? In the course of this discussion, the term "altmetrics" was introduced as a collective term for all those indicators that contain previously unnoticed information from the Internet—especially concerning social media. Altmetrics shed light on the reception of scientific publications in news websites as well as in scientific blogs, policy papers, and other web-based content. This chapter deals with the current state of the art of altmetrics, focusing on the present discussion about the informative value of altmetrics. Furthermore, we investigate to what extent altmetrics can be used in scientific evaluations. We conclude our chapter with an outlook on the potential prospects for success of altmetrics in different fields of application.

Keywords: altmetrics, bibliometrics, informative value, scientific evaluation, social media

1. Introduction

Similarly to many areas of private life and business, increasing numbers of processes, results, and discussions in science are shifting to the digital sphere. For example, the scientific output is shared and discussed in established social media such as Twitter and Facebook. In addition, platforms created specifically for scientists, such as Academia.edu, ResearchGate, or Mendeley [1, 2], are also growing in numbers. The "Science 2.0" [3] era is progressing and this simultaneously increases the demand for indicators capable of measuring web-based impact. A pure consideration of the citation numbers from classical bibliometrics appears outdated since they reflect only a limited picture of the impact of scientific publications [4].

To date, web-based impact in social media has been measured mainly by the number of downloads or clicks, or by using indicators created by the operators themselves, such as ResearchGate's (RG) score [5]. These web-based metrics get the umbrella term "alternative metrics," or "altmetrics" [6]. Collecting and analyzing altmetrics is gaining relevance, and not only in science. Political decision makers, too, are attaching corresponding importance to the issue. Thus, the German Federal Ministry of Education and Research (BMBF), for example, has launched the first study evaluating the possibilities and limitations of using altmetrics for impact measurements [7]. Furthermore, BMBF has initiated a funding line for quantitative science research, in which the further investigation of altmetrics plays a central role.

The present chapter gives an overview of the current stance of scientometric research on altmetrics. We show example metrics and discuss what conclusions can be drawn from them. It will become apparent that altmetrics do not meet the expectation of measuring scientific impact because the data are too heterogeneous, their interpretation has not yet been sufficiently clarified, and an indicator system with meaningful and reliable benchmarks does not yet exist. Furthermore, we will investigate what strategies scientific institutions can pursue in using altmetrics and provide information on prospects for success.

2. Scientific discussion of altmetrics

2.1. Basic scientific context of altmetrics

The introduction of alternative indicators for the quantification of scientific output and the associated resonance on the Internet can be traced back to a discussion by Priem et al. in 2010 [6]. They questioned whether focusing on the classical bibliometric indicators adequately reflects the scientific and social significance of research in the era of the Internet. During the course of this discussion, the expression "altmetrics" was coined as a collective term for alternative metrics, which include web-based information on scientific publications. Therefore, altmetrics can be regarded as a complement to classical bibliometric indicators providing new information that was previously unavailable, predominantly from the social media sector. This new information makes it possible to examine the reception of scientific publications, for example, on news sites, in science blogs, policy papers, and other web-based sources.

The altmetrics community can now look back on almost 7 years of research. On the one hand, the "visibility and presence of altmetrics are quite impressive" [8] because they are used as marketing tools by many scientific publishers, more than 300 publications on the subject have appeared, and there are even conferences dedicated solely to altmetrics. On the other hand, there is no uniform definition, and therefore no consensus on what exactly is measured by altmetrics and what conclusions can be drawn from the results [8–10]. The only consensus regarding the term definition is that the indicators discussed are intended to measure the attention paid to scientific output where bibliometrics reaches its limits—that is, on the Internet [6]. There is, however, a lack of any further and more detailed differentiation of such metrics.

2.2. Tension between altmetrics and bibliometrics

Due to the fact that the base communities are the same, there is a certain tension between altmetrics and bibliometrics. Both (sub-)disciplines are intended to fulfill the same purpose, to generate a picture of scientific impact, but based on different influencing factors. Almost like a reflex, the two fields are often set in relation to each other, compared, or set up as an either/or selection.

In contrast, within the community itself, there is a general consensus that both disciplines complement each other instead of one excluding the other [11]. Altmetrics are not intended to replace the peer review process or bibliometrics; rather, they should be viewed as a second opinion [10] and a "new perspective on communication by and about science in social media" [7]. A report by the expert group on altmetrics on behalf of the European commission also argues for classical bibliometrics that they "offer complementary approaches to evaluation" together with alternative metrics [12]. The expert group furthermore sees potentials for including a wider audience beyond the closed science system and for collecting information considerably faster than with conventional metrics. Furthermore, the idea of this approach is not limited to conventional scientific publication formats but offers the perspective of making data sources such as software and data sets accessible (e.g., as part of research data management).

The big difference between bibliometrics and altmetrics is the aspect that scientific publications are the traditional and indispensable main output of science. Thus, bibliometrics measures something that is at the center of the scientific reward system. The communication of science to society—that is, what is measured by altmetrics—is not part of the scientific reward system as yet. Creating incentives and expanding this reward system at this point would likely lead to increased use of social media by science and thus also strengthen altmetrics.

2.3. Use of altmetrics in science evaluations

With regard to the practical application of altmetrics in research policy, science evaluations, and management, the scientific community is mostly skeptical. Bornmann and Haunschild [13] stress the problematic nature of the matter, namely that altmetrics should first confirm with the Leiden Manifesto for research metrics [14] before being applied on a greater scale. The central difficulties associated with altmetrics are presented, namely that there are currently no standardized indicators, that altmetric data are for the most part not accessible in a transparent and open manner, and that numbers can be manipulated through "gaming." Gaming is a term for the targeted manipulation of data for the purposes of achieving better altmetric values. Such gaming activities are negative side effects of an orientation along user statistics in evaluation practice [9]. However, in spite of the difficulty in consistently unambiguously distinguishing gaming from marketing, altmetrics service providers are trying to minimize such effects. For example, altmetric.com manually removes obvious manipulations of altmetric scores or limits them by means of spammer lists [15].

Gaming is also a problem beyond the sources assessed by altmetric service providers. In a study by Meier and Tunger, it became apparent that it is possible to considerably influence the metrics specially developed by the ResearchGate platform, the RG score [16]. The RG score is intended to measure the "scientific reputation" of ResearchGate users. It is influenced by the impact of a user's own scientific publications but also by their social activities on the platform (see https://www.researchgate.net/RGScore/FAQ). Meier and Tunger found that it is possible within a relatively short time to achieve an RG score that is higher that the RG scores of half of all RG users solely by gaming without any scientific publications.

In another study for the European commission, Kim Holmberg found that altmetrics are not yet practically applied in the EU for the purposes of scientific evaluation. In his view, such practice on a wide scale would be premature as long as what altmetrics actually measure remains unclear [17].

3. Problems associated with collecting and interpreting altmetrics

A semantic analysis of contributions in social media is lacking for the most part, which is a major issue making the evaluation of altmetrics counts so difficult. References are mostly counted based on identifiers such as the DOI; however, which references should be evaluated as positive and which as negative cannot be handled, which means that a "performance paradox" develops [18]. This paradox also exists in a similar form in classical bibliometrics and must be considered as an inherent problem of quantitative metrics in use [19].

Furthermore, the coverage of scientific publications is relatively low and the distribution varies heavily both across disciplines and across platforms. Haustein et al. found that 21.5% of all scientific publications in Web of Science in 2012 were mentioned in at least one Tweet, while the proportion of these publications in other social media was mentioned less than 5% [20]. In percentage comparison, 67% of the publications were cited in Web of Science at least once. A feasibility study conducted by BMBF shows strong variation concerning coverage on altmetric.com between the scientific disciplines: publications from the field of medicine are represented considerably more often than, for example, publications from the engineering sciences [7]. Differences in coverage appear to benefit the humanities sciences in particular. While these are scarcely considered in established databases such as Web of Science, their coverage is considerably better in the field of altmetrics, according to a study conducted by Hammarfelt: over 61% of the investigated publications in this field have at least one reader on Mendeley and more than 20% have already been discussed on Twitter [21].

In general, the data basis underlying altmetrics is often problematic: the reproduction of data is almost impossible because data providers change, modify their data stock, or disappear completely [4]. For example, platforms such as Weibo or LinkedIn, which are included in the sources covered by altmetric.com, are now no longer analyzed since these data providers no longer grant access. Quality control, such as a validity check of accounts or the clean-up of duplicates, rarely occurs for social media platforms and complicates the aggregating and filtering of data for altmetrics service providers [22].

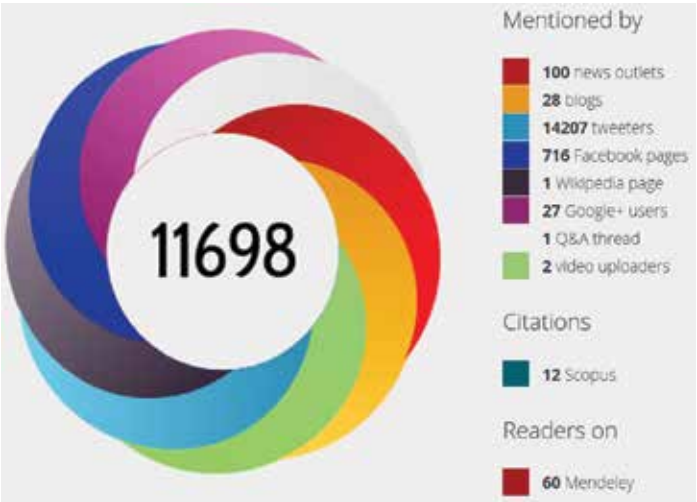


Figure 1. Example of the representation of the Altmetric Donut and its composition.

Furthermore, Fraumann et al. ascertained that duplicates can be found in several types of sources on altmetric.com, which makes the credibility of the attention score uncertain [23]. This attention score is currently used by many scientific publishers and institutions as a marketing tool in the form of the "Altmetric Donut" (see **Figure 1**). The Altmetric Donut is implemented on the websites of the journals *Nature* and *Science* among others, and in the repositories of the universities of Cambridge and Zürich. The composition of the attention score is based on an algorithm that adds up the attention—weighted differently—of scientific output in diverse sources. This trend is regarded skeptically in science, viewing the Altmetric Donut as a successful gimmick that is meaningless for science [9]. In general, simply adding up counts in a single metric is "impossible and undesirable" [12]. Thus, benchmarks such as the attention score do not represent the impact of scientific performance, but are suited solely to filter out those articles that have sparked interest in social media [24].

4. Requirements of altmetrics

To date, the European Commission ascribes high significance to altmetrics, particularly against the backdrop of open science. This is also reflected in the establishment of the associated expert group. The efforts have so far led to a compilation of twelve recommendations within the open science context. In the political context of the European Union's supranational level, the importance of guidelines for the conscientious application of metrics is emphasized. These guidelines are interlaced in the following with the demands from the Leiden Manifesto for research metrics.

The Leiden manifesto emphasizes the aspect of complementarity as a central principle and basis of any evaluation practice. According to it, for the existing qualitative practices, the aim

should be to complement each other in an advantageous manner. Peer review and expert assessment—this is the ambition—could be reinforced by the appropriate use of quantitative metrics, and further aspects beyond the traditional science system could be illuminated: "quantitative evaluation should support qualitative, expert assessment" [14].

Another aspect is the openness and transparency of all steps in the analysis process: "keep data collection and analytical processes open, transparent and simple" [14], that is, analyses should be verifiable and the indicators should not be unnecessarily complicated. At the same time, this does not mean that simple indicators (e.g., pure absolute numbers) with no significance should be used instead.

This recommendation is particularly important against the backdrop of the altmetric attention score since this composite indicator always combines data from many different sources. Their individual significance is unknown so that the score value can only contribute rudimentary information on the visibility of a publication in social media and therefore not be used for evaluation. At this point, attention should also be drawn to the inappropriate use of the journal impact factor, which occurs in a cumulative form particularly in medical science: its incorrect use as a citation indicator instead of as a simple journal indicator shows that it is immensely difficult to eliminate a metric once it has been established. Metrics in the scientific context must be reliable, reproducible, and significant.

5. Future potential of altmetrics in various fields of application

To what extent altmetrics will establish themselves in research policy depends fundamentally on empirical values from practical application in the sense of a learning experimental system. Therefore, potential fields of application are briefly outlined in the following paragraphs.

5.1. Science evaluation, performance assessment, and measurement of social impact

Due to the explorative development stage of altmetrics (as described above), they must be used carefully with regard to their application in the performance assessment of institutions and single scientists, for example within the scope of scientific evaluation. In particular, there is a lack of studies investigating how valid and reliable the evaluation of science based on altmetrics is. In the scientific discourse, a deeper understanding of the heterogeneity and the significance of the data must be achieved. In addition, useful indicators must be developed and benchmarking studies have to be conducted. According to current opinion, altmetrics will in the near future be more of a complementary component rather than an independent indicator for the assessment of scientific performance.

In addition, some research topics are more in the focus of society than others without necessarily displaying a larger social impact. In this context, attention should be drawn to the news values theory: it describes factors why some topics are reasonably sure to be reported and some are unlikely to become objects of journalistic reports in mass media [25]. Against this backdrop, altmetrics can be viewed as an incomplete indicator for social visibility. To what

extent this circumstance will change over time cannot currently be predicted and depends more on the social discourse on science and the opening of the science system than on further methodological developments.

5.2. Public relations, visibility, and advertising of activities

A part of communication on science and its visibility in the public sphere is represented by altmetrics. In any case, it should be noted that there is a rising trend in social media activity measured by the frequency of contributions and the number of people involved. Thus, it is becoming increasingly important to use social media platforms in order to proactively draw attention to research, that is, advertise it.

As an example in this context, institutional efforts such as those undertaken by universities or the European Commission, can be observed, which strategically position their own publications and activities. Against the backdrop of the explorative state of these efforts, altmetrics could serve as feedback, for example, to test various approaches aimed at new target groups in society. With regard to research policy, particularly activities with a strong social relevance and their visibility could represent an interesting field of application complementing current evaluation approaches for analyzing media feedback. Initial network analyses are already delivering promising results and their application to research policy issues could be examined. Using specific issues associated with communication propagation, attention could be focused, for example, on the identification of relevant multipliers—for example, science journalists and representatives from politics, industry, and interest groups—in the dissemination of information. Identifying such mechanisms and transmission channels in pilot studies would be promising research priorities in this respect in addition to medial feedback already addressed through established investigation designs.

Publishers already use the altmetric score mentioned in Section 3 as feedback on articles, albeit in a strongly aggregated and simplified form. Similar efforts are also apparent at universities and research institutions, which are testing the implementation of the Altmetric Donut both with and without the score, although the added value of these efforts has yet to be clarified. As part of a pilot measure, the OECD is currently investigating to what extent the altmetric explorer and the implementation of the altmetric score are suited to determine the social range of policy documents.

Science institutions can also use altmetrics within the scope of science marketing: it is conceivable that altmetrics could be used to focus attention on those publications by an institution that is widely discussed, shared, tweeted, or used in news pieces. This would permit the interface between science and society to be better addressed.

Whether there is any benefit from altmetrics in economics or politics beyond science has not yet been verified. From our viewpoint, there would be benefits if more sources of economic or policy-relevant sources were covered by the altmetrics databases. In this case, it would be possible to regard or measure the contribution of science in economy or policy. With bibliometric instruments, such as publication or citation analyses, it is not possible to measure this contribution since the economic or political world does not publish articles in scientific outlets. With

altmetrics one would be able to have a look at, for example, mentions of scientific publications in documents, which influence politics or discussions on the application of scientific research in economics or companies. Generally, it would be worthwhile to identify the impact of scientific contributions on individual groups more easily, if one could associate contributions on social media platforms to particular fields of application.

5.3. Reporting reputation

For scientists, the visibility of their publications is essential. The reputation resulting from the use by others of their scientific output in the form of ideas, statements, calculations, and findings is an essential part of the science system. Only the use of the generated output creates sustainable value for an individual scientist, be it in other scientific publications or in web-based communication, social media, or news pieces. Bibliometrics and altmetrics help scientists document the visibility of their work. Thus, the majority of the almost 700 scientists who participated in a survey on the RG platform stated that it is important to them to have a high RG score.

Altmetrics permit scientists to record, regulate, and document their own visibility to a greater extent than was previously possible. Particularly for early-career scientists, there is thus a great opportunity to increase attention and reputation independently from the traditional publication system. In the longer term, altmetrics could assume the function of documenting the mediation of science to society and of making it more transparent.

5.4. Support from libraries

Academic libraries are usually where contacts can be found within a scientific institution for issues related to publication data and bibliometric processes/indicators. Librarians' clean data, compile publication profiles, and collect data within the scope of evaluations. They are thus specialists for handling data, particularly data related to publications, user statistics, and stock management.

This is where altmetrics represent a connecting element as they illuminate the use of publications in social media. Thus it is plausible for libraries to be directly involved whenever the issue of altmetrics is addressed at an institution. This makes sense because librarians are in contact with many areas of a scientific institution and offer advice on using information products. Roemer and Borchardt [26] identified this central role of libraries and summarize: "[…] librarians serve as natural leaders when it comes to altmetrics […]" [26]. They argue that this is due to the resources and data knowledge of libraries as well as their central position as contact partners for various target groups [27, 28].

6. Conclusion

In conclusion, altmetrics are currently still at an explorative stage and have far to go before they can make a regular contribution to quantitative science indicators of bibliometrics [29].

We show that there are still problems with the indicators and associated benchmarks. This is why the use of altmetrics in the context of science evaluations is not yet conceivable. Simultaneously, however, this insight could function as an incentive to enhance application maturity and to create the political boundary conditions for advancing further developments. Thanks to initial applications of altmetrics in the academic context, important experience is being gained. The scientific debate over the past few years has thus led to altmetrics achieving the validity and application maturity required for initial applications. However, they must be further developed for applications that are more thorough; particular indicators have to go beyond the level of individual publications and should also aggregate data on various levels. Additionally, the problems of altmetric indicators have to be addressed especially regarding coverage, representativeness, gaming, and validity.

Interviews of the bibliometrics team at Forschungszentrum Jülich with experts in the field of bibliometrics and altmetrics confirm the above-mentioned findings [7]. These experts gave statements about the meaningfulness and application maturity of altmetrics. They stated that the significance of altmetrics indicators is located at a low to medium range only. The initial euphoria in the field, with the focus on the far-reaching potentials up to the measurement of the social impact and the performance evaluation of science, seems to have subsided.

There was a consensus between the experts that altmetrics is not an alternative to bibliometrics, but a new perspective on communication from and about science in social media: Perception and "popularity" are in the foreground. However, scientific quality or excellence is marginally represented by altmetrics, since it correlates only partially positively with perception. In principle, this contradicts bibliometrics, which is based on an inherent and peer review-based approach for the evaluation of science.

In contrast to the meaningfulness, the experts' assessments differ more strongly with regard to the maturity for application of altmetrics. This is sometimes due to the fact that expectations diverge: should these metrics be a purely quantitative indicator or do they provide the starting point for qualitative analyses? Furthermore, the areas of application are very broad and also include marketing activities that have so far been of secondary importance for research policy. Against this background, there is still unanimity that altmetrics can currently not be interpreted as a standalone and quantitative indicator. In particular, it was unanimously emphasized that altmetrics does not conform to a scientific database that is a prerequisite for the assessment of scientific work.

The appreciation of what role policymakers should play and how altmetrics can be used for research policy are divergent. However, in most of the interviews, the experts think that politicians should play an active role in shaping the implementation of altmetrics. Politicians could create a superordinate and binding framework for the application of altmetrics, for instance, by anchoring demands and formulating research questions.

In the long term, the increasing involvement of science in social media platforms will have a positive effect on the application of altmetrics. In addition, data providers are designing sources systematically and increasingly semantically. Current developments appear promising and point toward an expansion of source selection for English-language

policy documents and news articles [15]. This would mean that in addition to the relevant news target groups, two complementary transmission channels of science into politics and industry can be covered.

Author details

Dirk Tunger[1]*, Marcel Clermont[2] and Andreas Meier[1]

*Address all correspondence to: d.tunger@fz-juelich.de

1 Forschungszentrum Jülich, Jülich, Germany

2 Institute of Management Control and Business Accounting, Technische Universität Braunschweig, Brunswick, Germany

References

[1] Ortega JL. Relationship between altmetric and bibliometric indicators across academic social sites. The case of CSIC's members. Journal of Informetrics. 2015;**9**:39-49. DOI: 10.1016/j.joi.2014.11.004

[2] Thelwall M, Kousha K. Academia.edu: Social network or academic network? Journal of the Association for Information Science and Technology. 2014;**65**:721-731. DOI: 10.1002/asi.23038

[3] Shneiderman B. Human-computer interaction redefines science [Internet]. 2008. Available from: www.sciencedaily.com/releases/2008/03/080306170924.htm [Accessed: October 23, 2017]

[4] Haustein S, Peters I, Sugimoto CR, Thelwall M, Larivière V. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. Journal of the Association for Information Science and Technology. 2014;**65**:656-669. DOI: 10.1002/asi.23101

[5] Kraker P, Lex E. A critical look at the ResearchGate score as a measure of scientific reputation. In: Quantifying and Analysing Scholarly Communication on the Web (ASCW '15); 2015

[6] Priem J, Taraborelli D, Groth P, Neylon C. Altmetrics: A Manifesto [Internet]. 2010. Available from: http://altmetrics.org/manifesto [Accessed: October 23, 2017]

[7] Tunger D, Meier A, Hartmann D. Machbarkeitsstudie altmetrics [Internet]. 2017. Available from: http://hdl.handle.net/2128/16419 [Accessed: March 08, 2018]

[8] Haustein S. Vier Tage für fünf Jahre Altmetrics: Bericht über die Konferenz 2AM und den Workshop altmetrics15. b.i.t. Online. Vol. 19; 2016. pp. 110-112

[9] Franzen M. Digitale Resonanz: Neue Bewertungskulturen fordern die Wissenschaft heraus. WZB Mitteilungen. 2017;**155**:30-33

[10] Butler JS, Kaye ID, Sebastian AS, Wagner SC, Morrissey PB, Schroeder GD, Kepler CK Vaccaro AR. The evolution of current research impact metrics: From bibliometrics to altmetrics? Clinical Spine Surgery. 2017;**30**:226-228. DOI: 10.1097/BSD.0000000000000531

[11] Wouters P, Thelwall M, Kousha K, Waltman L, de Rijcke S, Rushforth A, Franssen T. The metric tide: Literature review (Supplementary report I to the independent review of the role of metrics in research assessment and management) [Internet]. Available from: http://www.dcscience.net/2015_metrictideS1.pdf [Accessed March 08, 2018]

[12] Wilsdon JR, Bar-Ilan J, Frodeman R, Lex E, Peters I, Wouters P. Next-generation metrics: responsible metrics and evaluation for open science [Internet]. 2017. Available from: http://eprints.whiterose.ac.uk/113919 [Accessed March 08, 2018]

[13] Bornmann L, Haunschild R. To what extent does the Leiden Manifesto also apply to altmetrics? A discussion of the manifesto against the background of research into altmetrics. Online Information Review. 2016;**40**:529-543. DOI: 10.1108/OIR-09-2015-0314

[14] Hicks D, Wouters P, Waltman L, de Rijcke S, Rafols I. Bibliometrics: The Leiden manifesto for research metrics. Nature. 2015;**520**:429-431. DOI: 10.1038/520429a

[15] Altmetric.com. Personal interview on August 14-15, 2017

[16] Meier A, Tunger D. Investigating the transparency and influenceability of altmetrics using the example of the RG score and the ResearchGate platform. Working Paper

[17] Holmberg KJ. Altmetrics for Information Professionals: Past, Present and Future. Amsterdam: Chandos Publishing; 2015

[18] Meyer MW, Gupta V. The performance paradox. Research in Organizational Behavior. 1994;**16**:309-369

[19] Holbrook J, Barr K, Brown KW. We need negative metrics too. Nature. 2013;**497**:439. DOI: 10.1038/497439a

[20] Haustein S, Costas R, Larivière V. Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. PLoS One. 2015;**10**: e0120495. DOI: 10.1371/journal.pone.0120495

[21] Hammarfelt B. Using altmetrics for assessing research impact in the humanities. Scientometrics. 2014;**101**:1419-1430. DOI: 10.1007/s11192-014-1261-3

[22] Thelwall M. A brief history of Altmetrics. Research Trends. 2014;**37**:3-4

[23] Fraumann G, Zahedi Z, Costas R. What do we know about Altmetric.com sources? A study of the top 200 blogs and news sites mentioning scholarly output [Internet]. Available at: http://hdl.handle.net/1887/48266 [Accessed: March 08, 2018]

[24] Galtung J, Ruge MH. The structure of foreign news. Journal of Peace Research. 1965;**2**:64-90. DOI: 10.1177/002234336500200104

[25] Warren HR, Raison N, Dasgupta P. The rise of altmetrics. Journal of the American Medical Association. 2017;**317**:131-132. DOI: 10.1001/jama.2016.18346

[26] Roemer RC, Borchardt R. Altmetrics and the role of librarians. Library Technology Reports. 2015;**51**:31-38

[27] Gimpl K. Evaluation von ausgewählten Altmetrics-Diensten für den Einsatz an wissenschaftlichen Bibliotheken [Internet]. Available from: https://publiscologne.th-koeln.de/frontdoor/index/index/docId/1034 [Accessed: March 08, 2018]

[28] Forschungszentrum Jülich. Almetrics. Metrics for information on the dissemination of scientific publications [Internet]. Available at: http://www.fz-juelich.de/zb/EN/altmetrics [Accessed: March 28, 2018]

[29] Haustein S. Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. Scientometrics. 2016;**108**:413-423. DOI: 10.1007/s11192-016-1910-9

Section 3

Patent Analysis

Chapter 9

Patent Research in a Period of Industry Transformation: A Focus on Electromobility

Zhao Qu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.75579

Abstract

Patent, as a valuable collection of technical information, is gaining momentum as proxy measures of innovative activities and is ascribed a unique role in tracking the rise of emerging technologies. The last 30 years have seen a dramatic transformation of the world's manufacturing landscape, for instance, a greening development in the automotive sector. A typical example of this practice is the emergence of electromobility (e-mobility)—an integrated approach addressing issues from sustainable transportation to revolutionary driving behavior adopted to circumvent problems concerning both resources and pollution while meeting mobility demands. Since novel technologies covered by e-mobility are not yet entirely attainable in the market, the only metric particularly is patent data. However, a correspondingly bright light seems not to be shined on e-mobility patent research, even in the area of engineering. This paper employs bibliometric and sentence-by-sentence analysis coupled with visualization tools to illustrate how the patent examines e-mobility-oriented issues in a contextualized and multivalent way. The conclusion reached is that patent research on e-mobility still has more spaces to move up, not only in improving its efficiency in plotting evolution of technologies but with regard to interpreting patents across the historical background of the industrial revolution.

Keywords: patent research, electromobility, bibliometric analysis, methods and design practices, conclusive and citing parts, sentence-by-sentence analysis

1. Introduction

As reported by WIPO, after 7 years of straight increases, global patent filings reached new highs in 2016—3.1million with annual growth of 8% [1]. It has correspondingly caught considerable interests ranging from social science research to economic analysis [2–4],

accumulating more than 16,000 international publications over the past 10 years. Conversely, patent study on e-mobility failed to go up by the same proportion even though it plays a tremendous role in promoting the development of technological innovation while being the focal point of a thoroughly international academic discourse. Referring to vehicles that rely on plug-in electricity for their primary energy [5], e-mobility is currently supported as a favorable approach to transform road transport by reducing carbon emissions and discussing drivers of change in the automotive industry [6, 7]. For accelerating technological progress, e-mobility has been extensively explored from commercial, political, and social network perspectives [8–10], indicating that this field represents a significant technical challenge and requires complex social changes. Even some deficiencies inherent in patent research are not to be neglected, for instance, not every technical invention is patented [11, 12]. Patent data as a special type of literature still has the advantage of being more retrievable and well organized in research for supporting scientific and technological decision, creating preferential development domains and protecting enterprise rights. Amid the rising concern and limited publication counts, new questions arise: How are patents integrated into e-mobility studies? How do e-mobility studies in turn shape them? And how, if at all, might scholars intervene in these processes?

In retrospect, patent documents have been assessed in conjunction with data extracted from scientific publications and industry products to examine recent developments and research progress on cold startup of automotive proton exchange membrane fuel cells (PEMFC), complete oxidation of methane at low temperature over noble metal, powertrain architectures, adsorbed natural gas technologies, and robust battery pack for electric vehicles (EVs) [13–17]. To trace the commercial pathway for ultra-capacitor technology, patents, especially the assignment information, are analyzed combined with investment figures [18]. However, the reason for applying patents to those studies has not been pointedly outlined and reviewed. Recent articles focus on patent-based indicators as to counts, families, portfolios, and citations in evaluating the effectiveness of e-mobility technology forcing policies and identifying technological changes, particularly around EVs [19–22], while the existing literature lacks details in conclusion on specific approaches and findings. The methods and design practices of e-mobility patent studies deserving of greater attention are the ones that place references at the forefront of the discussion about technology-driven innovation.

The present study, with a data set of 48 journal papers, is developed to review the patent research in the field of e-mobility by integrating a bibliometric overview on keywords and citations for insights into relevant research topics and knowledge base, then to trace back to the texts for an in-depth understanding of patent-use in practice and its contexts for answering the question: Does a lower share of international publications correspond to a less useful or more difficult intersection of patent analysis, especially into a field like e-mobility involving both traditional and emerging technologies? This special issue is a bridging effort to bring together patent study and bibliometric analysis by putting a spotlight on research progress, limitations, and potential topics in a period of industry transformation.

http://dx.doi.org/10.5772/intechopen.75579

2. E-mobility and significance of patent research

The invention of automobiles has been perceived as the promotion of global economic development and improvement of living standards by enabling mobile freedom. However, the growing concern for energy, environment, traffic safety, industrial competitiveness, and technology improvements raises the question of whether this freedom of mobility would be sustainable in the new era [7, 23]. The current and renewed interest in e-mobility can be explained in accordance with drivers of change earlier in the automotive industry. This term is not entirely new, and its central idea is urban electric cars which can be traced back to 50 years earlier than the first petrol-powered internal combustion engine vehicle (ICEV) [24, 25]. For stimulating technological progress, EVs in principle should now have a bright future; however, a lighter, cleaner, and smarter automobile era with adoption of wireless connections is in the movement [8, 9]. There is still considerable concern that efforts to date on making conventional powertrains more fuel efficient and less-polluting are insufficient [26]. This study thus is more inclined to adopt an expanded scope of e-mobility technologies other than the single category of EV-based technologies. Academics in this field are traditionally identified as having strong connections to governments and industries, as it is associated with the shift to a broader network of actors and stakeholders, ranging from automotive giants to battery-charging services providers [10, 27]. Thus, research on e-mobility not only seeks to answer the question of technology updating but is designed to give a sharp focus on changes caused by automotive industry transformation.

A wealth of technological, geographical, and industry information provided by patent has generated it to be a frequently used measure for studying basic research and anticipating emerging trends in automobile innovation [19, 21]. Bibliographic data extracted from patent documents is largely publicly available and quantitatively measurable [28, 29], which offers clear benefits in comparison with other indicators, for example, the one built upon R&D, to identify and measure patterns concerning innovative activities in uncertain technological fields [21]. Despite the controversial debate on the use of patent statistics to evaluate technological progress, the advantages prevail and empirical studies, particularly in research-intensive areas like e-mobility, support the application in obtaining an adequate output with a minimized input [11, 30]. The current publication counts contrast starkly with the significance of delving into patent issues of e-mobility. Hence, to drive further adoption of patent analysis as for e-mobility, scrutinizing related articles for progress, limitations, and potential topics is causally necessary.

3. Data and methodology

Advanced bibliometric analysis is regarded as a powerful method to answer questions, such as “How can we keep track of the increasing number of scientific articles? Are there specific patterns hidden in this mass of published knowledge at a meta-level, and if so, how

can these patterns be interpreted?" which enable us to analyze structures and dynamics of fields [31, 32]. Forty-eight articles in English identified by merging the query of terms[1] in the scope of e-mobility (e.g., electric vehicles, hybrid electric vehicle, etc.) with the topic search of patent (TS = patent*) from the Web of Science™ Core Collection (WoS) database up to 2017 are discussed in this chapter aiming to investigate the current progress of patent research on e-mobility. Visualizations are addressed throughout the discussion by explaining how they are produced and how they can be interpreted. Extrinsic data to the text such as the publication year, keywords, and citations are synthetically measured in a co-occurrence analysis, a technique that captures the frequency of pairs of words, phrases, or references in and between articles [33]. The first step is to represent the association of research topics and to observe the progress along with the time, source, and flow of knowledge, eventually to understand the development of scientific fields. The common base and expansion of knowledge are structured through backward and forward references by performing a co-citation and bibliographic coupling analysis, respectively, and the former depends on the frequency when two documents are cited together whereas the latter occurs when two works reference a common third work [33, 34]. Then, intrinsic information regarding the reason for performing patent analysis of e-mobility issues, research limitations, and trends dug out from abstracts, methods, conclusive parts, and recent highly cited papers are collected and categorized on a sentence-by-sentence basis in order to advocate for greater attention to article content in addition to the bibliometric analysis.

4. Patent research on e-mobility

4.1. A bibliometric overview

4.1.1. General trend and main research topics

The earliest publication involved in the present data set could be traced back to 1993 [35] and the pressing environmental concern has renewed calls for a shift toward internationalization of e-mobility research since 2012. Even the number of patent researches is limited, the sources are relatively scattered (132 authors from 58 affiliations), and their collaborations could not be captured in an extensive network (only 3 authors are engaged in the co-authorship more than once). In **Figure 1** 323 authors' keywords are cleaned up manually considering the problem of variants (e.g., "electric vehicles" and "EVs") and then 48 terms meeting the threshold (occurrences ≥2) are processed into 6 main clusters and colored using average publication year by VOSviewer. Patents have been employed to reveal the hype cycle for emerging technologies between 2001 and 2013 and then are discussed by introducing novel techniques (e.g., Google search) after 2014. Terms of "innovation" and "EV" as key nodes have been continuously

[1]Each term in the present query is sorted and filtered based on expert advice received from National Science Foundation of China (No.71673036) and Consulting Project of Chinese Academy of Engineering (2016-XZ-03-05).

http://dx.doi.org/10.5772/intechopen.75579

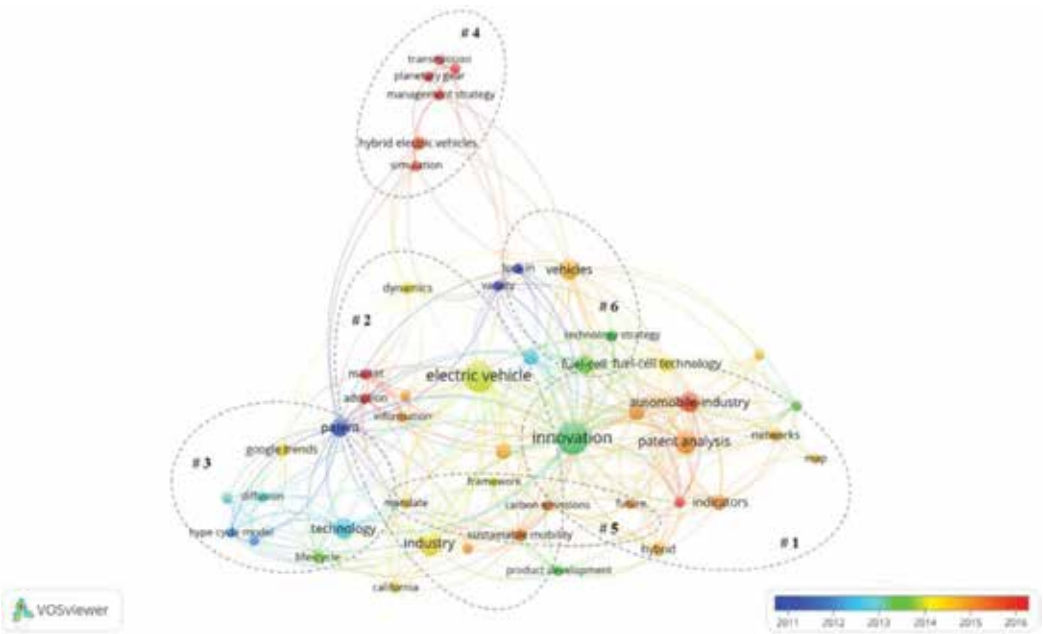


Figure 1. Co-occurrence analysis of keywords with average publication year.

explored ranging from the topics of patent-based indicators and approaches to technologies and the automotive industry as well as green products and market since 2013. Recently, the focus of patent research lies with the emergence of hybrid devices. E-mobility issues are inevitably tied to carbon emissions, efficient strategy, and sustainable development, which is proved in cluster 5 and 6.

4.1.2. Citation-based knowledge flows

Filtered by the minimum number of citations, 51 of 1788 references are collected and form four main clusters (**Figure 2**). The paper published by van den Hoed in 2005 [36] is represented as a key node among a group of emerging technology-based studies in red at the interface between discussions on emerging eco-innovation evaluation (green nodes on the top) and the cluster of papers adopting patent-based indicators in measuring technological change (yellow nodes in the middle). Note that citation is more frequent and probably more disciplined on the overall innovation performance research side, which also provides us with different kinds of evidence for the deficiency in e-mobility patent study. Among the technological forecasting-focused papers on the right side of **Figure 2** (blue nodes), the co-citation analysis highlights authors [37–39] who have engaged in discussions with the joint use of bibliometric and patent analysis. The first cluster indicated in red is led by research from van den Hoed and Bakker [36, 40], who share an interest in the development of fuel cell technology. Citations categorized into the second cluster have an earlier average publication year than that of the first cluster, including studies on e-mobility innovation coupled with policy, economic, and

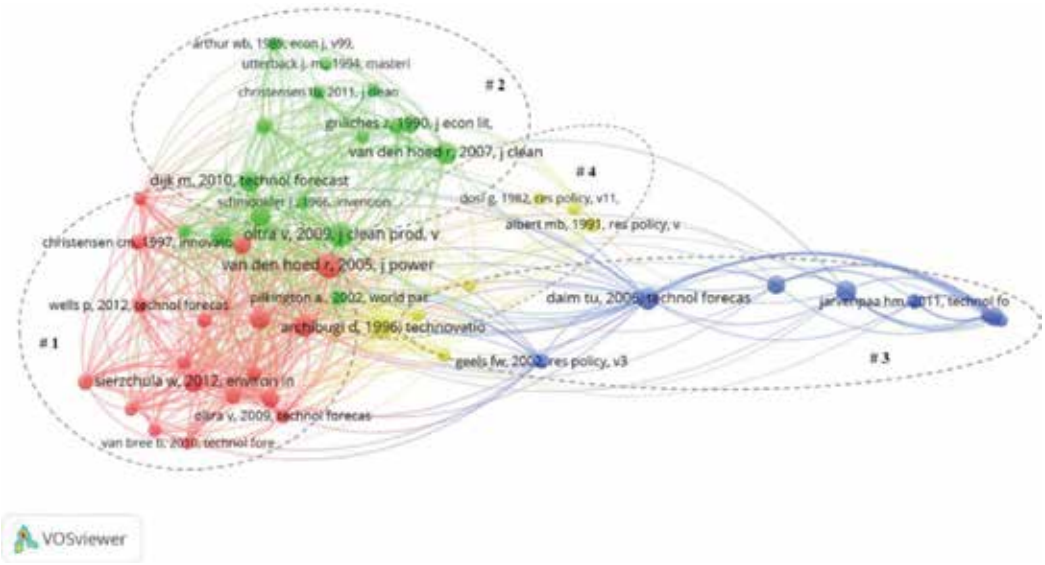


Figure 2. Co-citation analysis of references listed in 48 papers.

technology analysis [41–43]. Patent citation analysis is the central topic of discussion involved in the fourth cluster with the earliest average publication year of 2000 [44–46]. Namely, the core documents providing a common knowledge base for e-mobility patent study are relatively new and it is partly an effect of emerging technologies and the changing field.

By contrast, an analogous network structure could not be found through a bibliographic coupling analysis of 462 citing papers owing to the unclear layout of nodes and their links. It may indicate that the knowledge of these 48 patent studies is expanded to a much broader scientific field with more creative and diverse approaches to exploring e-mobility issues. Moreover, the share of self-citations is comparatively high and implies that there has existed specific groups of authors in this research area and their studies possess certain coherence.

4.2. The usage of patent

For answering the question of how patents are involved in e-mobility study and thus providing a reference point for further adoption of patent research, three main categories of the introduction of patents differed in research perspectives, and data formats are identified. **Table 1** outlines the direct use of patent-based indicators (e.g., counts, families, and citations) or patent documents (e.g., abstracts and literature) in view of reasons, including main routes as tracing the technological and industry trend, evaluating policies and innovation, as well as improving patenting activities and patent-based analysis [19, 22, 47–50]. Besides, the importance of patented technologies is highlighted while the rise in patenting activities and their commercialization indicate that a clean technology revolution staged by e-mobility is approaching [51, 52].

Reasons	Patent as the proxy for innovative output is the most common way in the automotive industry to protect intellectual property.	Patent documents include potential information on developed technologies
Measures	Patent—applications, counts, grants, families, origin countries and priority years, publications, citation networks, assignees, organizations, portfolios, keywords, International Patent Classifications (IPC)	Patent—pending applications, abstracts, literatures
Aims	To assess industry structure; examine the patterns of technological change; forecast diffusion or adoption patterns of new technologies; understand technology maturity; evaluate the effectiveness of technology-forcing policies; measure the incentive and opportunity to innovate; operationalize the R&D and commercialization aspects of innovation strategies; study the relationship between competitive forces and technological development; propose a predictive model of the patent registration time; find differences between patents and research publications for technology road mapping; filter the irrelevant patent citations; and verify components of the hype cycle	To describe technology in detail; explore technology clusters; give an indication as to main technical challenges for the relevant technology; assess technological evolvement or accelerate literature-based science discovery

Table 1. The usage of patent in an e-mobility study extracted from abstracts and methods.

4.3. Limitations and potential topics

4.3.1. Recent research limitations

Some of the deficiencies inherent in patent research are synthesized and divided into groups of limitations regarding patents or data sources [21, 53, 54] and patent-based indicators or approaches [47, 55, 56], respectively, thereby pinpointing areas of improvement in the further study on e-mobility. However, the following limitations should not be viewed in isolation, and the specificity of e-mobility field, especially the novelty and complexity of technologies [19, 21], needs to be considered in addition to patent-oriented issues (**Table 2**).

4.3.2. Potential topics

A series of up-to-date topics captured from citing articles based on recent highly cited ones in **Table 3** could be classified into the extension of the specific technology discussion, patent-based analysis, and research on innovation system or policies in the field of e-mobility [47, 57–62]. The classification of additional research perspectives is inevitably influenced by the usage of patent in highly cited papers (**Table 1**). More specifically, a review on patented technologies has developed the base for further experimental studies, and papers adopting patent-based indicators could arouse growing interests in examining the pattern of technological change [21, 63]. Patents combined with other format of data, such as scientific literatures, surveys, interviews, or press releases, may contribute to a more comprehensive understanding of relevant policy and innovation system research [48, 54].

Category	Limitations of patents or data sources in itself	Shortcomings of patent-based indicators or approaches
Limitations and drawbacks	-Not all technological knowledge is covered by patent data as not all inventions are patentable and not all patentable innovations are really being patented. -Web search is relied on secondary, and other sources are emerging, which may cause a shift in the category of search engine users. -Firms may exhibit differences in their tendencies to patent and their willingness to publish strategic decisions, thus affecting patent database. -It is difficult to cover every value chain step with relevant IPC codes.	-The analysis of the revealed technological advantage (RTA) is always subject to consideration of absolute numbers within the technologies. -Forming technology clusters through affinity propagation (AP) is susceptible and the interpretation of the technology clusters by using extracted core keywords is qualitative judgment. -Identifying important groups and the total groups that form a particular technology affects forecast. -Patent counts are not always representative for the success of the invention as its commercialization is not guaranteed.

Table 2. Recent research limitations stated in conclusive parts of publications (since 2015).

5. Discussion and conclusion

The present findings drawn from the bibliometric and sentence-by-sentence analysis of 48 journal papers indicate that patents, as indicators or references alike, still occupy an irreplaceable position in tracking the rise of emerging technologies. Since 2001, a sequence of structural data extracted from patents, like counts, grants, or classifications, has been employed to assess industry structure and examine the patterns of technological change. Assignees, organizations, and portfolios involved in patents are analyzed to measure the R&D and commercialization aspects of innovation strategies. New technologies' forecasting has been increasingly produced by keywords and patent citations accompanied with the emergence of advanced data search and mining techniques. Details of e-mobility technologies, ranging from batteries to smart grids, are scrutinized as references based on patent documents. Contents are continuously being specified and updated in line with the overall trend in the development of e-mobility, accounting for elements behind pure statistics. Even the patent study on e-mobility has not already accumulated a remarkable number of publications; the potential topics revealed by extended use of recent highly citied papers are researchable, including the analysis of automotive supply industry, technology diffusion, and landscape as well as the evaluation of green innovation system and policies. A specific focus in latter the e-mobility patent study is to expound key technical problems regarding free-piston linear generator. Drivers, the current momentum, and policies are constantly analyzed to answer the question of how EV development is accelerated. However, limitations rooted in patent data concerning patentability, search engine, willingness to publish, and the IPC-based bias with the one-sidedness of certain patent-based indicators mentioned as earlier should be noticed in further adoption of patents in e-mobility analysis, especially for improving its efficiency in plotting evolution of technologies and interpreting patents in a specific context. The limitation of such a study is that characteristics of patent study could not be fully identified because they are only identified from publications in an emerging field and the relevant search terms have not been unified in the past research. Nevertheless, it could be a sign of renewal when issues highlighted by those articles are explored in depth.

Times cited/Tittle (purpose)	Cited idea	Further study
18/Recent commercial free-piston engine developments for automotive applications (reviewing commercial developments in free-piston engine systems by looking at recent publications and patent documents)	"…discussed the basic features of a free-piston engine generator and the dynamics of the engine …"	Experimental study of a free-piston linear generator
15/Identifying trends in battery technologies with regard to electric mobility: evidence from patenting activities along and across the battery value chain (applying patent families as technological indicators in order to analyze the research activities of each step of the designed battery value chain individually and in comparison with each other to identify and discuss trends regarding the technologies associated with electric vehicles)	"…lithium-ion batteries have dominated very quickly…" "…the booming consumer electronics industry rapidly changed the economic and social landscape…" "…patent based studies have noted strong interest in radical innovation paths…"	Patent study on the automotive supply industry, technologies' diffusion, technological distance, technology landscape, and innovation network
9/Business strategies of incumbents in the market for electric vehicles: Opportunities and incentives for sustainable innovation (analyzing how environmental regulation and the firm's incentive and opportunity to innovate affected EV sales)	"…finds the same contrasting technology strategies employed by electric passenger vehicle manufacturers and companies need both incentives and opportunities…"	Review the proposal of conceptual framework regarding green innovation system; evaluate environmental policy and innovation
5/On the relation between communication and innovation activities: A comparison of hybrid electric and fuel cell vehicles (analyzing the relation between research and innovation activities and communication activities in the automotive industry using patent statistics, press releases, and interviews)	"…vehicle electrification can ease environmental problems …" "…the changing perception of the hybrid technology led to a 'hybrid race' testified by the significant increase in patents…"	analysis on drivers, the current momentum, and policies for accelerating EV development

Table 3. Highly cited publications and cited ideas (times cited ≥5 since 2015).

Acknowledgements

The author would like to thank for financial support provided by the China Scholarship Council (No.201506060153), the National Science Foundation of China (No.71673036), and the Consulting Project of Chinese Academy of Engineering (2016-XZ-03-05).

Author details

Zhao Qu[1,2]*

*Address all correspondence to: zhaoqu@dzhw.eu

1 Department of Social Sciences, Humboldt University of Berlin, Berlin, Germany

2 German Centre for Higher Education Research and Science Studies (DZHW), Berlin, Germany

References

[1] WIPO. World Intellectual Property Indicators - 2017 [Internet]. 2017. Available from: http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2017.pdf [Accessed: 2018-01-15]

[2] Jaffe, Adam B, Gaétan de Rassenfosse. Patent citation data in social science research: Overview and best practices. Journal of the Association for Information Science and Technology. 2017;**68**(6):1360-1374. DOI: 10.1002/asi.23731

[3] Lai Jessica C. The changing function of patents: A reversion to privileges? Legal Studies. 2017. DOI: 10.1111/lest.12176

[4] Gambardella A, Harhoff D, Verspagen B. The economic value of patent portfolios. Journal of Economics and Management Strategy. 2017;**26**(4):735-756

[5] Fédération Internationale de l'Automobile. Towards E-Mobility: The Challenges Ahead. [Internet]. 2011. Available from: http://www.lowcvp.org.uk/assets/reports/emobility_full_text_fia.pdf. [Accessed: 2017-11-10]

[6] Figenbaum E, Kolbenstvedt M. Electromobility in Norway-experiences and opportunities with Electric Vehicles (No. 1281/2013); 2013

[7] Sanden B. Systems Perspectives on Electromobility 2013. Chalmers University of Technology; 2013

[8] Fernandes SJV. Electric Vehicles: Technology, Policy and Commercial Development. Routledge; 2013

[9] Aichele C, Doleski OD, editors. Smart Market: Vom Smart Grid Zum Intelligenten Energiemarkt. Springer-Verlag; 2014. DOI: 10.1007/978-3-658-02778-0

[10] Capgemini. Managing the Change to E-Mobility [Internet]. 2012. Available from: https://www.capgemini.com/wp-content/uploads/2017/07/Managing_the_Change_to_e-Mobility___Capgemini_Automotive_Study_2012.pdf [Accessed: 2017-12-15]

[11] Qu Z, Zhang S, Zhang C. Patent research in the field of library and information science: Less useful or difficult to explore?. Scientometrics 2017;**111**(1):205-217. DOI: 10.1007/s11192-017-2269-2

[12] Tijssen R, Buter R, Van Leeuwen T. Technological relevance of science: An assessment of citation linkages between patents and research papers. Scientometrics. 2000;**47**(2):389-412

[13] Amamou AA, Kelouwani S, Boulon L, Agbossou K. A comprehensive review of solutions and strategies for cold start of automotive proton exchange membrane fuel cells. IEEE Access. 2016;**4**:4989-5002. DOI: 10.1109/ACCESS.2016.2597058

[14] Gélin P, Primet M. Complete oxidation of methane at low temperature over noble metal based catalysts: A review. Applied Catalysis B: Environmental. 2002;**39**(1):1-37

[15] Wu G, Zhang X, Dong Z. Powertrain architectures of electrified vehicles: Review, classification and comparison. Journal of the Franklin Institute. 2015;**352**(2):425-448

[16] Nie Z, Lin Y, Jin X. Research on the theory and application of adsorbed natural gas used in new energy vehicles: A review. Frontiers of Mechanical Engineering. 2016;**11**(3):258-274. DOI: 10.1007/s11465-016-0381-2

[17] Arora S, Shen W, Kapoor A. Review of mechanical design and strategic placement technique of a robust battery pack for electric vehicles. Renewable and Sustainable Energy Reviews. 2016;**60**:1319-1331. DOI: 10.1016/j.rser.2016.03.013

[18] Schultz LI, Querques NP. Tracing the ultracapacitor commercialization pathway. Renewable and Sustainable Energy Reviews. 2014;**39**:1119-1126. DOI: 10.1016/j.rser.2014.07.145

[19] Sierzchula W, Nemet G. Using patents and prototypes for preliminary evaluation of technology-forcing policies: Lessons from California's Zero Emission Vehicle regulations. Technological Forecasting and Social Change. 2015;**100**:213-224. DOI: 10.1016/j.techfore.2015.07.003

[20] Lee SL, Chen PC, Chan WC, Hung SW. A three-stage decision-making model for selecting electric vehicle battery technology. Transportation Planning and Technology. 2015;**38**(7):761-776. DOI: 10.1080/03081060.2015.1059122

[21] Golembiewski B, vom Stein N, Sick N, Wiemhöfer HD. Identifying trends in battery technologies with regard to electric mobility: Evidence from patenting activities along and across the battery value chain. Journal of Cleaner Production 2015;**87**: 800-810. DOI: 10.1016/j.jclepro.2014.10.034

[22] Yuan F, Miyazaki K. Trajectory identification as proxies for discerning the dynamic nature of technological change—The case of electric vehicles industry. International Journal of Innovation and Technology Management. 2017;**14**(01):1740006

[23] Chan CC. The rise & fall of electric vehicles in 1828-1930: Lessons learned [scanning our past]. Proceedings of the IEEE. 2013;**101**(1):206-212

[24] Adam M. Accelerating E-Mobility in Germany: A Case for Regulation. Springer; 2016. DOI: 10.1007/978-3-319-44884-8

[25] Santini DJ. Electric vehicle waves of history: Lessons learned about market deployment of electric vehicles. In: Electric Vehicles-The Benefits and Barriers. InTech; 2011

[26] Liesenkotter B, Schewe G. E-Mobility: Zum Sailing-Ship-Effect in Der Automobilindustrie. Wiesbaden: Springer Gabler; 2014. DOI: 10.1007/978-3-658-06310-8

[27] Kaltenbrunner W. Situated Knowledge Production, International Impact. Minerva: Changing Publishing Practices in a German Engineering Department; 2017. DOI: 10.1007/s11024-017-9337-x

[28] Narin F. Patent bibliometrics. Scientometrics. 1994;**30**(1):147-155

[29] Lerner J, Seru A. The use and Misuse of Patent Data: Issues for Corporate Finance and Beyond (No. w24053). National Bureau of Economic Research; 2017. DOI: 10.2139/ssrn.3071750

[30] Oltra V, Saint Jean M. Sectoral systems of environmental innovation: An application to the French automotive industry. Technological Forecasting and Social Change. 2009;**76**:567-583. DOI: 10.1016/j.techfore.2008.03.025

[31] Van Raan AF. Advances in bibliometric analysis: Research performance assessment and science mapping. Bibliometrics. Use and Abuse in the Review of Research Performance. 2014:17-28

[32] Ebrahim NA. Analysis of Bibliometrics Information for Selecting the Best Field of Study; 2016

[33] Wyatt S, Milojević S, Park H, Leydesdorff L. Quantitative and Qualitative STS: The Intellectual and Practical Contributions of Scientometrics; 2015. DOI: 10.2139/ssrn.2588336

[34] Zhuge H. Discovery of knowledge flow in science. Communications of the ACM. 2006;**49**(5):101-107. DOI: 10.1145/1125944.1125948

[35] Bai L, Qu DY, Conway BE, Zhou YH, Chowdhury G, Adams WA. Rechargeability of a chemically modified MnO2/Zn battery system at practically favorable power levels. Journal of the Electrochemical Society. 1993;**140**(4):884-889. DOI: 10.1149/1.2056222

[36] van den Hoed R. Commitment to fuel cell technology?: How to interpret carmakers' efforts in this radical technology. Journal of Power Sources. 2005;**141**(2):265-271. DOI: 10.1016/j.jpowsour.2004.09.017

[37] Daim TU, Rueda G, Martin H, Gerdsri P. Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technological Forecasting and Social Change. 2006;**73**(8):981-1012. DOI: 10.1016/j.techfore.2006.04.004

[38] Geels FW. Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. Research Policy. 2002;**31**(8):1257-1274. DOI: 10.1016/s0048-7333(02)00062-8

[39] Watts RJ, Porter AL. Innovation forecasting. Technological Forecasting and Social Change. 1997;**56**(1):25-47. DOI: 10.1016/s0040-1625(97)00050-4

[40] Bakker S. The car industry and the blow-out of the hydrogen hype. Energy Policy. 2010;**38**(11):6540-6544. DOI: 10.1016/j.enpol.2010.07.019

[41] Oltra V, SaintJean M. Variety of technological trajectories in low emission vehicles (LEVs): A patent data analysis. Journal of Cleaner Production. 2009;**17**(2):201-213. DOI: 10.1016/j.jclepro.2008.04.023

[42] Frenken K, Hekkert M, Godfroij P. R&D portfolios in environmentally friendly automotive propulsion: Variety, competition and policy implications. Technological Forecasting and Social Change. 2004;**71**(5):485-485. DOI: 10.1016/s0040-1625(03)00010-6

[43] Sierzchula W, Bakker S, Maat K, Van Wee B. Technological diversity of emerging eco-innovations: A case study of the automobile industry. Journal of Cleaner Production. 2012;**37**:211-220. DOI: 10.1016/j.jclepro.2012.07.011

[44] Archibugi D, Planta M. Measuring technological change through patents and innovation surveys. Technovation. 1996;**16**(9):451519-451468. DOI: 10.1016/0166-4972(96)000314

[45] Von Wartburg I, Teichert T, Rost K. Inventive progress measured by multi-stage patent citation analysis. Research Policy. 2005;**34**:1591-1607. DOI: 10.1016/j.respol.2005.08.001

[46] Pilkington A, Dyerson R. Innovation in disruptive regulatory environments: A patent study of electric vehicle technology development. European Journal of Innovation Management. 2006;**9**(1):79-91. DOI: 10.1108/14601060610640032

[47] Borgstedt P, Neyer B, Schewe G. Paving the road to electric vehicles–A patent analysis of the automotive supply industry. Journal of Cleaner Production. 2017;**167**:75-87. DOI: 10.1016/j.jclepro.2017.08.161

[48] Budde B, Alkemade F, Hekkert M. On the relation between communication and innovation activities: A comparison of hybrid electric and fuel cell vehicles. Environmental Innovation and Societal Transitions. 2015;**14**:45-59. DOI: 10.1016/j.eist.2013.11.003

[49] Jun S, Uhm D. A predictive model for patent registration time using survival analysis. Applied Mathematics & Information Sciences. 2013;**7**(5):1819-1823. DOI: 10.12785/amis/070520

[50] Yeh HY, Sung YS, Yang HW, et al. The bibliographic coupling approach to filter the cited and uncited patent citations: A case of electric vehicle technology. Scientometrics. 2013;**94**(1):75-93. DOI: 10.1007/s11192-012-0820-8

[51] Linnenluecke MK, Smith T, McKnight B. Environmental finance: A research agenda for interdisciplinary finance research. Economic Modelling. 2016;**59**:124-130. DOI: 10.1016/j.econmod.2016.07.010

[52] Coates D, Ferreira E, Charkey A. Development of a long cycle life sealed nickel-zinc battery for high energy-density applications. IEEE Aerospace and Electronic Systems Magazine. 1997;**12**(6):35-38. DOI: 10.1109/62.587056

[53] Jun SP, Sung TE, Park HW. Forecasting by analogy using the web search traffic. Technological Forecasting and Social Change. 2017;**115**:37-51. DOI: 10.1016/j.techfore.2016.09.014

[54] Wesseling JH, Niesten E, Faber J, et al. Business strategies of incumbents in the market for electric vehicles: Opportunities and incentives for sustainable innovation. Business Strategy and the Environment. 2015;**24**(6):518-531. DOI: 10.1002/bse.1834

[55] Kim G, Lee J, Jang D, et al. Technology clusters exploration for patent portfolio through patent abstract analysis. Sustainability. 2016;**8**(12):1252. DOI: 10.3390/su8121252

[56] Nagula M. Forecasting of fuel cell technology in hybrid and electric vehicles using Gompertz growth curve. Journal of Statistics and Management Systems. 2016;**19**(1):73-88. DOI: 10.1080/09720510.2014.1001601

[57] Hou X, Zhang H, Yu F, Liu H, Yang F, Xu Y, et al. Free piston expander-linear generator used for organic Rankine cycle waste heat recovery system. Applied Energy. 2017;**208**:1297-1307. DOI: 10.1016/j.apenergy.2017.09.024

[58] Abdalla II, Zainal AE, Ramlan NA, Aziz ARA, Heikal MR. Cogging force investigation of a free piston permanent magnet linear generator. IOP Conference Series: Materials Science and Engineering. 2017;**257**(1):012055. DOI: 10.1088/1757-899X/257/1/012055

[59] Karvonen M, Kapoor R, Uusitalo A, Ojanen V. Technology competition in the internal combustion engine waste heat recovery: A patent landscape analysis. Journal of Cleaner Production. 2016;**112**:3735-3743. DOI: 10.1016/j.jclepro.2015.06.031

[60] Binz C, Truffer B. Global Innovation Systems—A conceptual framework for innovation dynamics in transnational contexts. Research Policy. 2017

[61] Özel FM, Davies H, Wells P. What works? An ANFIS-based policy evaluation framework for electric vehicle technology development. International Journal of Electric and Hybrid Vehicles. 2017;**9**(3):222-252

[62] Bakker S, Farla J. Electrification of the car–Will the momentum last?: Introduction to the special issue; 2015

[63] Hanipah MR, Mikalsen R, Roskilly AP. Recent commercial free-piston engine developments for automotive applications. Applied Thermal Engineering. 2015;**75**:493-503. DOI: 10.1016/j.applthermaleng.2014.09.039

Chapter 10

Exploring Characteristics of Patent-Paper Citations and Development of New Indicators

Yasuhiro Yamashita

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77130

Abstract

In this study, the characteristics of "papers cited in patents" are examined and impact indicators of them based on existing bibliometric indicators are developed. First, the nature of patent-paper citations is examined for Japanese scientific papers as the basic knowledge for developing indicators. Second, the patent-paper citation index (PPCI) indicator, which was proposed in the previous study, is revised. Third, a set of indicators, named High Feature Valued Patent-Paper Citation Index, which is based on three feature values of citing patents, is proposed. Evidence using our new indicators is presented and the tendency of patent-paper citations of Japanese three sectors such as university, public institute, and corporation is discussed. Finally, issues to be addressed are discussed.

Keywords: patent-paper citations, impact indicators of papers, bibliometrics, institutional sectors, normalized citation impact, patent-paper citation index, technological impact, high feature valued patent-paper citation index

1. Introduction

Today, scientific research is expected not only to create knowledge but also to contribute to the development of industrial technology and the solution of social problems. Citations of scientific papers from patents (hereafter patent-paper citations) are rare data representing knowledge flows between coded scientific knowledge (scientific papers) and coded technological knowledge. Although there have been controversies over what is meant by patent-paper citations, it is deemed as data representing knowledge flows and used in the public statistics at present (e.g., see [1–3]).

As an indicator representing the relationship between science and technology, the number of cited scientific documents per patent (it is known as "science linkage") has been widely used. It is relatively straightforward to introduce science linkage, since it does not require identification of each scientific paper cited in patents and match to a specific record in databases of academic papers, such as Web of Science (WoS) and Scopus. However, science linkage only provides information on vicinity of science from technology, not vicinity of technology from science.

Along with the research utilizing science linkage as an index as described above, the nature of patent-paper citations itself has been studied. Such studies needed identification of bibliography of papers which appeared in patent documents. For example, Branstetter and Ogura [4] used data of patent-paper citations provided by CHI Research and analyzed the relationship between probabilities of occurring patent-paper citations and some variables obtained from both patents and papers for California. Such research had been relatively scarce, since they required a large-scale data set with identified paper data. However, in recent years, Ahmadpoor and Jones analyzed a large citation network, which consisted of patent-patent, paper-paper, and patent-paper citations, based on a large data set of US patents and scientific papers indexed in the Web of Science database provided by Clarivate Analytics and comprehensive patent-paper citation data [5]. They dealt with both patent-patent and paper-paper citations symmetrically and handled patent-paper citations like it bordered between these two networks and then uncovered differences in various aspects of them. Fukuzawa and Ida [6] analyzed the features of patent-paper citations from the paper side for 100 top researchers who were awarded the twenty-first-century COE. They found some important characteristics of patent-paper citations, such as the time lag of the former was longer than the latter, and the more the papers were cited from other papers, the more they tended to be cited from patents.

While these findings are important for practical use of patent-paper citations, there are almost no existing studies on the development of impact indicators of papers cited in patents.

On the other hand, the demand for methods of analysis and empirical indicator data of "papers cited in patents" in practical context has been expanded recently. For example, the Fifth Science and Technology Basic Plan which is the current Japanese five-year national plan for the promotion of science and technology between FY 2016 and 2020 requires monitoring of the performance. "Scientific papers cited in patents" is one of the key performance indicators of the plan. However, an effective method for showing performance using patent-paper citations is still unclear; therefore, it is indispensable to develop valid indicators of patent-paper citations.

My motivation is to develop impact indicators for scientific papers to show technological impact at meso (institutional sector in a country, research funding, and so on) to macro levels (country), based on the statistical nature of patent-paper citations. In the field of bibliometrics, many indicators have been developed and verified by many researchers (see [7]) and practical uses such as Leiden Ranking and Scimago Journal & Country Rank. Therefore, by developing robust impact indicators based on patent-paper citations symmetrical to existing bibliometric impact indicators, it should be possible to overview both the scientific and technological impacts of researches at the same time.

Moreover, from the view of patents, there have been many indicators for measuring patent quality (major indicators were written in [8]). For evaluating scientific papers from the aspect

of contributions to technological development, citations of scientific papers from "high-impact" patents seem to be good indicators of scientific papers. As far as my survey, I could not find any empirical study of indicators from the view mentioned above.

According to the aforementioned problem consciousness, I develop the new impact indicators of papers in the aspect of patent-paper citations. To secure the validity of new indicators, we investigate the nature of patent-paper citations in the dataset prior to the development of the indicators.

This article consists of the following sections. In Section 2, I explain data and time scheme of the study. I analyze relationships between probabilities of occurrence of paper citations from the patents and feature values of the scientific papers, using logistic regression analysis in Section 3. Based on the result of the analysis in Section 3, I improve the patent-paper citation index which we developed recently [9] (Section 4) and develop a set of new indicators from the aspect of patents' feature values (Section 5). Then, issues to be tackled are discussed in Section 6.

2. Data and their process

I utilized data sources and decided time scope in the study in the following process.

2.1. Patent data

I used worlds' patent data contained in the 2016 spring edition of the Patstat database produced by European Patent Office (EPO). The database contains patent applications filed until January 2016 and publications published until February 2016.

To avoid overrating the same inventions, patent data were counted by the DOCDB patent family. Only patent families which contain published patents, neither utility models nor design patents, were included in the dataset for securing consistencies of their statistic natures. Patent families are counted by their application year. The application year of the patent family was defined as the earliest filing year of the applications that constituted the family. Patent families which no application belonged to any of technology field defined in [10] were excluded, since percentiles of patent-patent citations were calculated by technology field.

2.2. Data of scientific papers

The Science Citation Index Expanded collection of the WoS database was used for this study. The WoS database contained bibliographic records of scientific papers which were published between 1981 and 2015. Each scientific paper in the WoS was classified to 1 of 22 scientific disciplines of the Essential Science Indicators. As for journals classified in "Multidisciplinary" by Clarivate Analytics, each of their papers was classified into 1 of the other 21 disciplines using their information on both forward and backward citations. Papers which were not classified into any of the 21 disciplines by the process were classified into "Multidisciplinary." They were excluded from the study because most of them obtained no or only a few citations and tended to be overestimated in the calculation of percentiles in the "Multidisciplinary" discipline. Disciplinary classification used in the study is shown in **Table 1**. Hereafter, I designated the codes for disciplines in the figures in this article.

Code	Discipline	Code	Discipline
AGS	Agricultural sciences	MTS	Materials science
BBI	Biology and biochemistry	MIC	Microbiology
CHE	Chemistry	MOL	Molecular biology and genetics
CLM	Clinical medicine	NEB	Neuroscience and behavior
CPS	Computer science	PHT	Pharmacology and toxicology
ECB	Economics and business	PHY	Physics
ENE	Environment/ecology	PLA	Plant and animal science
ENG	Engineering	PSS	Psychiatry/psychology
GSC	Geosciences	SPA	Space science
IMU	Immunology	SSS	Social sciences, general
MAT	Mathematics		

Table 1. Disciplinary classification of the study.

2.3. Linking non-patent literatures in the Patstat to specific papers in the WoS

All non-patent literatures appeared in the TLS214_NPL_PUBLN table of the Patstat and were matched to each bibliographic record of the WoS, so that citation links between them were identified. As a result of this process, 11,753,856 patent-paper citation links from Patstat to the WoS were identified. Number of WoS papers cited in the Patstat were 2,669,386, excluding duplications.

2.4. Attribution of institutional sectors to authors' organizations

Institutional sectors of authors' organizations were needed to be attributed to analyze tendencies of patent-paper citations by institutional sector in the following sections. The Connection Table between "Web of Science Core Collection" (WoSCC) and "NISTEP Dictionary of Names of Universities and Public Organizations" publicly provided by National Institute of Science and Technology Policy, Japan, was used for the purpose. The table consists of IDs of scientific papers in the WoS (UT), organization names, and sector and some other information extracted from the NISTEP Dictionary of Names of Universities and Public Organizations. The table contains UTs of Japanese papers published between 1998 and 2015 of which document types were "Article" or "Review." Therefore, the scope of data used in the study was limited to these document types and publication years.

The sectoral classification of the research was derived by combining the categories of the NISTEP table as shown in **Table 2**.

2.5. Time scheme of the study

As a result of the preprocess mentioned above, a scheme of time periods for analysis was set as **Figure 1**. A 6-year citation window (7 years including publication year of the scientific papers)

Sector classification in the study	Sector classification in NISTEP table
University	National university, public university, private university, interuniversity research institute
Public Institute	National institute, government-affiliated public corporation/independent administrative institution, institute of local government
Corporation	Corporation

Table 2. Institutional sector classification in the study.

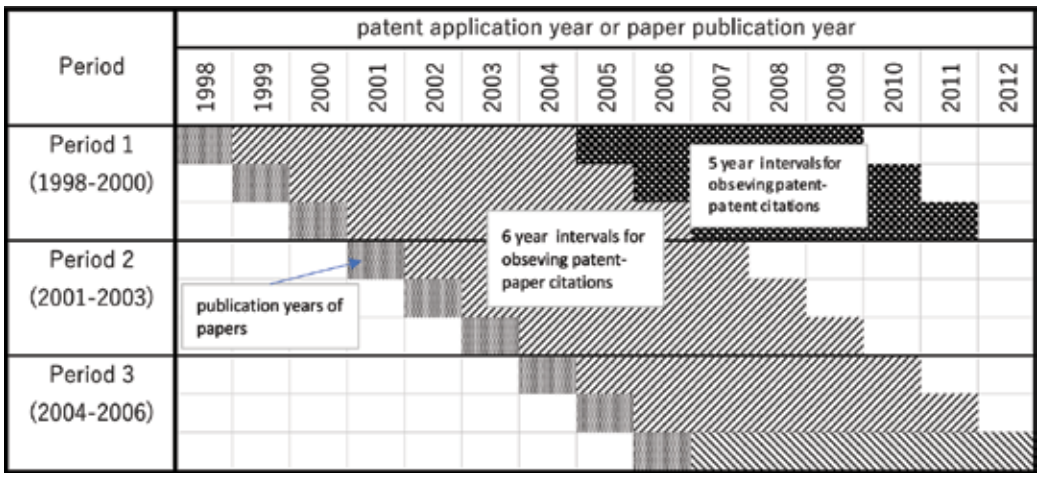


Figure 1. Time scheme of the study.

was secured for both patent-paper and paper-paper citations. The 6-year citation windows were defined in our previous study based on the criterion that at least a half of observable patent-paper citations could be grasped [9]. As for the earliest period (Period 1), 5-year citation windows were set according to [8] for observing citations from patents to patents citing target papers.

2.6. Basic statistics of the dataset

As a result of the abovementioned process, a dataset for the study, which consisted of 6,962,541 records of the worlds' scientific papers published between 1998 and 2006, was obtained. The number of Japanese papers by institutional sector counted fractionally by the number of addresses appearing in each paper in the dataset was shown in **Figure 2**. Japanese universities published 72.4% of Japanese papers; public institutes and corporations published 13.3 and 8.6%, respectively. When rate of papers cited in patents in papers of each sector was calculated, the above orders were reversed; the rate of papers cited in patent of corporation, public institutes, and universities was 21.6, 11.2, and 10.2%, respectively.

Number of the worlds' papers published between 1998 and 2006 by discipline was shown in **Figure 3**. Both clinical medicine and chemistry showed large numbers of papers, and that

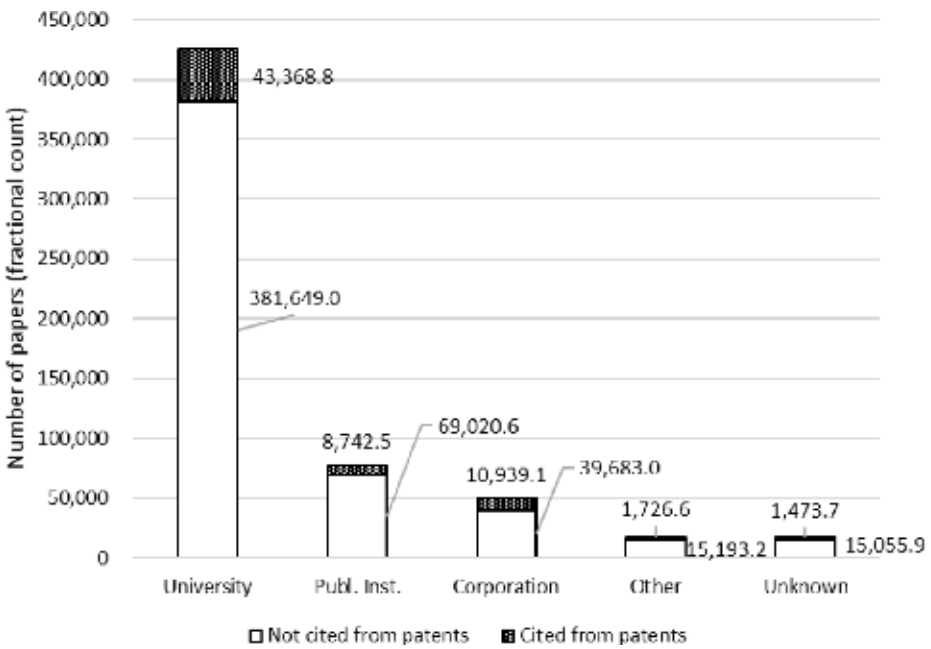


Figure 2. Number of Japanese papers by sector in 1998–2006.

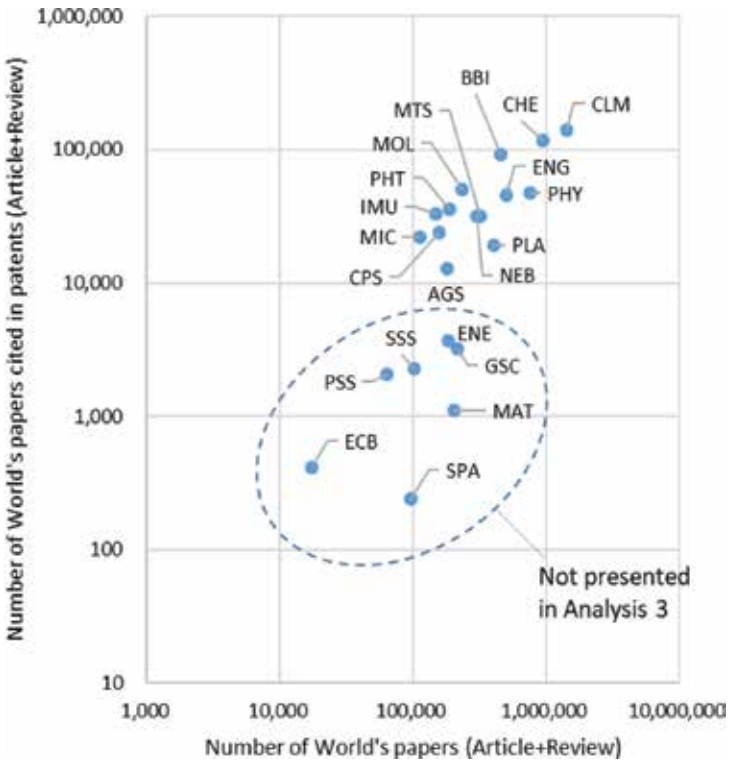


Figure 3. Number of publications and papers cited in patents between 1998 and 2006.

was cited in the patents. Biology and biochemistry showed relatively smaller numbers of papers but showed comparatively close number of that cited in patents to clinical medicine and chemistry. Therefore, it showed a relatively higher rate of papers cited from patents per their papers. Seven disciplines surrounded by the dotted circle in **Figure 3** showed both small number of papers and that was cited in the patents. These disciplines were excluded from presentation in analysis 3 (Section 5), in which analysis was executed and presented by discipline. However, these seven disciplines were included in the calculation as in other analyses, i.e., analysis 1 (Section 3) and analysis 2 (Section 4).

3. Relationships between feature of papers and patent-paper citations (analysis 1)

3.1. Research question

Patent-paper citations are different from paper-paper citations in their statistic nature, such as their small amount compared to that of the latter. Therefore, some indicators developed in bibliometrics cannot be applied to patent-paper citations. To develop valid indicators, many aspects of their tendencies, especially which kind of papers were preferred to cite in patent, should be grasped. Although some studies tackled this question partially [4–6, 11], their analyses were restricted to the US patents [4, 5, 11] or limited numbers of "top" researchers [6].

Moreover, it is still unknown how papers were cited from patents of which feature values were relatively high (hereafter, they are called as high-feature-valued patents). Branstetter [12] addressed the question whether patents citing papers tended to be high feature valued. However, his approach was done from the patent side, not the paper side. Patent-paper citations from high-quality patents seemed to be more valuable from the view of possibility of occurrence of innovation in many cases.

Here, I tried to grasp statistical tendencies of relationship between patent-paper citations from both all patents and those with high feature values. I intended to show the difference between them and to obtain basic knowledge of paper citations from high-feature-valued patents to develop valid indicators and show tendencies of (Japanese) scientific research from multi-aspects of patent-paper citations in the following sections.

3.2. Relationship between feature values of patents and their patentability

Although many "quality indicators" have been proposed, it might be questionable whether all of them exactly reflect patent quality. Since they each focused on different aspects of patents, they might represent different features of patents, not all of which represent "quality." To facilitate a precise understanding of the results of analysis of patent-paper citations from patents with high-"quality indicators" (hereafter they are called as "feature values" since they were not necessarily representative of quality), and the meaning of the new indicators proposed in Section 5, here I tried to show differences in meaning of the various major patent feature values.

In this subsection, I focused on the relationship between the three major feature values of patents: patent family size, forward citations (hereafter it is called as patent-patent forward

Independent variable	Coefficient	Std. err	Z value	Pr(>|z|)	Signif. codes
Intercept	0.474038	0.004213	112.51	<2e-16	***
Patent family size	0.257991	0.001038	248.62	<2e-16	***
Patent-patent forward citation (Top 1%)	0.029541	0.000276	107.02	<2e-16	***
Patent Generality Index	−0.188278	0.006765	−27.83	<2e-16	***

Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, " "1.

Table 3. Result of logistic regression analysis of patent feature values.

citations to distinguish it from other kinds of citations), and patent generality index. They are three of the four components of "composite index 4" presented in [8]. "Claims," which was the rest of the four, was not included in the study because it was not included in the Patstat comprehensively (only the US patents and European patents comprehensively included it exceptionally). As for "patent-patent forward citations," a dummy variable which distinguished whether patents obtained the top 1% of citations from other patents or not (it was presented as a "breakthrough" indicator in [8]) was used. The percentile of patent-patent citations was calculated by each of the 35 technology fields defined in [10].

Here, logistic regression analysis, of which independent variables were three patent feature values mentioned above, was executed. "Granted" flag in TLS201_APPLN table in the Patstat was selected as dependent variable, since it should represent an aspect of patent quality. Please note that this analysis was executed in the initial stage of the study before the specification of dataset was decided; therefore, all types of patents (such as utility models) were included.

The results are shown in **Table 3**. All coefficients of the three independent variables were significant at 0.1 percent level. Two of them (patent family size and patent-patent forward citations) were positive, and the rest was negative. As far as grant of patents was regarded as representative of patent quality, the former represents some aspects of patent quality. Patent family size could be thought of as quality assessed by applicants themselves (self-assessed quality), since "applicants might be willing to accept additional costs and delays of extending protection to other countries only if they deem it worthwhile" (p. 14) [8], while patent-patent forward citations could be deemed as quality assessed mainly by other applicants or examiners. On the other hand, the patent generality index seemed not to represent patent quality in the aspect of patentability.

3.3. Relationships between features of scientific papers and their citedness from all/high-feature-valued patents

In this subsection, I explored which features of papers affect their citedness from patents to grasp basic nature of patent-paper citations which might influence the nature of indicators presented in the following sections. Since we utilized information on patent-patent citations in which patents citing papers obtained, the analysis in this section was executed for Period 1 (PY1998–2000) in **Figure 1**.

Independent variable	(a) Cited/not		(b) Large patent family (> = 15)		(c) High patent-patent forward citation (top 1%)		(d) High patent generality index (> = 0.85)	
	Coefficient		Coefficient		Coefficient		Coefficient	
(Intercept)	−2.504476	***	−4.30065	***	−4.91949	***	−5.27141	***
Review	0.125596	*	0.29487	*	0.30569	*	0.20671	.
Int Coauthored					−0.09241		0.08866	.
IF	0.269865	***	0.15193	***	0.14490	***	0.13507	***
Top 10%	1.417856	***	1.42854	***	1.65927	***	1.59834	***
University	−0.281680	***	−0.42518	***	−0.35581	***		
Publ Inst	−0.038220	.	−0.36932	***	−0.11208	.	0.17765	***
Corporation	0.837952	***	0.83858	***	0.81681	***	0.62083	***
Other								
AGS	−0.268111	***	−0.39318	*	−0.91885	**		
BBI	0.895510	***	0.33564	***	0.57985	***	0.85431	***
CHE	0.044250	.	−0.35768	***	0.13846	*	0.76895	***
CPS	0.296150	***	−2.09236	***	0.80914	***		
ECB	−0.806569							
ENE	−1.403637	***	−2.40567	***	−1.92041	**	−1.21992	*
ENG	−0.144508	***	−3.77031	***	0.29438	***	0.27416	**
GSC	−3.268167	***	−15.37014		−2.16536	***	−3.29013	**
IMU	1.074738	***	1.19463	***	0.89635	***	0.49992	***
MAT	−4.296640	***	−15.18047		−13.65028		−13.66243	
MTS	−0.426886	***	−2.19212	***	0.25507	**	0.76189	***
MIC	0.829376	***	0.31394	*	−0.71977	*	0.29761	
MOL	1.063727	***			0.53478	***	0.94839	
NEB					−0.22540	.	−0.20025	
PHT	0.402472	***	0.71171	***	0.19559	.		
PHY	−0.559729	***	−3.77438	***				
PLA	−0.475982	***	−1.35393	***	−0.49401	***	−1.22287	***
PSS	−1.228774	***	−1.60205		−13.72792		−13.74086	
SPA	−4.640363	***	−15.22141		−13.74022		−13.82943	
SSS	−1.694540	***	−2.23959	*	−1.75911	.	−13.78140	

Signif. codes: "***" 0.001, "**" 0.01, "*" 0.05, "." 0.1, ""1.

Table 4. Result of logistic regression of rate of patent-paper citations.

I tried to include broad feature values of papers which might affect their citedness from patents as widely as possible to grasp characteristics of patent-paper citations comprehensively. Six feature values (document type, international co-authorship, impact factor (hereafter IF), paper-paper citations, institutional sectors and disciplines) shown in **Table 4** were selected from [13]. In **Table 4**, the variable "Review" and "Int-Coauthored" represents the feature value "document type" and "international co-authorship," respectively, and the variables "University" to "Other" and "AGS" to "SSS" represent "institutional sectors" and "disciplines," respectively.

I executed logistic regression analyses of which independent variables were six feature values of papers mentioned above and dependent variables were distinct from whether papers were cited from (all or high-feature-valued) patents (1) or not (0). To ignore the shape of distributions of patent-paper citations, I discarded information on the number of citations but used distinction of cited or not.

IFs were obtained from the Journal Citation Reports produced by Clarivate Analytics. Since IFs changed every year, years of IFs were defined as publication years of papers. This was because I intended to use them as the journals' quality indicators independent of the target papers. IFs in a year Y were calculated using papers published in years Y-1 and Y-2; therefore, they did not contain the target papers in the calculation. As it was well known, values of IFs differed largely by discipline; therefore, they were normalized by the following process: (1) IFs were attributed to each paper in the WoS (but IFs could not be given to some papers exceptionally); (2) mean values of IFs attributed to papers by ESI discipline were calculated for each year; (3) IF attributed to each paper was normalized by mean IF of its ESI discipline.

The threshold values of feature values of patents were decided according to the criteria: number of papers cited in high-feature-valued patents should be almost the same. As the number of papers cited from the top 1% patent-patent forward citation patents was predetermined, it was used as the reference value of number of papers cited from high-feature-valued patents. Threshold values were set to 15 for patent family size, 0.85 for patent generality index. Therefore, patents of which patent-patent forward citations were within top 1% or patent family sizes or patent generality indexes were equal to or more than the abovementioned thresholds were defined as high-feature-valued patents in this study.

Document types "Article" and discipline "Clinical Medicine (CLM)" were set to reference, since they were classified exclusively.

The results of the logistic analyses were shown in **Table 4**. Since patent-paper citations from high-feature-valued patents ((b), (c), (d)) were subsets of the whole patent-paper citations, they showed somewhat similar tendencies.

As for document type, reviews showed positive relationships to probabilities of being cited from both patent ((a)) and all three types of high-featured-valued patents ((b)-(d)). The result on patent ((a)) reinforced the result by Hicks et al. [11]. This result showed that indicators should be weighted by document type as far as possible.

International co-authorship showed no statistically significant relationship to any kinds of paper citedness. While Japan's co-authorships with any country were combined into the same

flag, it might show a statistically significant difference if difference of countries was taken into account. However, the number of international co-authored papers was limited, so we did not divide them into specific countries.

IF showed positive relationships with all kinds of patent-paper citations. This result reinforced analysis of Guan and He [14]. They showed nine of ten journals most frequently appeared as non-patent literatures in Chinese inventors' US patent were ranked within the top ten in their categories in the Journal Citation Report. Therefore, papers published in prestigious journals tended to be more cited than those published in lesser known journals.

The top 10% of paper-paper citations also showed positive relationships with all kinds of patent-paper citations, as many previous studies [5, 6, 11].

Institutional sectors showed some interesting tendencies; corporations showed relatively strong tendencies to be cited from all four kinds of patents ((a)-(d)). Although university and public institutes tended not to be cited from patents generally, they were not so from patents with high patent generality indexes. Latter tendencies might be explained that universities and public institutes produce generic knowledge, not focus on specific industrial applications, so patents citing them tended to also have a generic nature.

As for disciplines, some of the life sciences (biology and biochemistry, immunology, microbiology, molecular biology and genetics, pharmacology and toxicology) showed tendencies to be more cited (than clinical medicine, which was a reference discipline), while most physical sciences (engineering, materials science, physics) showed opposite tendencies. Similar results were reported in previous studies, such as [11]. However, it also showed some interesting tendencies when citations from high-feature-valued patents were focused on. For example, computer science tended to be more cited relatively, while they tended to be less cited from large patent families; engineering and materials science tended not to be cited from patents, while they tended to be cited from patents of top 1% patent-patent forward citations; microbiology showed an opposite tendency in that they tended to be cited from patents, while they tended not to be cited from patents of top 1% patent-patent forward citations. What caused such differences? To answer this question, further investigation from the patent side is needed.

4. Improvement of the patent-paper citation index (PPCI) (analysis 2)

4.1. Definition of improved PPCI

In the previous study, we proposed an impact indicator of patent-paper citations, named patent-paper citation index (PPCI) [9]. PPCI is based on rates of the papers cited from patents in the targets' publications. We proposed a method to overview targets' research activities from both scientific and technological impacts compared to the world average by using normalized citation impact (NCI) [13] in combination. Differences in both document types and disciplines were ignored in the previous study [9]. However, the analysis in Section 3.3 revealed their effects on papers' tendencies to be cited from patents. Therefore, I propose an improved version of PPCI in this section.

NCI, which was the basis of PPCI, is the ratio of the number of paper-paper citations which the target paper got to the expected value of that of the same cohort papers in the world. NCI is calculated for paper by paper, so when it is applied to an aggregate, such as institutions or countries, the average per their publications' NCI is applied. On the other hand, PPCI is based on the rate of papers cited in patents in targets' publications. Indeed, it is preferable to apply the same definition as NCI to secure symmetry; we applied the abovementioned definition to avoid influence of limited highly cited papers, since the rate of papers cited from patents was relatively smaller than that from papers.

Improved PPCI was defined as Eq. (1):

$$p_{ijd} = \frac{\left(n'_{ijd}/n_{ijd}\right)}{\left(N'_{id}/N_{id}\right)} \tag{1}$$

where.

n_{ijd}: number of target *j*'s papers with document type *d* published in discipline *i*;

n'_{ijd}: number of target *j*'s papers cited in patents with document type *d* published in discipline *i*;

N_{id}: number of total papers with document type *d* published in discipline *i*; and

N'_{id}: number of total papers cited in patents with document type *d* published in discipline *i*.

Target j's field weighted PPCI was calculated as follow:

$$P_j = \frac{\sum_i \sum_d p_{ijd} \times n_{ijd}}{\sum_i \sum_d n_{ijd}} = \frac{\sum_i \sum_d \left(N_{id} \times n'_{ijd}/N'_{id}\right)}{\sum_i \sum_d n_{ijd}} \tag{2}$$

To increase visibility, we normalized PPCI by Eq. (3):

$$\text{Normalized } P_j = \frac{\left(P_j - 1\right)}{\left(P_j + 1\right)} \tag{3}$$

Hereafter, improved Normalized PPCI (Eq. (3)) is merely called as PPCI.

While the whole counting method was used to count Japanese sectors' publications in the previous study [9], the fractional counting method by number of addresses which appeared in each paper was used. The whole counting method always attributed one count to each target appeared in a paper, so they are easy to understand intuitionally; however, it often causes overrating to multiauthored papers.

4.2. Chronological changes of NCI and PPCI of Japanese sectors

Next, I tried to apply PPCI to three Japanese sectors (university, public institute, corporation) to show how PPCI could describe the scientific and technological impact of aggregate of meso (sector) level. This was mainly aimed to figure out on which level of aggregates PPCI could be used. The chronological change of both NCI and PPCI was shown in **Figure 4**.

http://dx.doi.org/10.5772/intechopen.77130

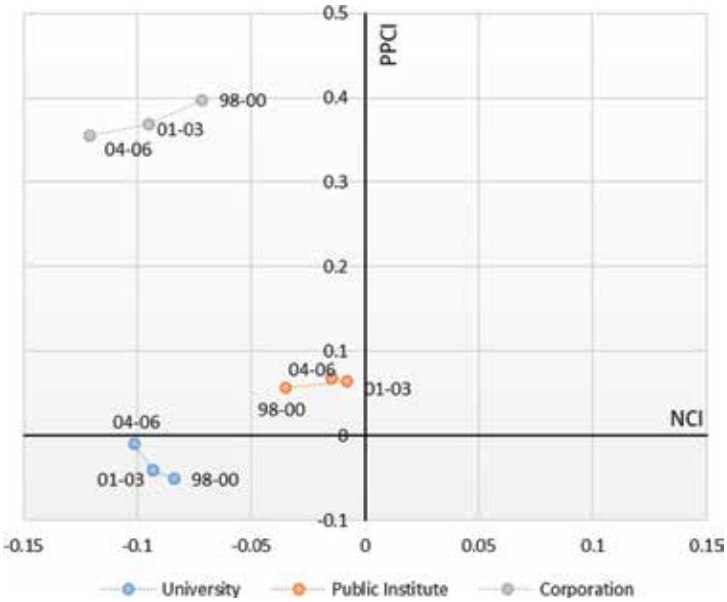


Figure 4. Chronological change of NCI and PPCI of three Japanese sectors.

All three sectors were located on the left half of the plane, which meant average scientific impacts of them were below world average during three periods. Two sectors, public institute and corporation, were located on the second quadrant; therefore, their average technological impacts were above the world average. In particular, corporation showed a remarkably high PPCI values and seemed to have been specializing in technological impact only period by period. University, which published most of the Japanese papers, was located on the third quadrant, which meant both scientific and technological impacts were below world average. However, their PPCI had been increasing period by period.

5. Development of high-feature-valued patent-paper citation index (analysis 3)

5.1. Definition

I showed that tendencies of paper citations from high-feature-valued patents differed from whole patents in some cases. It is suggested that indicators based on high-featured-valued patents might reveal hidden structure of the targets' research performance.

I tried to develop another indicator symmetrical to the PPCI to use them in combination. Here, we introduced the indicators based on paper citations from high-feature-valued patents, named high-feature-valued patent-paper citation index (HFPPCI). HFPPCI is a generic name of set of indicators, since there were many kinds of patent feature values. Of the many kinds of patent feature values, I will show the analysis of three patent feature values (patent

family size, patent-patent forward citations, and patent generality index) of Japanese sectors to examine the nature of HFPPCI as well as to show the tendencies of the Japanese sectors.

HFPPCI of target *j* in discipline *i* was defined as Eq. (4):

$$p_{ij}^{h} = \frac{\left(m_{ij}^{'}/n_{ij}\right)}{\left(M_{i}^{'}/N_{i}\right)} \tag{4}$$

where.

n_{ij}: number of target *j*'s papers published in discipline *i*;

$m_{ij}^{'}$: number of target *j*'s papers cited in high-feature-valued patents published in discipline *i*;

N_{i}: number of total papers published in discipline *i*; and

$M_{i}^{'}$: number of total papers cited in high-feature-valued patents published in discipline *i.*

To increase visibility, we normalize HFPPCI by Eq. (5):

$$\text{Normalized } P_{ij}^{h} = \frac{\left(p_{ij}^{h}-1\right)}{\left(p_{ij}^{h}+1\right)} \tag{5}$$

Here, the difference in document types was ignored, since the number of review papers cited from high-feature-valued patents was very few. Eq. (2) could be applied to aggregate p_{ij}^{h} into the whole target level; however, the selection of disciplines was inevitable because paper citations from high-feature-valued patents occurred rarely and $M_{i}^{'}$ might be zero in some cases.

5.2. Japanese sectors' PPCI and HFPPCI by discipline

In this subsection, I tried to analyze the Japanese three sectors' technological impacts by discipline in Period 1 (1998–2000). HFPPCIs of three patent feature values were called as large patent family paper citation index (LPFPCI) for large patent family, high forward citation patent-paper citation index (HFCPCI) for the patents of high patent-patent forward citations, and high generality patent-paper citation index (HGPCI) for patents with a high patent generality index. Definition of high-feature-valued patents was same as Section 3.3: equal or more than 15 for patent family size, top 1% for patent-patent forward citations, and equal or more than 0.85 for patent generality index. In the following subsections, document types were ignored in the calculation of both PPCI and HFPPCI. Both PPCI (X-axis) and HFPPCI (Y-axis) were plotted in bubble charts, and the number of papers cited from high-feature-valued patents was presented as size of the circles in **Figures 5–13**.

5.2.1. University

For LPFPCI, each discipline in **Figure 5** was positioned in line to some extent. This roughly means that large patent families of most of the disciplines in the sector appeared in proportion to papers cited in patents. In this case, there were not very much special information that could be obtained from the LPFPCIs, because PPCI contained almost the same information as LPFPCI. However, it was suggested that the LPFPCI functioned robustly, since there were only few deviating cases.

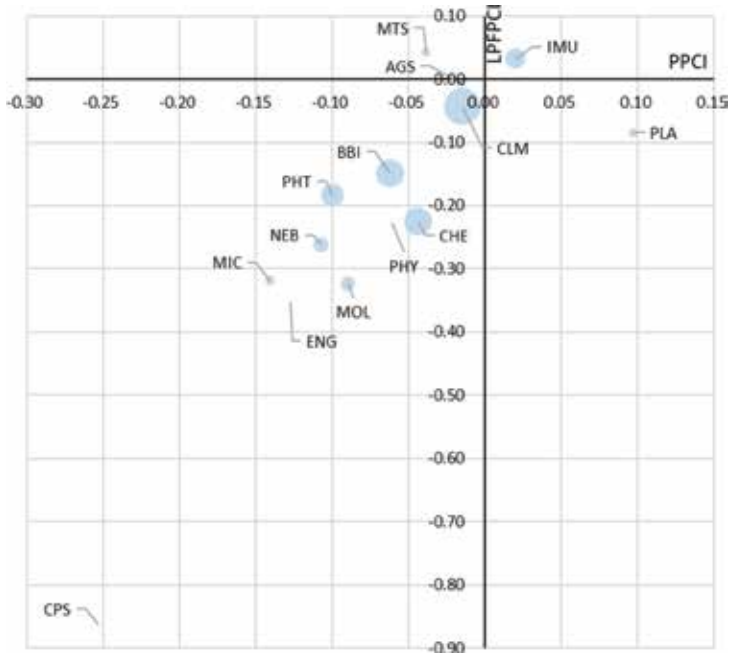


Figure 5. PPCI and LPFPCI of Japanese university sector by discipline (1998–2000).

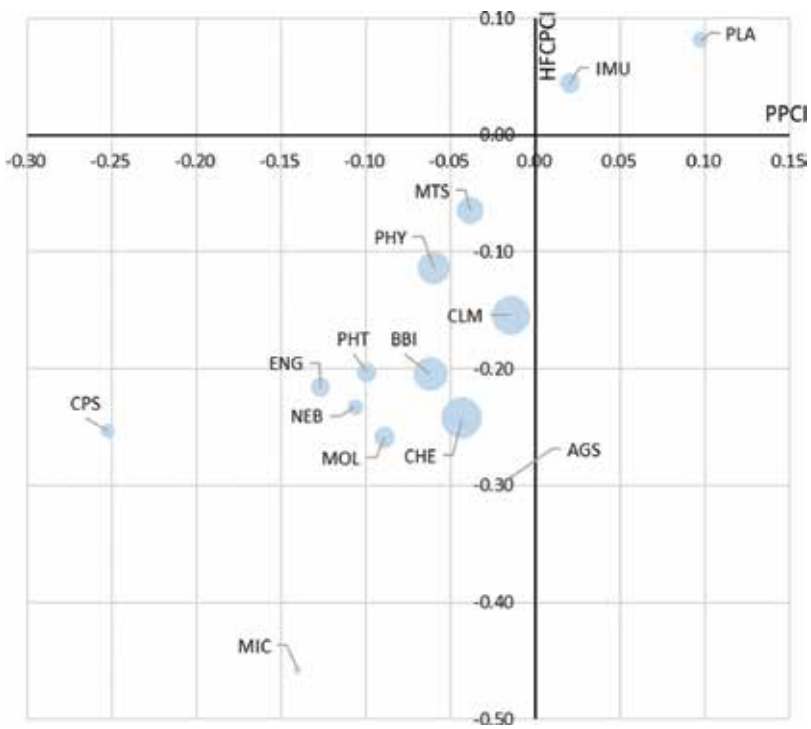


Figure 6. PPCI and HFCPCI of Japanese university sector by discipline (1998–2000).

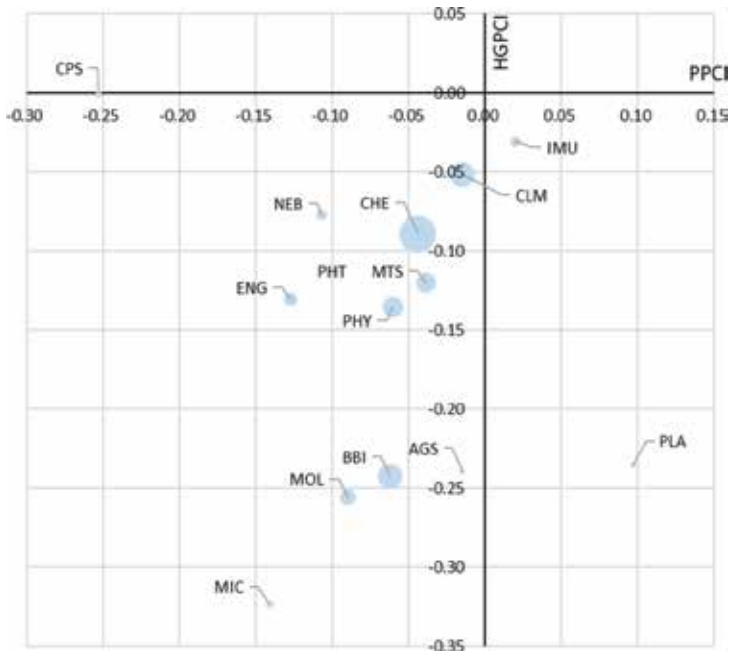


Figure 7. PPCI and HGPCI of Japanese university sector by discipline (1998–2000).

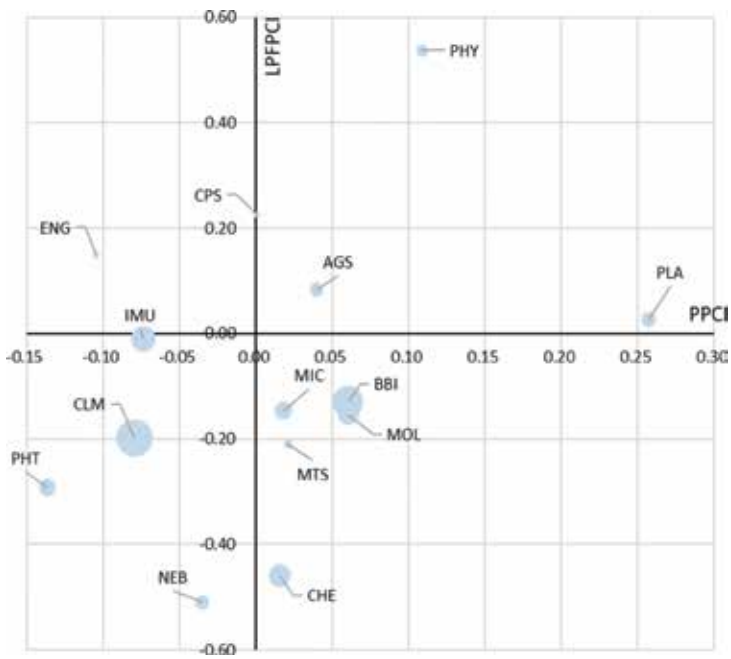


Figure 8. PPCI and LPFPCI of Japanese public sector by discipline (1998–2000).

http://dx.doi.org/10.5772/intechopen.77130

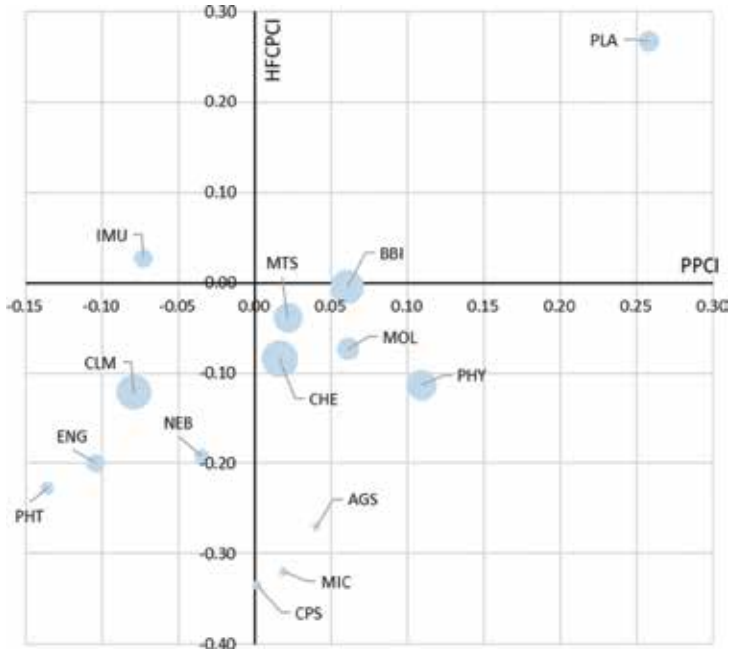


Figure 9. PPCI and HFCPCI of Japanese public sector by discipline (1998–2000).

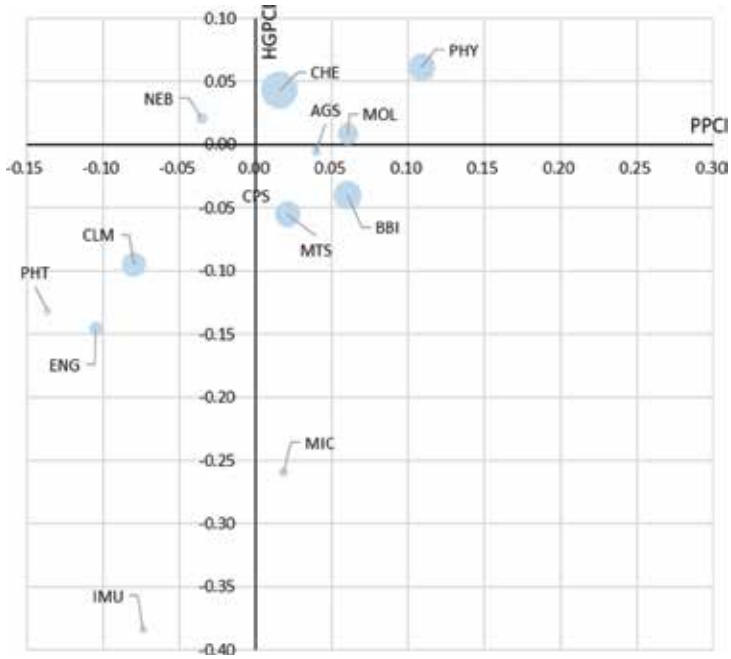


Figure 10. PPCI and HGPCI of Japanese public sector by discipline (1998–2000).

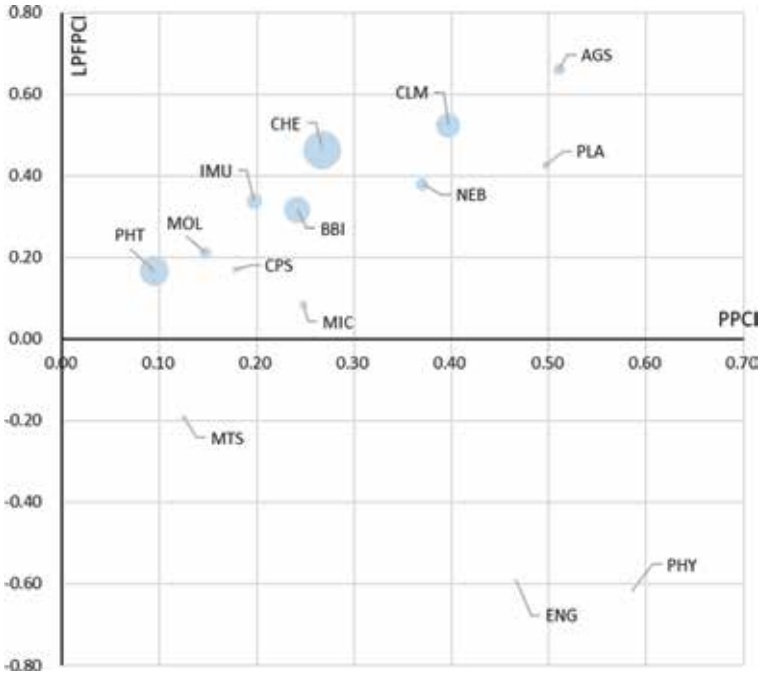


Figure 11. PPCI and LPFPCI of Japanese corporation sector by discipline (1998–2000).

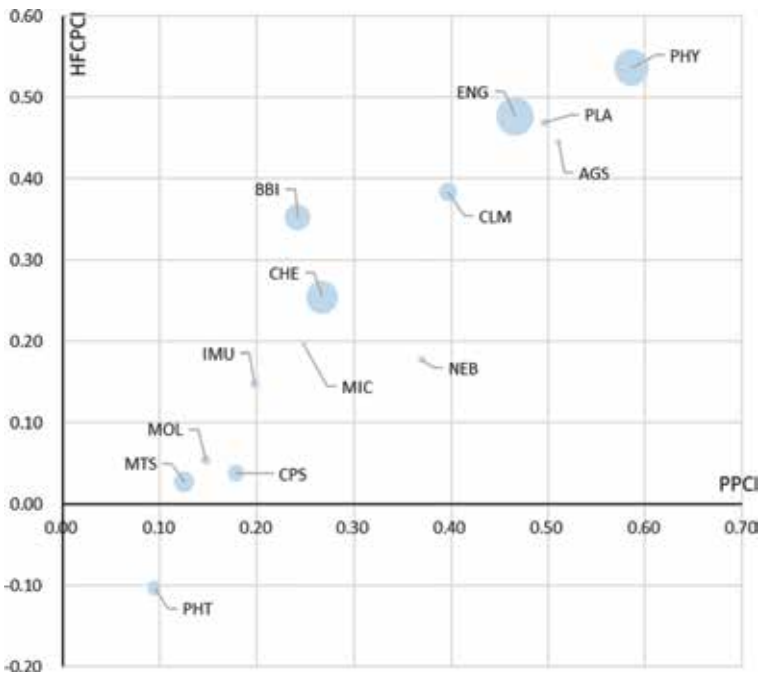


Figure 12. PPCI and HFCPCI of Japanese corporation sector by discipline (1998–2000).

http://dx.doi.org/10.5772/intechopen.77130

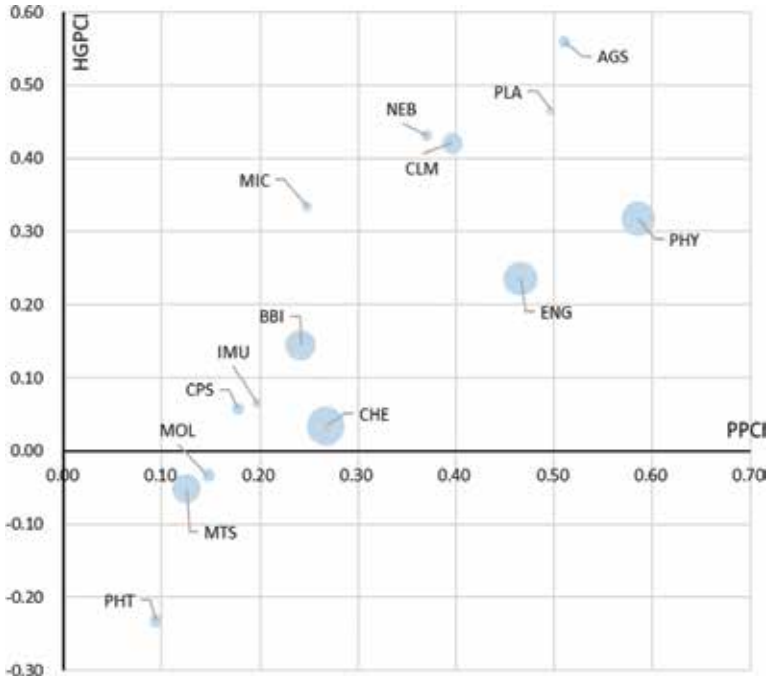


Figure 13. PPCI and HGPCI of Japanese corporation sector by discipline (1998–2000).

For HFCPCI, most disciplines seemed to distribute vertically, suggesting their relatively inconsistent natures in terms of HFCPCI within the sector (**Figure 6**). Two disciplines, immunology and plant and animal science, showed relatively high impact in both PPCI and HFCPCI.

For HGPCI, the university sector seemed to consist of two clusters divided vertically, except for two small disciplines, plant and animal science and computer science (**Figure 7**). The upper cluster consisted of both physical and life sciences, while the lower consisted of life sciences concerning biotechnology.

5.2.2. Public institute

There seemed to be almost no correlation between PPCI and LPFPCI shown in **Figure 8**, and it seemed interesting that relatively smaller circles were located above X-axis while larger circles were opposite. This arrangement was caused by the fact that the disciplines located above the X-axis tended not be cited from the large-sized patent families as a whole, so the Japanese public institute was positioned above average regardless of their small number of papers cited from large-sized patent families.

Most of the disciplines, of which number of papers cited in patents ranked within the top 1% patent-patent forward citations were relatively large, were located on the fourth quadrant (**Figure 9**). Therefore, papers' impact on highly cited patents seemed to be below X-axis totally. This agrees with the coefficient of public institute's patent-patent forward citations, which was below zero as shown in **Table 4**.

For HGPCI shown in **Figure 10**, two relatively large disciplines—Chemistry and Physics—which were located above the X-axis, seemed to make a trend of public institute, because the coefficient of the sector in the column of high patent generality index in **Table 4** was positive.

5.2.3. Corporation

Corporation's prominent performance in both PPCI and HFPPCIs could be seen in **Figures 11–13** in which most disciplines were located on the first quadrant. It was also interesting that all three figures showed a correlation between the two indicators, except for two disciplines (engineering and physics) in **Figure 11**. Therefore, three indicators functioned robustly, regardless of the limited number of papers cited in high-feature-valued patents and corporation's relatively small share of publications in Japan.

Engineering and physics showed opposite impacts in LPFPCI (**Figure 11**) compared to HFCPCI (**Figure 12**) and HGPCI (**Figure 13**). They showed very low values of LPFPCI and limited number of papers cited in large-sized patent families. However, they showed high values of both HFCPCI and HGPCI and relatively large numbers of papers which were cited in patents with the top 1% patent-patent forward citations and with high patent generality index. Although further analysis was needed to show the correct factors of the phenomenon, this might be caused by characteristics of the industries which cited these disciplines.

6. Discussion and conclusion

In this study, three issues were tackled: investigation of the statistical nature of patent-paper citations, development of indicators, and tendencies of Japanese sectors' characteristics concerning patent-paper citations. Here, I discuss the findings and issues needed to be addressed:

1. Investigation in the study revealed the statistical nature of patent-paper citations, i.e., review papers, papers published in high IF journals, and papers highly cited from papers tended to be more cited than papers not so. These characteristics had been reported by previous studies which utilized different datasets and methodologies. Therefore, these results should reveal precise characteristics of patent-paper citations and suggest that fostering excellent scientific research might serve not only science itself but also technological development to some extent.
2. Results of both the logistic regression analysis and analysis by new indicators showed corporation sector's prominence from the view of patent-paper citations. Why were their papers cited more frequently than that of other sectors? To know the reason, identification of patent applicants might be needed, since information on who cited their paper is important to guess the motivation of citations.
3. I showed that (improved) PPCI and HFPPCI could be used to obtain an overview of technological performance of target, whereas there were some problems intrinsic to the rare

and long-tailed nature of citations. If these indicators were used as monitoring tools, a long citation window would be a bottleneck for practical use. Exploring the possibilities of development of methods for shorter-time measurement and to show their availability and limitations should be an important theme.

4. HFPPCI might be inevitably sensitive to small changes in time sequence. Paper citation from high-feature-valued patents is a rarer phenomenon than that from all patents—even the latter is rare. Therefore, only a few citations might yield large changes to values of indicators. Chronological changes of HFPPCIs should be traced to grasp to what extent they are sensitive, and also possibilities for relaxing the threshold to increase samples should be addressed.

Acknowledgements

In the study, I used the Connection Table between "Web of Science Core Collection" (WoSCC) and "NISTEP Dictionary of Names of Universities and Public Organizations," produced by National Institute of Science and Technology Policy.

Author details

Yasuhiro Yamashita

Address all correspondence to: yasuhiro.yamashita@jst.go.jp

Japan Science and Technology Agency, Tokyo, Japan

References

[1] OECD. OECD Science, Technology and Industry Scoreboard 2013. Innovation for Growth. Paris: OECD Publishing; 2013. DOI: 10.1787/sti_scoreboard-2013-en

[2] OECD. OECD Science, Technology and Industry Scoreboard 2015. Innovation for Growth and Society. Paris: OECD Publishing; 2015. DOI: 10.1787/sti_scoreboard-2015-en

[3] NISTEP. Japanese Science and Technology Indicators 2017. NISTEP Research Material No. 261. Tokyo: National Institute of Science and Technology Policy; 2017. DOI: 10.15108/rm261

[4] Branstetter L, Ogura Y. Is academic science driving a surge in industrial innovation? Evidence from patent citations. NBER Working Paper No. 11561; Issued in August 2005

[5] Ahmadpoor M, Jones BF. The dual frontier: Patented inventions and scientific advance. Science. 2017;**357**:583-587. DOI: 10.1126/science.aam9527

[6] Fukuzawa N, Ida T. Science linkages between scientific articles and patents for leading scientists in the life and medical sciences field: The case of Japan. Scientometrics. 2016;**106**:629-644. DOI: 10.1007/s11192-015-1795-z

[7] Waltman L. A review of the literature on citation impact indicators. Journal of Informatics. 2016;**10**:365-391. DOI: 10.1016/j.joi.2016.02.007

[8] Squicciarini M, Dernis H, Criscuolo C. Measuring Patent Quality: Indicators of Technological and Economic Value. OECD Science, Technology and Industry Working Papers, No. 2013/03. Paris: OECD Publishing; 2013. DOI: http://dx.doi.org/10.1787/5k4522wkw1r8-en

[9] Yamashita Y, Jibu M. Exploration of new performance indicator of academic paper citations from patents. JAPIO Yearbook. 2017;**2017**:144-155

[10] Schmoch U. Concept of a technology classification for country comparisons. Final Report to the World Intellectual Organisation (WIPO); June 2008

[11] Hicks D, Breitzman A Sr, Hamilton K, Narin F. Research excellence and patented innovation. Science and Public Policy. 2000;**27**:310-320. DOI: 10.3152/147154300781781805

[12] Branstetter L. Exploring the link between academic science and industrial innovation. Annales d'Économie et de Statistique. 2005;**79**(80):119-142

[13] Thomson Reuters (present Clarivate Analytics) [Internet]. InCites Indicator Handbook. Available from: http://ipscience-help.thomsonreuters.com/inCites2Live/8980-TRS/version/default/part/AttachmentData/data/InCites-Indicators-Handbook-6%2019.pdf [Accessed: September 20, 2017]

[14] Guan J, He Y. Patent-bibliometric analysis on the Chinese science - technology linkages. Scientometrics. 2007;**72**:403-425. DOI: 10.1007/s11192-007-1741-1

Section 4

Content-based Analysis

Chapter 11

Mapping Science Based on Research Content Similarity

Takahiro Kawamura, Katsutaro Watanabe,
Naoya Matsumoto and Shusaku Egami

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77067

Abstract

Maps of science representing the structure of science help us understand science and technology development. Thus, research in scientometrics has developed techniques for analyzing research activities and for measuring their relationships; however, navigating the recent scientific landscape is still challenging, since conventional inter-citation and co-citation analysis has difficulty in applying to recently published articles and ongoing projects. Therefore, to characterize what is being attempted in the current scientific landscape, this article proposes a content-based method of locating research articles/projects in a multi-dimensional space using word/paragraph embedding. Specifically, for addressing an *unclustered* problem, we introduced cluster vectors based on the information entropies of technical concepts. The experimental results showed that our method formed a clustered map from approx. 300 k IEEE articles and NSF projects from 2012 to 2016. Finally, we confirmed that formation of specific research areas can be captured as changes in the network structure.

Keywords: map of science, content-based, paragraph vector, information entropy, clustering

1. Introduction

In 1965, Price [1] proposed studying science using scientific methods. Since then, research in scientometrics has developed techniques for analyzing research activities and for measuring their relationships and constructed maps of science, one of the major topics in scientometrics, that provides a bird's eye view of the scientific landscape. Maps of science have been useful tools for understanding the structure of science, their spread, and interconnection of disciplines. By knowing such information, science, and technology enterprises can

anticipate changes, especially those initiated in their immediate vicinity. Research laboratories and universities that are organized according to the established standards of disciplinary departments can understand an organization's environment. Furthermore, such maps are important to policy analysts and funding agencies. Since research funding should be based on quantitative and qualitative scientific metrics, they usually perform several analyses on the map with statistical analysis and careful examination by human experts. However, conventional approaches to understanding research activities focus on what authors told us about past accomplishments through inter-citation and co-citation analysis of published research articles. Thus, ongoing project and the recently published articles that do not have enough citations have not been analyzed.

Therefore, we propose to analyze them using a content-based method using natural language processing (NLP) techniques. Recently, word/paragraph embedding has been proposed for finding relationships between unstructured descriptions. Such embedding techniques represent words and paragraphs as real-valued vectors of several hundred dimensions. The distances between the descriptions are calculated from the similarities between vectors. Thus, we constructed a new mapping tool that represents the recent scientific trends, where nodes represent research projects or the articles that are linked by certain distances of the content similarity. Moreover, we drew a map from approx. 300,000 IEEE articles and National Science Foundation (NSF) projects, and then from its chronological changes we obtained some findings regarding the formation processes of research areas.

The remainder of this chapter is organized as follows. In Section 2 discusses related work, and Section 3 describes our proposed method for calculating the content similarity and its evaluations. Then, Section 4 introduces our tool, Mapping Science, and we confirm on the map the formation process of research areas such as the Internet of Things in Section 5, final conclusions and suggestions for future work are provided in Section 6.

2. Related work

Maps of Science (http://mapofscience.com/) are a well-known website. Katy et al. also provides Sci2Tool visualization tools [2] and maps of journals and documents [3]. In Japan, National Institute of Science and Technology Policy (NISTEP) provides Science Map (http://www.nistep.go.jp/wp/wp-content/uploads/ScienceMapWebEdition2014.html). In such studies, the similarity between journals and articles is calculated based on the cosine and/or Jaccard similarity of inter-citation and co-citation. These maps promote interdisciplinary research collaboration, but citation analysis cannot be utilized for ongoing projects and recently published articles, although project descriptions will eventually include articles in their research results.

Funding agencies and publishers generally have their own classification systems. Projects/articles have more than one code; thus, interdisciplinary projects can be found by searching multi-labeled projects. However, even if two projects/articles are assigned the same category, their similarity may not be found. Moreover, funding agencies and publishers use different categories, and there is no comprehensive scheme for characterizing projects or articles; thus, they cannot be compared between different agencies or publishers. For example, comparing

articles with Association for Computing Machinery classification (https://www.acm.org/publications/class-2012) with Springer Nature classification requires taxonomy exchanges.

Therefore, several content-based methods are proposed in the related literature. Previous studies have examined automatic topic classification using probabilistic latent semantic analysis (pLSA) [4] and latent Dirichlet allocation (LDA) [5]. One uses LDA to find the five most probable words for a topic, and each document is viewed as a mixture of topics [6]. This approach can classify documents across different agencies and publishers. However, the similarity between projects/articles cannot be computed directly. In this regard, the National Institutes of Health (NIH) Visual Browser [7, 8] (http://nihmaps.org/index.php) computed the similarities between projects as the mixture of classification probability to each topic based on pLSA, using the average symmetric Kullback-Leibler divergence function [9]. However, this similarity is a combination of probabilities; that is, it is not derived from sentence context. Other studies are also based on the similarity between sets of words (bag-of-word) included in documents like term frequency-inverse document frequency (TF-IDF), and not considering the sentence context.

By contrast, a word/paragraph vector, which is a distributed representation of words and paragraphs, is attracting attention in NLP. Assuming that context determines the meaning of a word [10], words appearing in similar contexts are considered to have a similar meaning. In the basic form, a word vector is represented as a matrix, whose elements are the co-occurrence frequencies between a word w with a certain usage frequency in the corpus and words within a fixed window size c from w. A popular representation of word vectors is word2vec [11, 12]. Word2vec creates word vectors using a two-layered neural network obtained by a skip-gram model with negative sampling. Specifically, word vectors are obtained by calculating the maximum likelihood of objective function L in Eq. (1), where T is the number of words with a certain usage frequency in the corpus. Word2vec clusters words with similar meanings in a vector space.

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} \log p(w_{t+j} \mid w_t) \tag{1}$$

In addition, Le and Mikolov [13] proposed a paragraph vector that learns fixed-length feature representations using a two-layered neural network from variable-length pieces of texts such as sentences, paragraphs, and documents. A paragraph vector is considered another word in a paragraph and is shared across all contexts generated from the same paragraph but not across paragraphs. The contexts are fixed length and sampled from a sliding window over the paragraph. The paragraph vectors are computed by fixing the word vectors and training the new paragraph vector until convergence, as shown in Eq. (2).

$$L = \sum_{t=1}^{T} \log p\left(w_t \mid w_{t-c}, \ldots, w_{t+c}, d_i\right) \tag{2}$$

where d_i is a vector for a paragraph i that includes w_t. Whereas word vectors are shared across paragraphs, paragraph vectors are unique among paragraphs and represent the topics of the paragraphs. By considering word order, paragraph vectors also address the weaknesses of bag-of-words models in LDA and pLSA. Therefore, paragraph vectors are considered more accurate representations of the context of the content. We can then input resulting vectors

into the analysis using machine learning and clustering techniques for finding similar articles in different academic subjects as well as the relationships between projects from different agencies. Thus, we tried to convert the natural sentences in project descriptions and article abstracts to paragraph vectors in this study.

3. Paragraph embedding using information entropy

This section introduces our proposed paragraph embedding method using entropy and then evaluates whether the similarity of the resulting vectors accurately represents the content similarity of documents.

3.1. Proposal of the paragraph embedding method

Before introducing the proposed method, we present a problem in applying the paragraph vectors for research project descriptions. We implemented the paragraph embedding technique using the Deep Learning Library for Java (https://deeplearning4j.org). Then, we constructed paragraph vectors for approx. 30,000 NSF projects mentioned in the next section. Although we need a more systematic way, but this time the hyperparameters were set empirically as follows: 500 dimensions were established for 66,830 words that appeared more than 5 times; the window size c was 10, and the learning rate and minimum learning rate were 0.025 –0.0001, respectively, with an adaptive gradient algorithm. The learning model is a distributed memory model with hierarchical softmax.

However, the result showed that projects are scattered and not clustered by any subject or discipline in the vector space. Most projects are slightly connected to a low number of projects. Thus, it is difficult to grasp trends and compare an ordinary classification system. Closely observing the vector space reveals some of the reasons for this *unclustered* problem: each word with nearly the same meaning has slightly different word vectors, and shared but unimportant words are considered the commonality of paragraphs. In fact, Le and Mikolov reported classification accuracy with multiple categories of less than 50% [13].

Therefore, for addressing this problem, we introduce the information entropy [14] for clustering word vectors before constructing paragraph vectors. The fact that synonyms tend to gather in a word vector space indicates that the semantics of a word spatially spread to a certain distance. This observation is also suggested in the related literature [15]. Therefore, to unify word vectors of almost the same meanings, excluding trivial common words, we generated clusters of the word vectors based on the semantic diversity of each concept in a thesaurus. We first extracts 19,685 hypernyms (broader terms) with one or more hyponym (narrower term) from the Japan Science and Technology Agency (JST) science and technology thesaurus [16]. The JST thesaurus primarily consists of keywords that have been frequently indexed in 36 million articles accumulated by the JST since 1975. Currently, this thesaurus is updated every year and includes 276,179 terms with English and Japanese notations in 14 categories from bioscience to computer science and civil engineering. Based on the World Wide Web Consortium (W3C) Simple Knowledge Organization System (SKOS), the JST thesaurus

also exists in W3C Resource Description Framework (RDF, https://www.w3.org/RDF/) format with semantic relationships SKOS: broader, SKOS: narrower, and SKOS: related. A broader or narrower relationship essentially represents an *is-a* subsumption relationship but sometimes denotes a *part-of* relationship in geography, body organ terminology, and other academic disciplines. The JST thesaurus is publicly accessible from Web APIs on the J-GLOBAL website (http://jglobal.jst.go.jp/en/), along with the visualization tool Thesaurus Map (http://thesaurus-map.jst.go.jp/jisho/fullIF/index.html). We then calculate the information entropy of each concept in the JST thesaurus from the dataset. Shannon's entropy in information theory is an estimate of event informativeness. We used this entropy to measure the semantic diversity of a concept [17]. After creating clusters according to the degree of entropy, we unify all word vectors in the same cluster to a cluster vector and constructed paragraph vectors based on the cluster vectors. The overall flow is shown in **Figure 1**.

Hereafter, the "word" is a word in the dataset, the "term" is a term in a thesaurus, and terms are classified into hypernyms, hyponyms, and their synonyms. The "concept" is defined as a combination of a hypernym and one or more hyponyms one level below the hypernym indicated as a red box in **Figure 2**. Given that a thesaurus consists of terms T_i, we calculated the entropy of a concept C by considering the appearance frequencies of a hypernym T_0 and its hyponyms $T_1 \ldots T_n$ as an event probability. The frequencies of synonyms $S_{i0} \ldots S_{im}$ of term T_i was summarized to a corresponding concept (synonyms S_{ij} include descriptors of terms T_i themselves).

$$H(C) = -\sum_{i=0}^{n}\left(\sum_{j=0}^{m} p\left(S_{ij} \mid C\right) \cdot \log_2 \sum_{j=0}^{m} p\left(S_{ij} \mid C\right)\right) \tag{3}$$

In Eq. (3), $p(S_{ij}|C)$ is the probability of a synonym S_{ij} given a concept and terms T_i. For each concept in the thesaurus, we calculated the entropy $H(C)$ in the dataset. As the probabilities of events become equal, $H(C)$ increases. If only particular events occur, $H(C)$ is reduced because of low informativeness. Thus, the proposed entropy of a concept increases when a hypernym

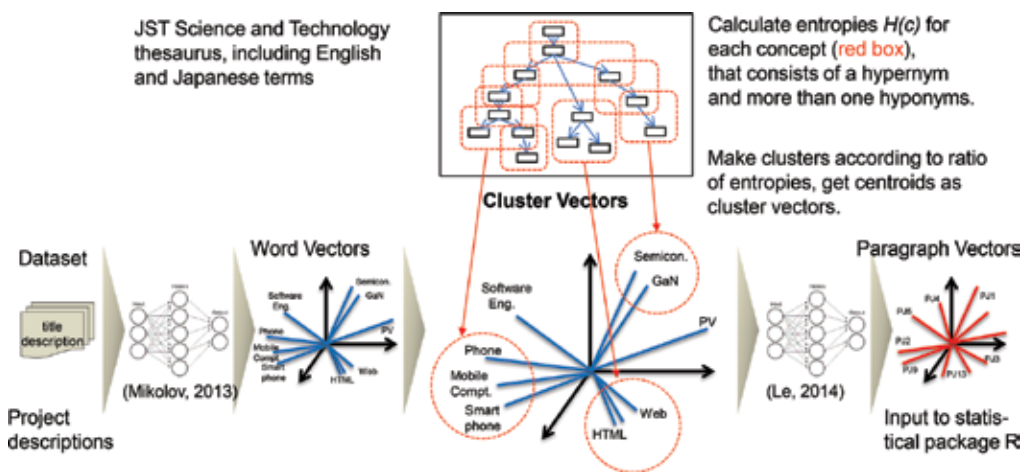


Figure 1. Construction of paragraph vectors based on cluster vectors.

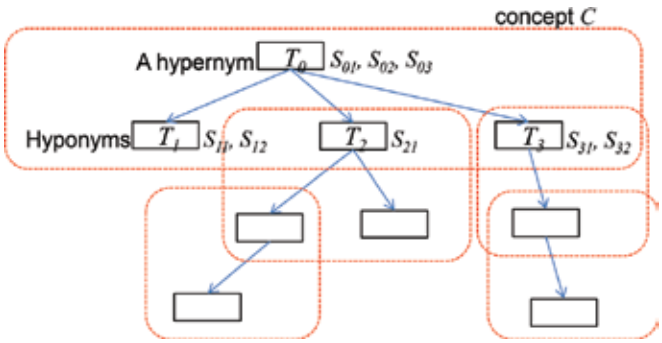


Figure 2. Concepts in a thesaurus.

and hyponyms that construct a concept separately appear with a certain frequency in the dataset. Therefore, the degree of entropy indicates the semantic diversity of a concept. Then, assuming that the degree of entropy and the spatial size of a concept in a word vector space are proportional to a certain extent, we split the word vector space into clusters. In fact, our preliminary experiment indicated that the entropy of a concept has high correlation $R = 0.602$ with the maximum Euclidean distance of hyponyms in the concept in a vector space, at least while the entropy is rather high. Specifically, we refined clusters by repeatedly subdividing them until the defined criterion was satisfied. In our method, we set the determination condition as shown in Eq. (4).

$$Cl(w_k) = \begin{cases} Cl(w_i) & \left(\frac{H(C(w_i))}{H(C(w_j))} > \frac{\| w_k - w_i \|}{\| w_k - w_j \|} \right) \\ Cl(w_j) & \text{(otherwise)} \end{cases} \tag{4}$$

This condition represents that the word vectors $w_0 \ldots w_T$ are subdivided into two clusters proportionally to the ratio of the highest two concept entropies $H(C(w_i))$ and $H(C(w_j))$, which are selected from all entropies of concepts in a cluster (an initial cluster is the whole vector space). $C(w_i)$ and $C(w_j)$ mean concepts C to which words w_i and w_j belong, respectively. The words w_i and w_j are words, whose lemmatized forms are identical to terms or synonyms in the thesaurus. However, note that the entropies of the other words whose correspondences are not included in the thesaurus are not calculated in Eq. (3). $Cl(w)$ means a cluster to which a vector of a word w should be classified.

The vector space is subdivided until the entropy becomes lower than 0.25 (the top 1.5% of entropies) or the number of elements in a cluster is lower than 10. These parameters were also determined empirically through the experiments. After generating 1260 clusters from 66,830-word vectors, we considered the centroid of all vectors in a cluster as a cluster vector. Then, we constructed paragraph vectors using the cluster vectors rather than word vectors, as shown in Eq. (5) that is an extension of Eq. (2). After all, each cluster vector represents a concept that has the highest entropy in all concepts included in the cluster.

$$L = \sum_{t=1}^{T} \log p\left(Cl(w_t) \mid Cl(w_{t-c}), \ldots, Cl(w_{t+c}), d_i \right) \tag{5}$$

3.2. Evaluation of paragraph vectors

Next, we evaluate the resulting vectors on the map constructed from the following dataset. In this article, the dataset includes titles and abstracts of 266,772 IEEE conference articles published from 2012 to 2016, including 2,290,743 sentences in total and titles and descriptions of 34,192 NSF projects from 2012 to 2016, including 730,563 sentences in total. Note that IEEE journal, transaction, symposium, and workshop articles are not included, and NSF project domains are limited to Computer and Information Science and Engineering, Mathematical and Physical Sciences, and Engineering in accordance with IEEE articles. All words in the sentences were tokenized and lemmatized by Stanford CoreNLP before creating the vector space.

In terms of the *unclustered* problem, we confirmed that the proposed method successfully formed several clusters compared with the original paragraph embedding method. For a quantitative comparison, in **Figure 3** shows the relationships between the cosine similarities and the number of edges, and the relationship between the degree centrality and the number of nodes (i.e., projects) in the case of the cosine similarities of >0.35. As a result, we confirmed that edges with a higher cosine similarity and nodes with higher degrees increase. The reason for this result is because, through the use of high-entropy concepts, which are significant in scientific and technological contexts excluding scientifically unimportant words—as elements between paragraph vectors, the paragraph vectors were able to comprise meaningful groups. Simultaneously, newly, unknown synonyms, and closely related words that are not defined in the thesaurus can be unified to a cluster vector, if they are in the same cluster. Taking the centroid vector as a representative vector in a cluster involves separating each cluster vector as much as possible to form a clear difference in the vector space.

In terms of the accuracy of content similarities, the evaluation encounters difficulty since, to the best of our knowledge, there is no gold standard for evaluating the similarity among scientific and technological documents. Therefore, we first evaluated the degree of the similarities based on a sampling method. We randomly extracted 100 pairs of projects with a cosine similarity of >0.5 (similarities less than 0.5 are not considered in the map layout), to make the distribution similar to the entire distribution. Each pair has two project titles and descriptions, and a cosine value that is divided into three levels: weak ($0.5 \leq \cos. < 0.67$), middle ($0.67 \leq \cos. < 0.84$), and strong ($0.84 \leq \cos.$). Some examples of two projects and their cosine value are shown in **Table 1**. Then, three members of our organization, a funding agency in Japan, evaluated the similarity of each pair. The members were provided the prior explanations for the intended use of the map

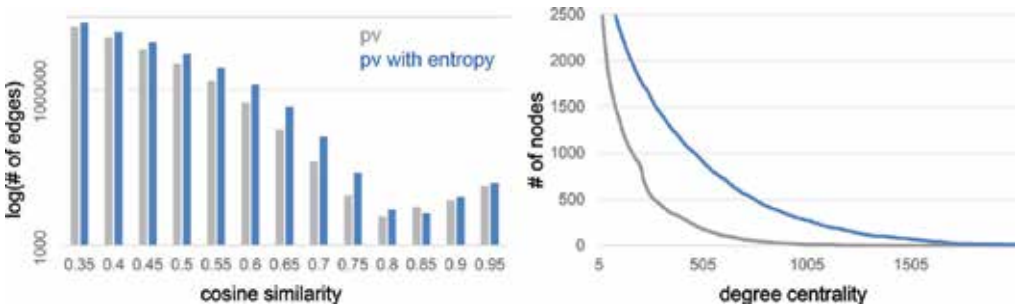


Figure 3. Comparison between paragraph vectors and those with entropy clustering.

title / (desc.)	*cos.*	*title / (desc.)*
understanding the physics of galaxy formation and evolution at high redshift / understanding the processes regulating galaxy…	0.50 (weak)	the birth of the first stars and galaxies / the aim of this proposal is to simulate the formation and evolution of galaxies within the…
asymptotic graph properties / many parts of graph theory have witnessed a huge growth over the last years, partly because of their relation to theoretical computer science and statistical physics. …	0.52 (weak)	benjamini-schramm approximation of groups and graphings / large graphs have become central objects in many fields in the last couple of decades: in neural sciences, network sciences…
a high intensity neutrino oscillation facility in Europe / the recent discovery that the neutrino changes type (or flavour) as it travels through space, a phenomenon referred to as neutrino oscillations,…	0.53 (weak)	probing fundamental properties of the neutrino at the sno+ experiment / i propose a comprehensive programme of research on sno+, a multi-purpose neutrino experiment that has the capacity…
systems biology of pseudomonas aeruginosa in biofilms / systems biology is a new and rapidly growing discipline . it is widely…	0.54 (weak)	cyclic-di-gmp: new concepts in second messenger signaling and bacterial biofilm formation / biofilms represent a multicellular…
investigation of human nucleoporins stoichiometry and intracellular distribution by quantitative mass spectrometry / the nuclear pore complex (npc) is one of the most intricate multi-protein…	0.56 (weak)	atlas of cell-type specific nuclear pore complex structures / the nuclear pore complex (npc) is one of the most intricate components of eukaryotic cells and is assembled from ~30 nucleoporins…
european science and technology in action building links with industry, schools and home / the aim of establish is to facilitate and implement an inquiry based approach in the teaching and learning…	0.67 (middle)	science teacher education advanced methods / helping teachers raise the quality of science teaching and its educational environment has the potential to increase student engagement,…
support to tenth european conference on turbomachinery - fluid dynamics and thermodynamics, lappeenranta, finland, 15-19 march 2013 / the european turbomachinery conference is…	0.99 (strong)	support to ninth european conference on turbomachinery - fluid dynamics and thermodynamics, istanbul, turkey, 21-25 march 2011 / the european turbomachinery conference is…

Table 1. Example of sampled projects/articles.

and some examples of evaluation. The members received the same data, and their backgrounds are bioscience, psychology, and computer science. As a result, we confirmed that 78% of the similarities matched majority votes of the members' opinions. Examples misjudged include, for example, the not related pairs of two projects that have the same acronyms with different meanings, and the stronger pairs of two projects that have only a few common words, but which are recent technologies attracting attention. We expect that those words will eventually have higher entropies and then the project similarities will be estimated to be stronger. We also plan to replace acronyms in project descriptions with full words before making vectors. By contrast, the accuracy of the similarities of the original paragraph embedding method was 21%. The evaluation results were determined to be in "fair" agreement (Fleiss' Kappa $\kappa = 0.29$) (**Table 2**).

Moreover, we evaluated the accuracy of content similarities using the artificial data, part of which is randomly replaced with the other projects/articles. We replaced 10, 20, …, 100% of

Similarity	**Weak**	**Middle**	**Strong**
Precision	77.5	83.3	100.0
Recall	98.6	33.3	83.3
F1 value	86.8	47.6	90.9

Table 2. Evaluation of similarity based on sampling (%).

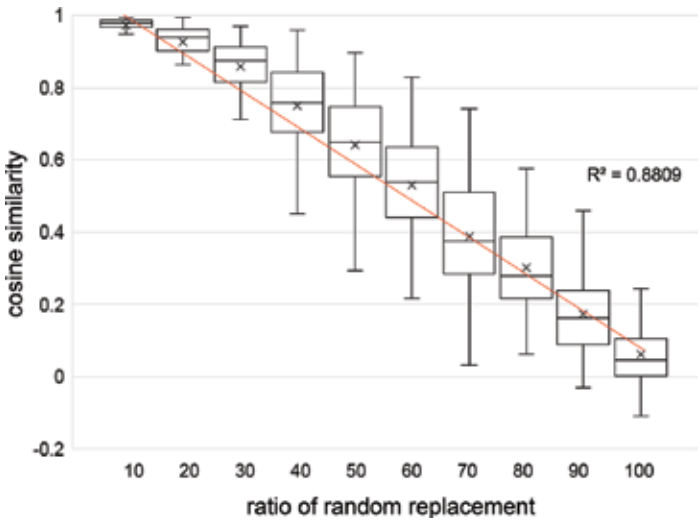


Figure 4. Cosine similarities of artificial data with partial replacement.

a project description or a article abstract with sentences randomly selected from the others. Then, we measured a cosine similarity between a vector generated from the artificial project/article and a vector of the original project/article. The projects/articles were randomly selected from all projects/articles, and then we evaluated 1000 pairs of the original project/article and the artificial project/article. The relationship of the replacement ratios and the cosine similarities is shown in **Figure 4**. As a result, we confirmed that there is an obvious correlation between content similarities of projects/articles and their cosine similarities with $R^2 = 0.89$. The paragraph vectors without the entropy clustering also had the same trend, but the vectors with the entropy clustering had higher similarities on average. This result matches the relationships between the cosine similarities and the number of edges shown in **Figure 3**.

4. Mapping Science

This section describes our content-based map of science, Mapping Science [18, 19]. After introducing its interface, we describe our clustering and layout method of articles and projects in the map and analytical functions provided.

4.1. Interfaces

In **Figure 5** shows three main views of the Mapping Science, which are a portfolio view, a clustered view, and analytic views.

In the portfolio view, five research areas, Information, Mathematics and Physics, Communication, Electronics and Mechatronics, and Power and Energy, to which the entire dataset has been divided by full-text search with predefined queries, are shown. The size of circles corresponds to the number of articles and projects in the area.

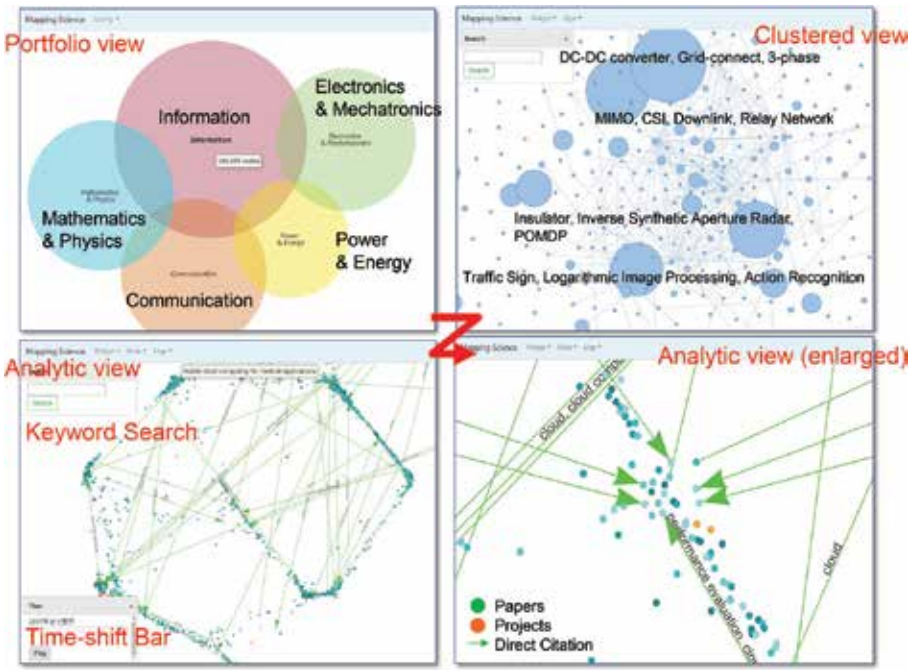


Figure 5. Interface of Mapping Science.

In the clustered view that opens when users click one of the areas in the portfolio view, the results of clustering all the articles and projects in the area are shown. The details of the clustering method are shown in the next section. This view is for taking a look at the technologies in the area. Each cluster has at most 10 labels, which are extracted as feature phrases using a probabilistic information retrieval method, BM25 [20].

In the analytic view that opens when users click one of the clusters in the clustered view, each node corresponds to a article or a project, and distances between the nodes are proportional to the cosine similarities between articles/projects, as much as possible. In addition, direct citation links between articles (citing → cited) are shown in light green edges with labels showing common phrases between two articles, which are also extracted by the BM25 method. When users click a node, the detailed information about the node (article or project) is shown on the map.

In all the views, the search box located at the upper-left corner provides full-text search for all articles and projects included in the current view, and the search results are highlighted in the view. Moreover, the analytic view provides the time-shift bar, which displays the cumulative changes in a cluster according to published/started years of articles/projects. The trial version of this map is publicly available at https://jipsti.jst.go.jp/foresight/.

4.2. Clustering and layout method of the nodes

In this section, we describe a method for generating the clustered view and the analytic view. There are too many nodes (articles and projects) even in a research area to explore a specific research topic (over 160,000 nodes included in the Information area in **Table 3**). We thus

	Information	Mathematics and Physics	Communication	Electronics and Mechatronics	Power and Energy
# of nodes	165,823	113,982	99,995	88,023	89,845
# of clusters	474	345	338	400	303
# of clusters (only by infomap)	2313	1614	1630	2807	1776

Table 3. # of nodes and clusters in each research area.

divided them into several hundred clusters and provided analytic functions described in the next section to explore articles and projects in each cluster.

A major concern in clustering and laying out the nodes is to reduce 500-dimensional paragraph vectors to a 2D network structure. In general, conventional clustering or dimension reduction techniques such as multi-dimensional scaling (MDS) have $O(n^3)$ computational complexity, which increases the calculation time in proportion to that. We thus, to accommodate the practical calculation time, generated a network structure only from the edges that are the 30 highest similarities (at least, 0.5 or more) to other nodes. Sci2Tool [3] also generated the network only from the 15 highest similarities edges and successfully created an informative map of journals.

Clusters in the clustered view are calculated by info map [21], which is one of modularity-based network clustering algorithms [22]. By increasing the modularity, the nodes are divided into clusters that have more edges within the clusters than edges between the clusters. Thus, articles or projects in a cluster have relatively high similarities and form meaningful sets. However, the simple application of the info map generated too many clusters to explore the clustered view (over 2800 clusters included in Electronics & Mechatronics area in **Table 3**). Therefore, we merged small clusters comprised of less than 50 nodes into the nearest cluster, which has the highest similarity pair between any of two nodes in the clusters. This operation corresponds to a single linkage clustering in agglomerative clustering. As a result, the numbers of clusters are reduced as in **Table 3**. Although the accuracy of the clustering result falls (the modularity decreases), nodes incorporated into the nearest cluster tend to form independent sets of nodes in the analytic view and can be distinguished in the view. The distances between clusters in the clustered view mean the distances in the single linkage-clustering.

The layout algorithm in the analytic view is OpenOrd (formally, DrL) [23]. This is a well-known force-directed layout algorithm and frequently used in other maps of science such as Sci2Tool. In **Figure 6** shows a comparison of layout algorithms for Internet of thing cluster (see the next section), which includes the OpenOrd (edge cut parameter: 0.88, 0.91, and 0.94), MDS with cosine dissimilarity, large graph layout (LGL) [24] and Fruchterman Reingold layout (FR) [25]. The LGL and the FR are also force-directed algorithms. We can obviously confirm several clusters in the OpenOrd, but those are not clear in the other algorithms. The number of clusters in the OpenOrd increase as the edge cut parameter increases. Thus, we empirically set the OpenOrd with the edge cut parameter: 0.91 in the analytic view by default. The other parameters were also empirically set to show the structural features as much as possible. However, as shown in the next section, the analytic view provides several other layout algorithms and parameters; thus, users can change the layout of nodes according to their needs.

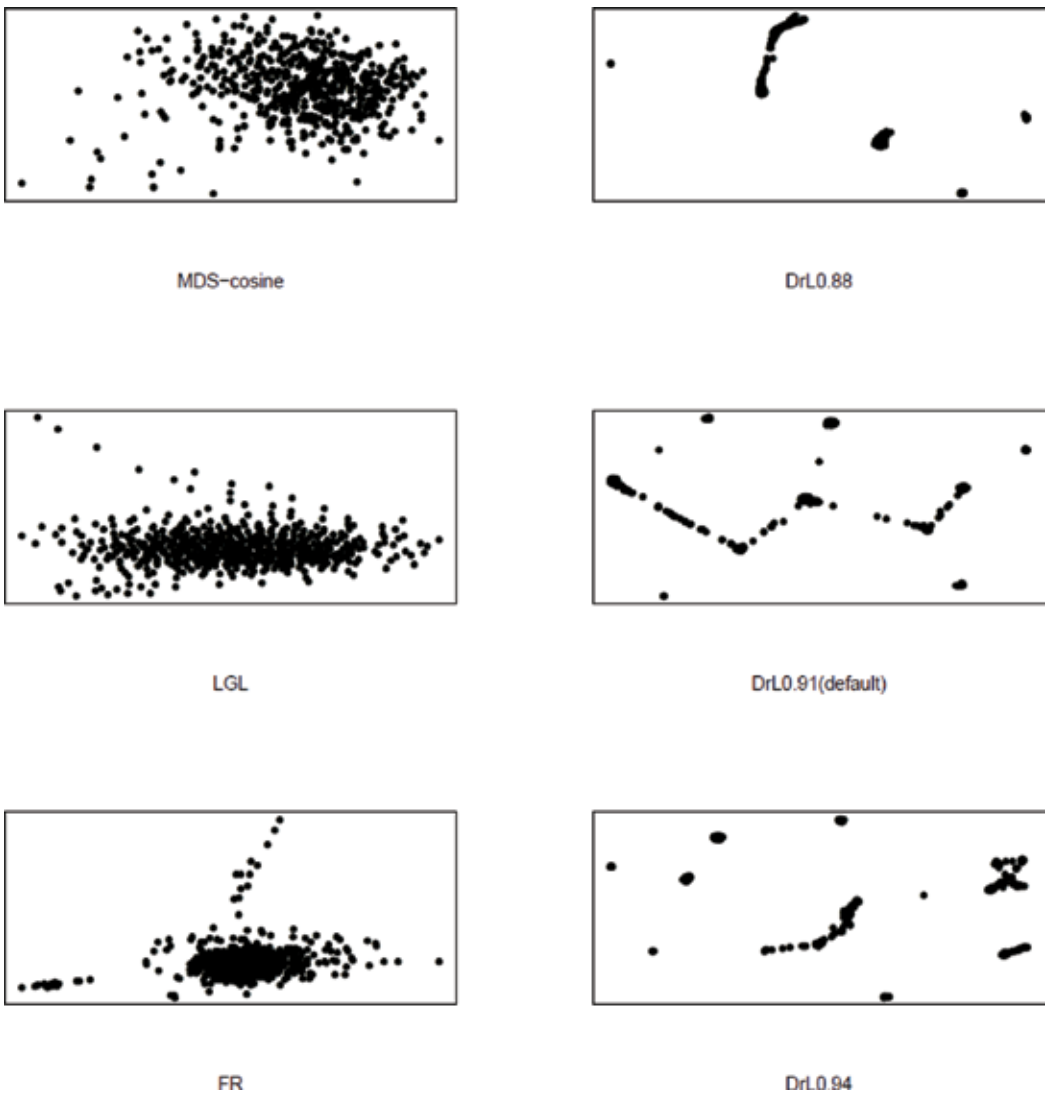


Figure 6. Comparison of graph layout algorithms.

4.3. Analytical functions provided on the map

In addition to the functions described in Section 4.1, the Mapping Science provides the following analytical functions: (1) translation of article abstracts and project descriptions, (2) visualization of statistical information, (3) summarization of feature phrases, (4) querying and exporting using SPARQL, (5) change of layout algorithms, and (6) generation of customized analytic views.

4.3.1. Abstract/description translation function

In the analytic views, users can see the detailed information, such as titles, article abstracts/ project descriptions, authors/project members, affiliations, and publication year/proposed

year. In addition, the abstracts/descriptions are translated into Japanese by clicking "Translate" buttons. The users can read the original abstracts/descriptions in the same pane for confirming the translation validity.

4.3.2. Visualization function of statistical information

As in **Figure 7**, the analytic view can visualize the summary of bibliometric information of the nodes contained in the view. There are several widgets, such as for citation (Impact Factor, SJR, and CiteScore) metrics, publications by year, citations by year, and publications by each country. Moreover, the users can select the nodes in a rectangle area and see the statistical information of the selected nodes. The upper part of the publication by country shows an article count (AC) (https://www.natureindex.com/faq). The AC means the country-level participation in a study, where a country is counted if one or more authors of the article are from the country. For example, if countries of three authors' affiliations in a article are A, B, and B, A is counted as one and B is also counted as one. In contrast, the lower part of the publication by country shows a fractional count (FC) that means the contribution of each country. In the above example, A becomes 1/3, B becomes 2/3.

4.3.3. Summarization function of feature phrases

As in **Figure 8**, the feature phrases of the selected nodes can be summarized in word clouds. At most 10 feature phrases of each node are extracted based on the BM25 method in advance. Then, if the users select the multiple nodes, the feature phrases with higher frequencies are displayed larger and placed closer to the center of the word cloud. This function is useful for understanding specific themes of the selected nodes in a cluster.

Figure 7. Statistical information.

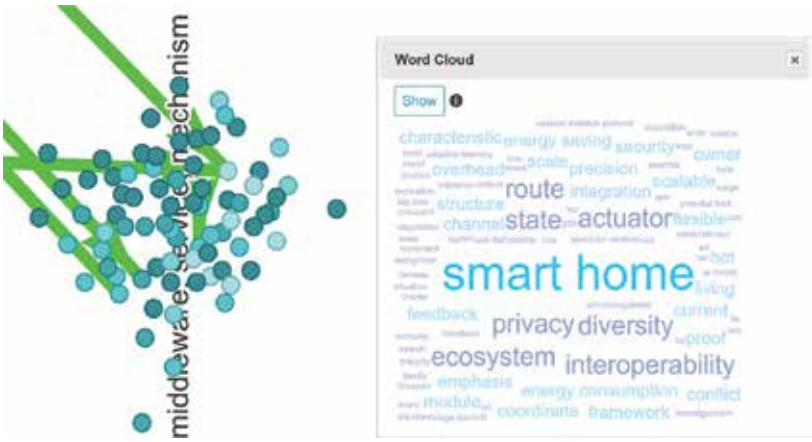


Figure 8. Feature phrases in the selected nodes.

4.3.4. Query function and export function

The background data in the Mapping Science have been converted to RDF data and stored in a graph database. Therefore, the analytic views provide a high-level search using a formal query language, SPARQL, as in **Figure 9**. For example, the users can search for articles, which have >0.8 similarities with articles cited 100+ times from journals with >10 impact factor (such articles might be obscure but important). When the users click a node ID in the result table,

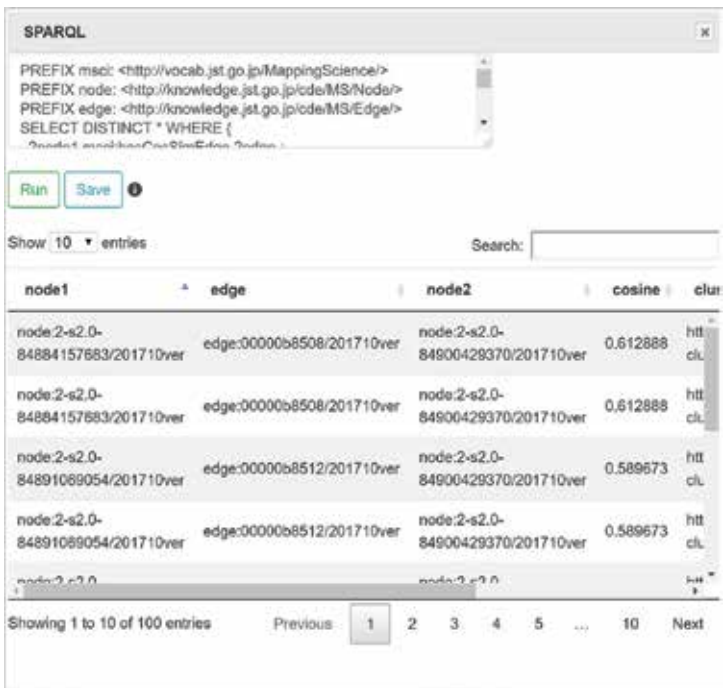


Figure 9. SPARQL search widget.

the node is highlighted and the viewpoint is automatically moved to the node. Moreover, the users can store their own SPARQL queries as macros. Therefore, users who are not familiar with SPARQL can simply call the macros and obtain the query results.

In addition, since we received requests for downloading the information displayed on the map, the information of the selected nodes and all nodes in a cluster can be exported in comma-separated values (CSV) format. The result of SPARQL queries can be also exported in CSV format.

4.3.5. Layout change function

As described in the previous section, the layout of the analytic view was calculated by the OpenOrd (edge-cutting value: 0.91). In addition to that, the analytic views can be redrawn by the OpenOrd (edge-cutting value: 0.94 or 0.88), LGL, Fruchterman-Raingold, or Kamada-Kawai [26]. When the users select a layout, the layout algorithm is executed in the background, the resulting layout information is stored and the view is redrawn. If the layout information is stored in advance, the layout is redrawn immediately. The layout calculation time depends on the number of nodes, and the average time is a few seconds to a few minutes.

4.3.6. Custom analytic view function

The analytic views were composed by the info map algorithm, but the users can create the customized Analytic views by keyword search. When the users enter keywords into the widget in the portfolio view, the nodes are extracted by the full-text search for all nodes in five research areas, and then the layout is calculated by the OpenOrd based on the cosine similarities of the extracted nodes. For example, an analytic view for studies related to neural networks and artificial intelligence across multiple research areas can be created by keywords such as "Artificial Intelligence [AND] Neural Network." This function could help find interdisciplinary studies. The calculation time depends on the number of nodes, and the average time is a few seconds to a few minutes. The information on the customized analytic views is stored in the background; the same view is immediately displayed for the second time. The customized analytic view can provide the same analytical functions, such as keyword search, visualization of statistical information, visualization of the cumulative changes by year, and layout change.

5. Case study for the formation process of research areas

In this map, we try to understand the formation processes of several research areas through chronological changes of network structure. This section describes two cases for the Internet of Things (IoT) and Brain-Computer Interface (BCI).

In **Figure 10** shows the analytic views for an IoT area from 2012 to 2016, which includes 574 nodes as of 2016. The last view is the analytic view in 2016 displaying >0.6 cosine similarities as edges.

In 2012, four islands (places, at which nodes are densely located) mainly for IoT frameworks and networks and for IoT system and security are barely found (labels of each island have been extracted by the summarization function of feature phrases).

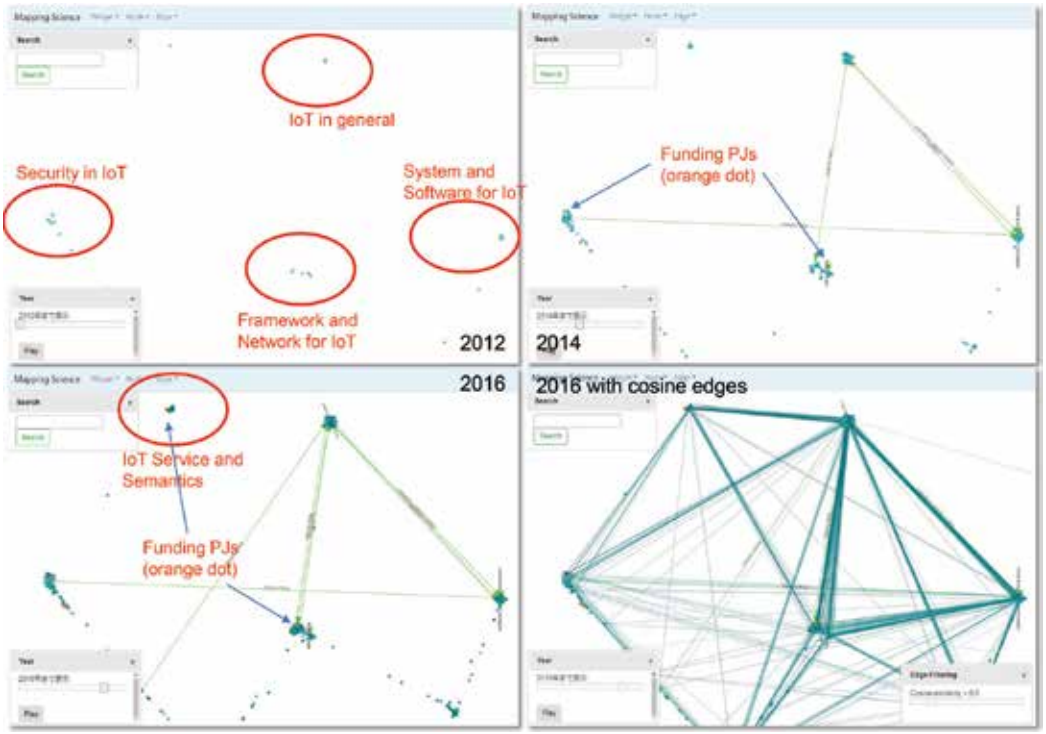


Figure 10. Formation of IoT areas.

In 2013, a funding project (orange node) was firstly established in the security, and then the corresponding island grew bigger, that is, the number of articles increased, although a causal relationship is unclear.

Then, in 2014, the island of the IoT frameworks and networks also had a funding project and grew bigger. At the same time, researchers of each island, which seem to correspond to the different research community, started to recognize with each other, and thus mutual citation links (light green edges) between islands began to be drawn.

In 2015 and 2016, this movement was accelerated; thus, we can confirm that the islands were getting bigger and denser, and mutual citation links increased. Moreover, the other islands than the first four islands, for example, an island for IoT services and semantics at the upper-left corner also gradually grew, and some of them are greatly increasing the articles by getting funding projects.

Finally, the edges of the cosine similarity 0.6 in the last view mean relatively weak similarity described in Section 3.2. In contrast, nodes which compose an island are mutually connected with stronger similarities, although they are too dense to confirm in the figure. Therefore, in this IoT area, there are several research communities dedicated to specific research themes, and they are mutually connected with their content similarity and citation relations. Thus, we can understand that they are developing each theme while forming the IoT area as a whole.

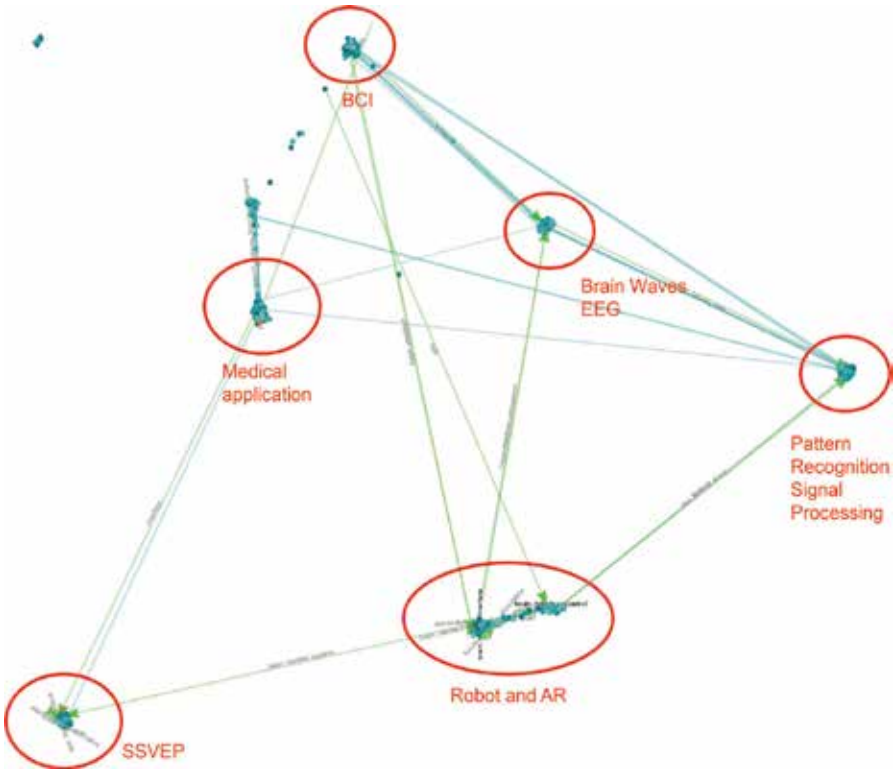


Figure 11. Formation of BCI.

We confirmed several other processes of research area formation in our case studies. For example, in **Figure 11** shows the analytic view for BCI in 2016. In this figure, an island at the top is growing while citing articles for several specific research themes, such as medical applications, brain waves, pattern recognition, and steady state visual evoked potentials (SSVEP). Thus, we can understand that the BCI has been simultaneously approached from several different conventional research themes, and is integrating them. In this manner, we confirmed that the formation processes of research areas can be captured by closely observing the map.

6. Conclusion and future work

In this study, we developed a map of science, Mapping Science based on the research content similarity for funding project descriptions and recently published articles, which have difficulty in applying the citation analysis. After improving the existing paragraph embedding technique with an entropy-based clustering method of word vectors, we confirmed the good face validity. Then, we introduced the map constructed from approx. 300 k IEEE articles and NSF projects from 2012 to 2016 with the clustering and layout method of articles/projects and analytic functions provided on the map. Finally, we confirmed that formation processes of some specific research areas can be captured as changes of network structure.

As the next step, we plan to have a comparison with citation-based methods on concrete scenarios and incorporate patent information on the map. In addition, by overlaying domestic funding projects with NSF and Horizon2020 through the JST thesaurus that has English and Japanese notations, we will identify the trend of public grants. Finally, we try to extract metrics from chronological changes of the network structure of research areas. Foresight and understand from scientific exposition (FUSE) program in Intelligence advanced research projects activity (IAPRA) already conducted a study for identifying emerging research area based on several metrics obtained from several maps of science from 2011 to 2015. We, JST, will also utilize such metrics in statistical analysis and machine learning techniques to detect emerging research areas in their early stage for the next science and technology policies.

Author details

Takahiro Kawamura*, Katsutaro Watanabe, Naoya Matsumoto and Shusaku Egami

*Address all correspondence to: takahiro.kawamura@jst.go.jp

Japan Science and Technology Agency, Tokyo, Japan

References

[1] Price D. Networks of scientific articles. Science. 1965;**149**:510-515

[2] Borner K. Sci2: A tool of science of science research and practice. In: Tutorial of the 10th International Conference on Scientometrics and Informetrics (ISSI 2011); 2011

[3] Boyack K, Klavans R, Borner K. Mapping the backbone of science. Scientometrics. 2005; **64**(3):351-374

[4] Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer T, McNamara D, Dennis S, Kintsch W, editor. Latent Semantic Analysis: A Road to Meaning. Hillsdale, NJ: Laurence Erlbaum; 2007

[5] Blei D, Ng A, Jordan M. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;**3**:993-1022

[6] Griffiths T, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences. 2004;**101**(suppl. 1):5228-5235

[7] Talley E, Newman D, Mimno D, Herr B II, Wallach H, Burns G, Leenders A, McCallum A. Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods. 2011;**8**:443-444

[8] Herr II B, Talley E, Burns G, Newman D, LaRowe G. The NIH visual browser: An interactive visualization of biomedical research. In: Proceedings of 13th International Conference on Information Visualisation (ICIV 2009); 2009. pp. 505-509

[9] Kullback S, Leibler R. On information and sufficiency. Annals of Mathematical Statistics. 1951;**22**:79-86

[10] Firth JR. A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis. 1957;**1952-59**:1-32

[11] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at the International Conference on Learning Representations (ICLR 2013); 2013

[12] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 13). Vol. 2; 2013. pp. 3111-3119

[13] Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). 2014;**32**(2):1188-1196

[14] Shannon C. A mathematical theory of communication. Bell System Technical Journal. 1948;**27**(379-423):623-656

[15] Vilnis L, McCallum A. Word representations via Gaussian embedding. In: Proceedings of International Conference on Learning Representations (ICLR 2015); 2015. pp. 1-12

[16] Kimura T, Kawamura T, Watanabe K, Matsumoto N, Sato T, Kushida T, Matsumura K. J-GLOBAL knowledge: Japan's largest linked data for science and technology. In: Proceedings of the 14th International Semantic Web Conference (ISWC 2015); 2015

[17] Santus E, Lenci A, Lu Q, Walde S. Chasing hypernyms in vector spaces with entropy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014); 2014. pp. 38-42

[18] Kawamura T, Watanabe K, Matsumoto N, Egami S, Jibu M. Funding map for research project relationships using paragraph vectors. In: Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI 2017); 2017. pp. 1121-1131

[19] Kawamura T, Watanabe K, Matsumoto N, Egami S, Jibu M. Science graph for characterizing the recent scientific landscape using paragraph vectors. In: Proceedings of the 9th ACM International Conference on Knowledge Capture (K-Cap 2017); 2017. pp. 9-16

[20] Jones KS, Walker S, Robertson SE. A probabilistic model of information retrieval: Development and comparative experiments. Information Processing and Management. 2000;**36**(6):779-808

[21] Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America (PNAS 2008). 2008;**105**(4):1118-01123

[22] Newman MEJ. Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America (PNAS 2006). 2006; **103**(23):8577-8582

[23] Martin S, Brown WM, Klavans R, Boyack K. OpenOrd: An open-source toolbox for large graph layout. In: Proceedings of SPIE, Visualization and Data Analysis (VDA); 2011. p. 786806

[24] Adai AT, Date SV, Wieland S, Marcotte EM. LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks. Journal of Molecular Biology. 2004;**340**(1):179-190

[25] Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software—Practice and Experience. 1991;**21**(11):1129-1164

[26] Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information Processing Letters. 1989;**31**(1):7-15

Chapter 12

The Impact on Citation Analysis Based on Ontology and Linked Data

Ming Xiao, Zeshun Shi and Shanshan Wang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76377

Abstract

Research focus: The aim of this chapter is to introduce a new citation analysis and service framework based on the semantic web technologies (e.g., ontology and linked data). *Research methods:* This research project is based on a review of relevant literature and a series of experimental results based on ontology and linked data. *Motivation:* Traditional citation analysis methods and tools are overly dependent on citation databases, and traditional citation information service may ignore the semantics of knowledge resources and lack ability to store and query data in a machine-readable mode. *Findings:* The findings underline that the new citation analysis and service system framework based on ontology and linked data are feasible, which can integrate information requirements and knowledge services, and provide users with more personalized and comprehensive services.

Keywords: citation analysis, methods and techniques, ontology, linked data, citation knowledge service system

1. Introduction

Citation analysis is a bibliometric analysis technique which reveals the quantitative characteristics and laws of scholarly publications. It involves the use of mathematical and statistical methods to analyze citations within journals, papers, authors, and other references. Citation analysis has seen substantial theoretical and practical progress over several decades of development and has been widely applied to evaluate scientific knowledge, identify scientific models, and explore new frontiers which being explored by the scientific community. It is of great significance in regard to technological innovation and scientific decision-making. Traditional

citation analysis methods and tools are overly dependent on citation databases, which have the following drawbacks:

1. All citation acts are treated as equally important.
2. All kinds of statistical indicators are based on specific instances of citation, which are annotated only by the author.
3. Citation databases can only reveal whether there is a reference shared between different papers but fail to reflect any deeper relationships among semantic citations.

Motivations and behaviors related to citation have been analyzed by researchers from various angles. In 2014, content-based citation analysis method [1] has also been proposed. In this chapter, we propose a new citation analysis framework based on ontology and linked data; our goal is to enhance the efficacy of citation analysis via semantic web technology.

2. Related work

Berners-Lee, Hendler, and Lassila [2] published the article "The Semantic Web" in 2001, marking a brand new approach to semantic web research. The World Wide Web Consortium (W3C) later established a series of technical specifications that promoted the further development of the semantic web; specifications such as RDF, OWL, and SPARQL have allowed the application of the semantic web to many research fields and, further, have laid a foundation for knowledge representation, knowledge organization, and information retrieval on the Internet. Ontology is one of the backbones of the semantic web and was widely used to specify standard concept vocabulary for exchanging data between systems, offer suggestions of answering queries, publish reusable knowledge bases, and provide services to facilitate operations across heterogeneous systems and databases [3]. In 2006, Berners-Lee [4] first proposed the concept of "linked data", which has since become a wildly popular research topic in the computer science (CS) and library and information science (LIS) fields. Linked data builds associations between objects through the resource description framework (RDF) structure, ultimately revealing the relationships and implicitly shared knowledge between heterogeneous sets of data. After more than 10 years of development, linked data has seen numerous breakthroughs in both theoretical and technical aspects. To date, the linking open data project [5] has successfully transformed billions of web data points (e.g., Wikipedia, geographic data, government data) into the RDF triples of linked data, creating one massive data network.

In recent years, researchers have begun to introduce semantic web technology to citation analysis in effort to exploit ontology, linked data, and other technologies to improve the description of citation behaviors and motivations. The most representative example is the semantic publishing and referencing (SPAR) ontologies created by Shotton, Portwin, Klyne, and Miles [6]. Citation Typing Ontology (CiTO) is the ontology SPAR used to describe the relationship between citing papers and cited papers; it provides reference information such as background, method, citation type (e.g., journals, books, reports), peer review, and more. CiTO's

citation types include factual relationships and rhetorical relationships. The current version (CiTO 2.4.6) allows authors to describe their citation motivations as references, thus helping to reveal indirect and implicit relationships at work in scholarly literature. Ciancarini et al. [7] presented an experiment to investigate which are the main difficulties behind CiTO and how the humans understand and adopt CiTO. Iorio et al. [8] proposed a tool called CiTalO, which could automatically annotate the nature of citations with properties defined in CiTO through the semantic web and NLP techniques. By contrast, Recupero et al. [9] created SHELDON to extract citation RDF data from text using a machine reader, and CiTO was also used to describe the citation relationship.

Other researchers, for example, Ding, Konidena, Sun, and Chen [10], have also explored the idea of semantic citation to suggest that individuals can use ontology and linked data to describe bibliographic data and publish it to RDF triples. Mahmood, Qadir, and Afzal [11] combined semantic web technology with credible citation analysis to establish a framework that provides openness and reliability validation for all stages of the citation behavior lifecycle. The framework requires the use of semantic metadata at all stages of academic publishing to annotate the citation behavior and generate machine-readable RDF triples. This kind of annotation makes author, publisher, database vendor, and citation analysis system work together and build a set of reliable reference information while eliminating any false or misleading citation actions in the literature. More recently, Peroni et al. [12] experimentally described references in a suitable machine-readable RDF formats to make reference lists freely available to all academics. The open citation corpus [13] is created to store citation data from open access databases.

Quickly moving into an unfamiliar field for researchers is difficult, due to the mass of scientific articles [14] that must be reviewed without prior knowledge of their research contents. In a traditional citation information service, the search results are generated by keywords and other information that match specific knowledge resources and the corresponding user's correspondence. Such a method is simple, but it often ignores the semantic level of the knowledge resources, causing it to miss a significant number of semantic knowledge resources [15]. It may yield search results from a large number of studies that still do not meet the user's personalized knowledge needs [16].

In 2001, Aronson [17] argued that query refinement based on ontology is more efficient than other methods that were available at the time. From the perspective of information organization, ontology is a new method of knowledge organization and processing, and it is also the basis of semantic webs. It can systematize and organize a large amount of relevant information. When applying ontology to information retrieval, it is necessary to apply ontological principles to the information resources, so that search reasoning is implemented by the logical rules contained in the ontology itself, and a high quality retrieval result is output. With respect to the shortcomings of traditional citation information services, the introduction of ontology may help users to improve their searches aimed at multiple citation retrieval. In 2012, Kara, Alan, Sabuncu, Akpınar, Cicekli, and Alpaslan [18] found that while thesauruses are concerned with meanings at the level of words, ontologies more specifically deal with meanings at the level of real-world entities denoted by words. That is, ontologies deal with the interpretation of words in terms of real-world entities.

In recent years, with the advance of ontology, related studies have revealed that ontology-based knowledge services have been developed in different areas, including personalized medicine [19], e-government [20], medicine [21, 22], smart homes [23], the digital library [24, 25], and so on.

The digital library is an important application area of ontology-based knowledge service research. In 2015, Patkar [26] indicated that ontology is one of the latest tools for information retrieval from libraries in this digital age. His paper discusses advances in information managing tools and concludes by highlighting the applications of ontology among the different fields.

Koutsomitropoulos, Solomo, and Papatheodorou [27] studied the semantic search service of the DSpace digital repository system. They argued that Semantic Search v2 introduces a structured query mechanism that makes querying easier and improves the design of the system, performance, and scalability. Queries based on the DSpace ontology were dynamically created, and DSpace was able to obtain structured knowledge from the available metadata. Empirical and quantitative evaluation has shown that such a system can conduct semantic searches that provide better services for inexperienced users, such as the use of new query dimensions, with clear benefits.

In 2015, Iorio and Schaerf [28] proposed a semantic model defined by the Sapienza Digital Library to describe resource metadata. The semantic model is derived from the metadata object description model (a digital library descriptive standard). A top-level conceptual reference model supports the implementation of semantic web technologies for digital library metadata.

3. Method and process

Any citation analysis method based on ontology and linked data mainly includes the following three steps: first, building citation ontology according to the bibliographic citation data and full-text citation information; second, using the citation ontology to normalize the reference information and publish the data to linked data according to the RDF model; and, third, in order to extract the required citation information, writing a specific SPARQL search query for a citation analysis dimension and executing the search query. The search results are then visualized to reach the citation analysis goals.

3.1. Citation ontology construction

From the perspective of citation analysis, bibliographic citation information and full-text citation information are not only two independent parts but also two important sources of data that are both necessary for citation analysis. Here, we construct the bibliographic citation ontology (BCO) and full-text citation ontology (FCO) based on the bibliographic citation information and the full-text citation information, respectively. This allows us to achieve comprehensive semantic annotation of the citation information at hand.

The most commonly used ontology construction methods are the IDEF-5 [29], skeletal methodology [30], KACTUS [31], TOVE [32], METHONTOLOGY [33], and seven-step methodology [34]. The purpose of this study was to construct a task-based ontology to describe citation

information, so we choose the seven-step method developed by the Stanford University. The seven steps are (1) defining the domain and category of the ontology, (2) examining the possibility of reusing existing ontologies, (3) listing the important terms in the ontology, (4) defining the hierarchical system of classes, (5) defining the properties of the classes, (6) defining the facets of the properties, and (7) creating the instance. We also use the most popular protégé as our ontology development tool.

The construction of BCO is based on references. From the list of references, information such as the author, periodical, document type, year, volume period, and page number are extracted as the classes of BCO. In order to extend the dimensions of citation analysis, we extend the subclass from the perspective of journal and author. The "reference number" class is also added to the article, and the importance of the reference is measured by the quantities of internal references and external references. For property definitions, we reused the already-existing ontology properties (e.g., "fabio: hasPublicationYear," "bibo: volume") and marked the newly added attributes in the form of "bco." An example of the BCO ontology's classes and properties is shown in **Figure 1**.

The construction of FCO begins with three aspects: citation function, citation sentiment, and citation position. The citation function represents the role of cited work to citing work, such as background development, data support, methodology support, extension, or refutation. Citation sentiment expresses the emotion attitude from citing work to cited work, such as positive, neutral, and negative. Citation position indicates the location of the paragraph where the reference behavior occurs, such as the "Introduction" section of the document. An example of the FCO ontology's classes and properties is shown in **Figure 2**.

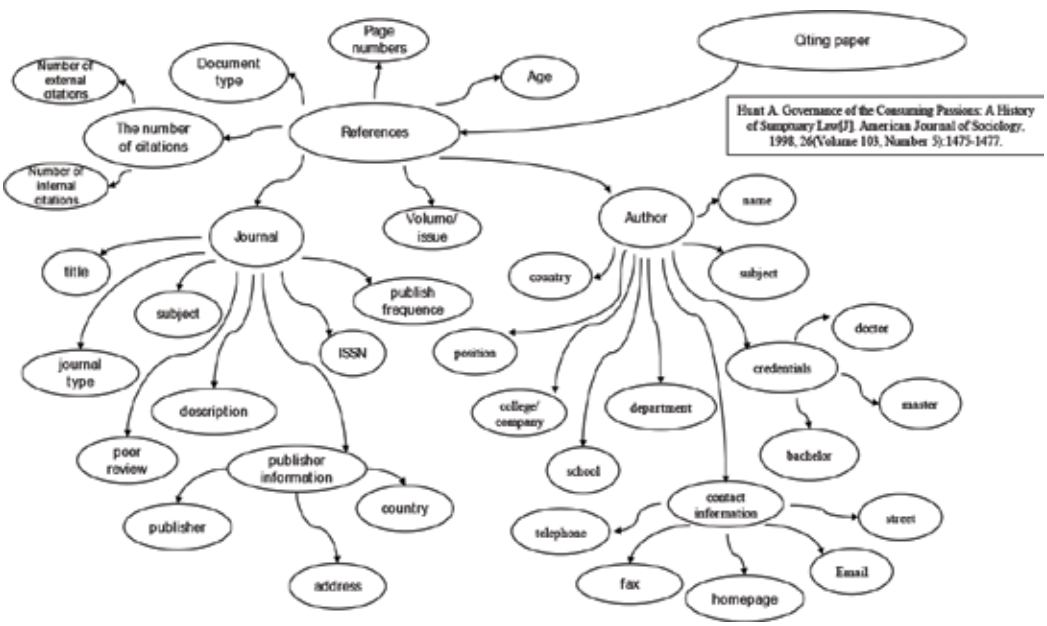


Figure 1. Example classes and properties of bibliographic citation ontology.

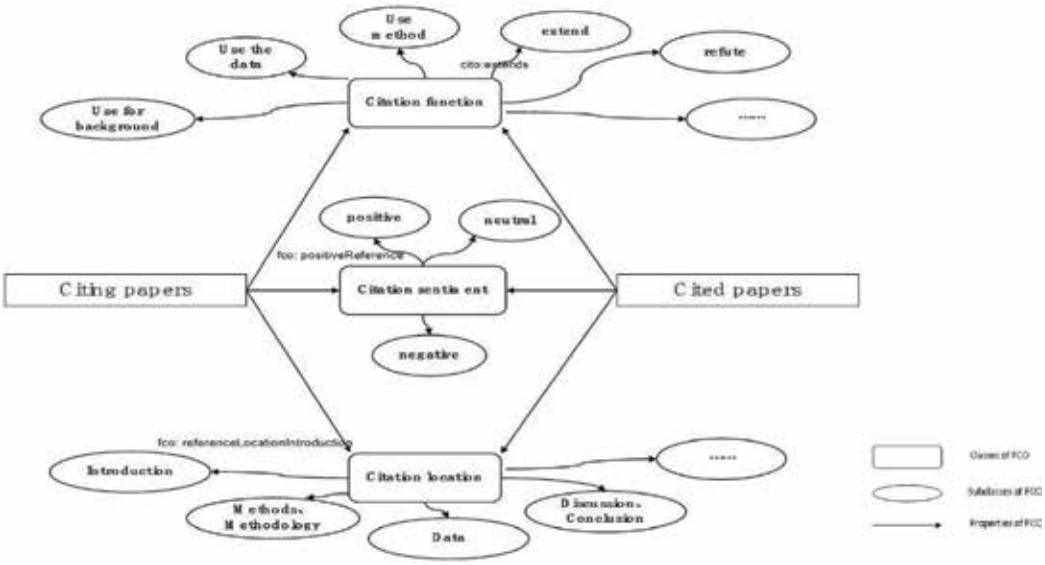


Figure 2. Classes and properties of full-text citation ontology.

3.2. Publishing linked data of the citation information

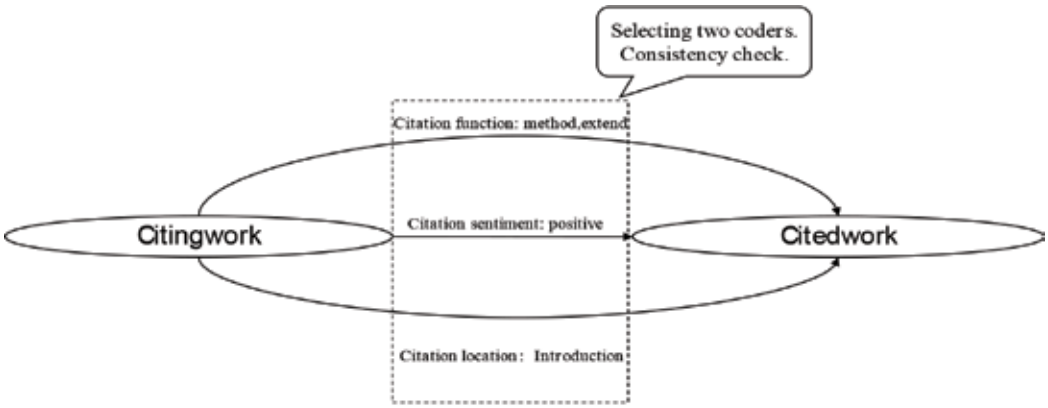
By using the citation ontology, we can publish the citation information linked data in the form of RDF triples. We used D2R as the linked data release software for this purpose. D2R is a very popular tool for linked data publication which serves to convert the massive, relational database format data into linked data RDF triples. We then imported the linked data into the semantic repository Virtuoso.

In terms of bibliographic citation data, we use the library, information science, and technology abstracts (LISTA) database as the data source. LISTA is a citation abstract database which contains the structured data of more than 600 core journals and 5000 core authors [35]. We have successfully published these data as linked data to form a strong foundation for subsequent citation analysis.

In the full-text citation information set, the most often-cited papers in the specific field were selected first as the citing work subset. The reference literatures were extracted as the cited work subset. On this basis, quoted sentences in the citing literature and cited literature were extracted, and the citation function, citation sentiment, and citation position information were marked by two trained coders. The full-text citation information was then organized into RDF triples as shown in **Figure 3**.

3.3. Citation analysis method implementation

The essence of the citation analysis method based on the linked data is to write the corresponding SPARQL query, which can be used to extract the citation information of specific dimensions. The search results are then calculated and visualized to analyze the citations




```
<http://example.com/citingwork>
   cito:cites <http://example.com/citedwork> ;
   cito:usesMethodIn <http://example.com/citedwork> ;
   cito:extends <http://example.com/citedwork> ;
   fco: referenceLocationIntroduction<http://example.com/citedwork> ;
   fco: positiveReference<http://example.com/citedwork> ;
   .....
```

Figure 3. Marking the full-text citation information.

per different dimension. In this chapter, we initially plan to implement 11 dimensions of citation analysis. Among them, citation quantity analysis, citation strength analysis, citation type analysis, citation language analysis, citation country analysis, citation age analysis, citation journal analysis, and co-citation analysis are based on the bibliographic data of traditional citation analysis, while the remaining three dimensions (citation function analysis, citation sentiment analysis, and citation position analysis) are based on a full-text citation analysis perspective.

The citation analysis process (for age and function, as examples) is shown in **Figure 7**. Citing literatures A and B constitute the citing subset, while references [1–7] serve as the cited paper subset. The relationship between them is complex and involves many factors. As mentioned above, the citation functions between them have been marked with "cito:extends," and the age information have been published as linked data. These citation relationships can thus be transformed into RDF triples as shown in **Figure 4**.

Once the triples are complete, we need to write a specific SPARQL search query to extract the specific citation information as shown in **Figure 5**.

The first SPARQL query is used to retrieve all the publication year information for the references cited by paper A, and the second query to retrieve all references to reference [4], which extends the function of document 4. The search results are then calculated and displayed as the final results. Visualization software (e.g., Power BI, Tableau) could also be applied to

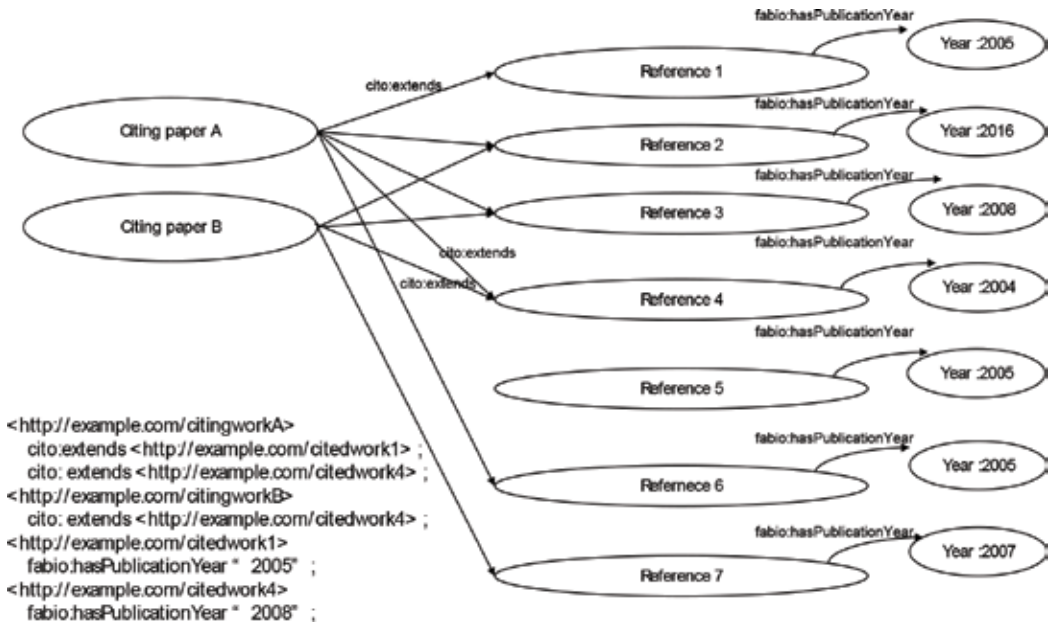


Figure 4. Example citation network for citation function analysis and citation age analysis.

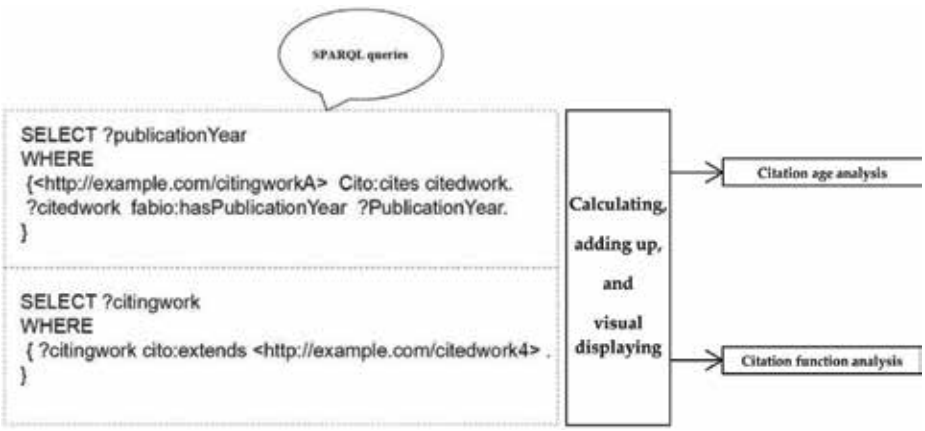


Figure 5. SPARQL queries and the corresponding citation analysis type.

simplify the display of results, and other dimensions of citation analysis can be implemented according to the same principle. As the quality of data is continually improved, more dimensions of citation analysis can also be achieved in follow-up experiments.

The citation knowledge service system based on ontology introduces ontology-related theory and technology into citation knowledge organization and knowledge retrieval and constructs an ontology-based citation knowledge service system. This system introduces a lightweight cube ontology to organize, store, and query citation knowledge data in a machine-readable

mode. It uses domain ontology to express the semantic representation of the citation knowledge base and to associate the citation knowledge data with the domain knowledge. According to user registration information and a user need survey, a user log flow provides users with targeted knowledge to ensure the effectiveness of the knowledge services.

4. Framework for a citation knowledge service system

In the process of creating a citation knowledge service, we construct the citation knowledge base, the lightweight citation ontology base, and the domain ontology base by using ontology and other technologies. We use the ontology to reorganize the citation knowledge unit, organize, store, and query citation data in a machine-readable mode. According to user's search habits that captured user behavior preferences and knowledge preferences, the system is able to understand user needs and establish a matching knowledge discovery mechanism [36].

This chapter presents a framework of an ontology-based citation knowledge service system, which contains four core layers, a data resource layer, an ontology layer, a semantic association layer, and a functional layer, as shown in **Figure 6**.

4.1. Data resource layer

The data resource layer is at the bottom of the knowledge base and contains the citation knowledge base and the user database.

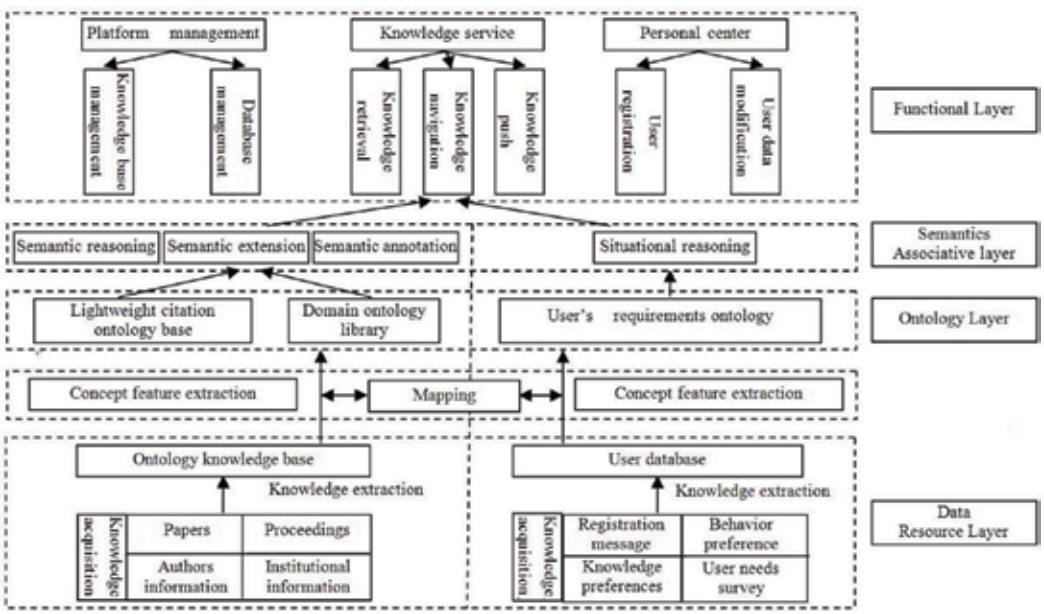


Figure 6. Citation knowledge service system framework.

The citation knowledge base provides data protection used for the construction of the domain ontology base, knowledge retrieval, knowledge recommendation, and other knowledge services. It stores the information about the citation resources gleaned during the knowledge acquisition. In addition to providing the user with the basis of the information, the citation knowledge base also contains other relevant information, such as the authors' personal profile, which allows it to reduce secondary retrieval.

The user database contains the registration information of the users. This system carries out a user demand investigation when the user registers. It can add user preferences and extract as a conceptual feature the input word phrase(s) of the user. It also performs ontological mapping.

4.2. Ontology layer

The ontology base contains the lightweight cube citation ontology base, the domain ontology base, and the user requirement ontology base. It simplifies the entity level, builds a convenient, simple citation ontology, organizes, stores, and queries data in a machine-readable mode. For example, the terms "dc: title," "fabio: hasTranslatedTitle," "bibo: pageStart," and "bibo: pageEnd," respectively, define the title of the journal, the English title, the start page, and other related attributes. These describe the citation in detail and realize the knowledge association of the citation information.

The domain ontology base contains the domain ontology, which includes class, property, and instance of domain ontology, as well as the ontological semantics of citation resources. Song and Zhang [37] agree that ontology can represent the complex semantic relations in the content of the information resources; it has a solid concept of hierarchical structure that supports logical reasoning. It is helpful for us to organize and retrieve information.

User requirements ontology conducts user need surveys for users and obtains user preferences directly. It analyzes users' search behavior, retrieves content, analyzes the users' preferences, obtains the user database, and builds the users' need ontology.

4.3. Semantic association layer

The semantic association layer will mainly analyze the content and related characteristics of the data, using Jena as the core processing tool, based on the pre-built domain ontology model. The information in the citation knowledge base is marked by Jena and uses Jena for reasoning. Finally, the SPARQL language is used to retrieve the information that has been marked. The semantic layer is based on the user requirement ontology and the user database and implements user requirements through scenario reasoning.

4.4. Functional layer

The functional layer provides the ontology-based citation knowledge service function, which currently includes a personal center module, a platform management module, and a knowledge service module.

The personal center provides new user registration and user data modification function. Platform management is the function of monitoring the entire knowledge service system that is used to operate the knowledge bases and databases. It mainly includes two modules: a knowledge base management module and a database management module. The knowledge service module includes the core functions of the knowledge service system, including ontology-based knowledge retrieval, knowledge navigation, and knowledge push modules.

4.5. The model for citation knowledge base

The functions of the proposed knowledge base include the following model: collecting literature from the citation database; extracting other relevant information from the authors' home page and organization page and other information carriers; introducing lightweight cube citation ontology to extract consensus citation elements; simplifying the entity level; organizing, storing, and querying data in a machine-readable mode to produce a list of concept features; establishing the relationship between the concept feature list and the domain ontology map; associating the citation knowledge data with the domain knowledge; performing the semantic processing of information after the semantic annotation, expansion, and synthesis; using ontology for formalization; mining implicit semantics through semantic reasoning; and forming a citation knowledge base ultimately. The model for citation knowledge base is shown in **Figure 7**.

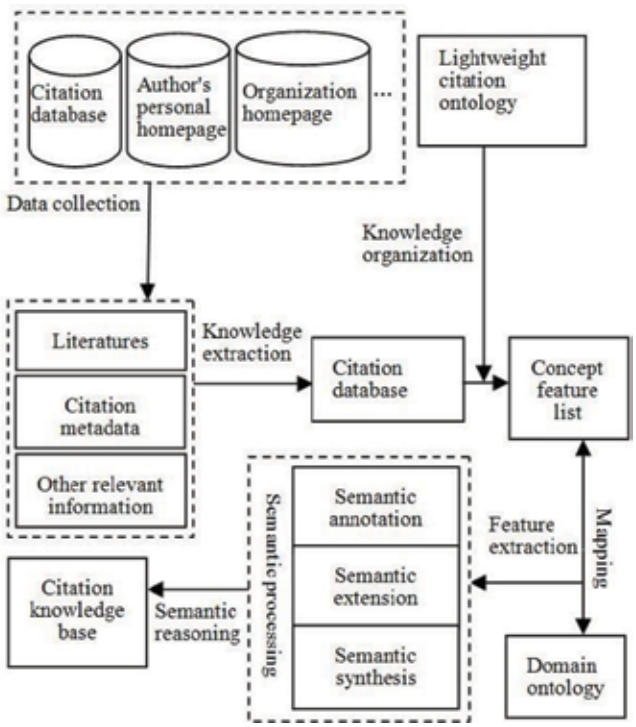


Figure 7. The model for citation knowledge base.

4.6. The model for user recommendation

This system uses the user registration information, the user need investigation, and the user log flow to obtain the user knowledge base. The user-related knowledge is analyzed, and the feature extraction is carried out. The users' requirement ontology is constructed and mapped with the domain ontology and users' characteristics. The mapping relationship is established, and the knowledge is organically related. This system uses the user knowledge base for knowledge extraction. The knowledge resources are classified and semantic associations are created. Entities are stored in the user requirement base, and ontology based on user recommendations is ultimately achieved, as shown in **Figure 8**.

In actual cases, data is usually collected in an ordered sequence. The distribution of the data is not static but changes over time. As certain factors change due to environmental factors, the regular pattern that the data has followed also changes; this is known as a concept change. The concept of "concept drift" [38] is that the rules that the data follows have changed throughout the sequence and the concept has drifted over time.

Because the users' actual operation is uncontrollable and does not follow any existing model, any new factors may have an impact on the users' operation, and concept drift in the users' data acquisition process is inevitable; therefore, the user model requires regular evolution [39].

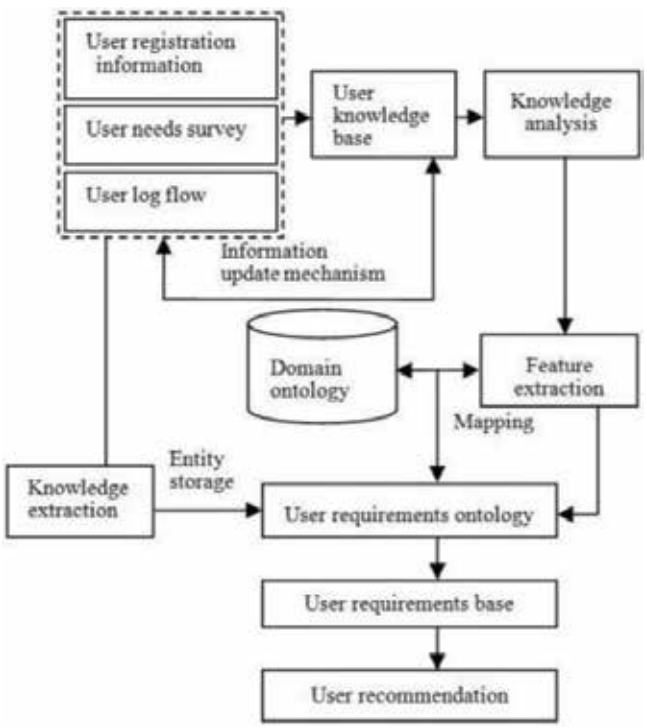


Figure 8. The model for user recommendation.

In order to reduce the effect of concept drift on the prediction effect, a triggering mechanism can be used to detect conceptual drift. Such change detection is based on statistics. It tracks the process of change in the user need concept set, removing the old data and re-adding the detected data to the users' requirement base to improve the prediction accuracy.

5. Conclusion

In this chapter, we proposed a new citation analysis framework based on ontology and linked data. By combining these technologies into a new semantic web with citation analysis method, we were able to improve the traditional citation analysis method (which relies heavily on citation databases). Rapid advancements in semantic publishing [40] and projects like the open citation corpus [41] have made it possible to mark massive amounts of citation information as machine-readable RDF triples. In the future, we plan to design further experiments to verify the feasibility of the proposed method. We hope that introducing ontology and linked data into citation analysis will yield optimal results while facilitating new technological developments and innovations.

An ontology-based citation knowledge service system uses ontology technology, knowledge navigation, knowledge recommendation, and other technologies and methods to organize, store, and query data in a machine-readable mode. It can successfully search knowledge across resource types and databases. Through the semantic relevance and knowledge navigation of various resources, we can render resources more granular, standardized, and automated. Using the methods of concept drift to track changes in users' needs achieves their information needs, and knowledge integration services provide users with more personalized and comprehensive services. This chapter constructs a framework of an ontology-based citation knowledge service system, aiming to provide new ideas for the development of knowledge services offered by traditional citation retrieval systems. We will focus on the realization of an ontology-based citation knowledge service system in the near future.

Acknowledgements

This work was supported by a grant from the national social science foundation of China (No. 16BTQ073).

Author details

Ming Xiao*, Zeshun Shi and Shanshan Wang

*Address all correspondence to: ming_xiao@bnu.edu.cn

School of Government, Beijing Normal University, Beijing, P.R. China

References

[1] Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C. Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology. 2014;**65**(9):1820-1833. DOI: 10.1002/asi.23256

[2] Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American. 2001;**284**(5):34-43. DOI: 10.1038/scientificamerican0501-34

[3] Gruber T. Ontology. In: Liu L, Özsu TM, editors. Encyclopedia of Database System. Boston, MA: Springer; 2009. pp. 1963-1965. DOI: 10.1007/978-1-4899-7993-3_1318-2

[4] Berners-Lee T. Linked Data [Internet]. 2006. Available from: https://www.w3.org/DesignIssues/LinkedData.html [Accessed: October 03, 2018]

[5] W3C. Linking Open Data [Internet]. 2007. Available from: https://xwww.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData [Accessed: October 03, 2018]

[6] Shotton D, Portwin K, Klyne G, Miles A. Adventures in semantic publishing: Exemplar semantic enhancements of a research article. PLOS Computational Biology. 2009; **5**(4):e1000361. DOI: 10.1371/journal.pcbi.1000361

[7] Ciancarini P, Iorio AD, Nuzzolese AG, Peroni S, Vitali F. Evaluating citation functions in CiTO: Cognitive issues. In: 11th Extended Semantic Web Conference (ESWC 2014), 25 May 2014 ; Crete; 2014. pp. 580e-594

[8] Iorio AD, Nuzzolese AG, Peroni S. Identifying functions of citations with CiTalO. In: The Semantic Web: ESWC 2013 Satellite Events (ESWC 2013), 26-30 May 2013; Montpellier; 2013. pp. 231-235

[9] Recupero DR, Nuzzolese AG, Consoli S, Presutti V, Mongiovì M, Peroni S. Extracting knowledge from text using SHELDON, a semantic holistic framework for linked ontology data. In: Proceedings of the 24th International Conference on World Wide Web, 19 May 2015; Florence; 2015. pp. 235-238

[10] Ding Y, Konidena D, Sun YY, Chen SS. 2009. Semantic Citation [Internet]. 2007. Available from: http://www.aswc2009.org/images/stories/documents/aswc2009_poster03.pdf [Accessed: October 03, 2017]

[11] Mahmood Q, Qadir MA, Afzal MT. Document similarity detection using semantic social network analysis on RDF citation graph. In: 2013 IEEE 9th International Conference on Emerging Technologies (ICET); 9-10 December 2013; Islamabad. New York: IEEE; 2013. pp. 1-6

[12] Peroni S, Shotton D, Vitali F. One year of the open citations corpus. In: 16th International Semantic Web Conference (ISWC 2017), 21-25 October 2017; Vienna, Cham: Springer; 2017. pp. 1e84-192

[13] Shotton D. Publishing: Open citations. Nature. 2013;**502**(7471):295-297. DOI: 10.1038/502295a

[14] Si Z, Dai G, Niu Z. Literature search framework by analyzing key aspects. In: International Conference on Advanced Communication Technology (ICACT 2016); South Korea. New York: IEEE; 2016. pp. 581-585

[15] Yang GY. Review of ontology–based knowledge retrieval research in China. Library Work and Study. 2015;**1**(6):18-21. DOI: 10.3969/j.issn.1005-6610.2015.06.004

[16] Băjenaru L, Smeureanu I. Learning style in ontology-based e-learning system. In: International Conference on Informatics in Economy (IE 2016). 2-3 June 2016; Cluj-Napoca, Cham: Springer; 2016. pp. 115-129

[17] Aronson AR. Effective mapping of biomedical text to the UMLS metathesaurus: The metamap program. Proceedings AMIA Symposium. 2001;**1**:17-21. PMID:11825149

[18] Kara S, Alan Ö, Sabuncu O, et al. An ontology-based retrieval system using semantic indexing. Information Systems. 2012;**37**(4):294-305. DOI: 10.1016/j.is.2011.09.004

[19] Chiang TC, Liang WH. A context-aware interactive health care system based on ontology and fuzzy inference. Journal of Medical Systems. 2015;**39**:1-25. DOI: 10.1007/s10916-015-0287-2

[20] Santos PM, Rover AJ. Knowledge representation through ontologies: An application in the electronic democracy field. Perspectivas em Ciência da Informação. 2016;**21**:22-49. DOI: 10.1590/1981-5344/2523

[21] Samwald M, Boyce RD, Freimuth RR. Pharmacogenomic knowledge representation, reasoning and genome-based clinical decision support based on OWL 2 DL ontologies. BMC Medical Informatics and Decision Making. 2015;**15**:1-10. DOI: 10.1186/s12911-015-0130-1

[22] Zhang YF, Gou L, Zhou TS, Lin DN, Zheng J, Li Y, et al. An ontology-based approach to patient follow-up assessment for continuous and personalized chronic disease management. Journal of Biomedical Informatics. 2017;**72**:45-59. DOI: 10.1016/j.jbi.2017.06.021

[23] Tao M, Zuo J, Liu Z, Castiglione A, Palmieri F. Multi-layer cloud architectural model and ontology-based security service framework for IoT-based smart homes. Future Generation Computer Systems. 2018;**78**:1040-1051. DOI: 10.1016/j.future.2016.11.011

[24] Yao Y. Library resource vertical search engine based on ontology. In: International Conference on Smart Grid and Electrical Automation (ICSGEA). May 27-28, 2017; Changsha. IEEE Computer Society; 2017. pp. 672-675

[25] Popa R, Vasilateanu A, Goga N, Doncescu A, Darminio P, Barbur RM, et al. Ontologies applied in medicine: A digital library creation framework for health literacy. In: The 6th IEEE International Conference on E-Health and Bioengineering (EHB 2017), 22-24 June 2017; Sinaia; 2017. pp. 137-140. DOI: 10.1109/EHB.2017.7995380

[26] Patkar V. A Passage to Ontology Tool for Information Organization in the Digital Age [Internet]. 2011. Available from: http://publications.drdo.gov.in/ojs/index.php/djlit/article/view/861 [Accessed: October 02, 2018]

[27] Koutsomitropoulos DA, Solomo GD, Papatheodorou TS. Semantic query answering in digital repositories: Semantic search v2 for Dspace. International Journal of Metadata Semantics & Ontologies. 2013;**8**(1):46-55. DOI: 10.1504/IJMSO.2013.054181

[28] Iorio AD, Schaerf M. A semantic model for content description in the sapienza digital library. In: Italian Research Conference on Digital Libraries(IRCDL); 1 January 2016. Cham: Springer; 2015. pp. 36-47

[29] Benjamin, PC, Menzel CP, Mayer RJ, Fillion F, Futrell MT, deWitte PS, Lingineni M. IDEF5 Method Report [Internet]. Knowledge Based Systems, Inc. 1994:1-175. Available from: https://www.scss.tcd.ie/Andrew.Butterfield/Teaching/CS4098/IDEF/Idef5.pdf [Accessed: October 03, 2018]

[30] Uschold M, King M. Towards a Methodology for Building Ontologies [Internet]. IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing. 1995;**98**(98):137-142. Available from: http://www.aiai.ed.ac.uk/publications/documents/1995/95-ont-ijcai95-ont-method.pdf [Accessed: October 03, 2018]

[31] Schreiber G, Wielinga B, Jansweijer W. The KACTUS View on the 'O' Word [Internet]. IJCAI Workshop on Basic Ontological Issues in Knowledge Sharing. 1995;**98**(98):159-168. Available from: http://www.math.vu.nl/~guus/papers/Schreiber95a.pdf [Accessed: October 03, 2018]

[32] Gruninger M, Fox MS. Methodology for the Design and Evaluation of Ontologies [Internet]. In: Workshop Notes of IJCAI-95, Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI1995). August, 1995; Montreal, Canada; 1995. pp. 6.1-6.10. Available from: http://stl.mie.utoronto.ca/publications/gruninger-ijcai95.pdf [Accessed: October 03, 2018]

[33] Fernández-López M, Gómez-Pérez A, Juristo N. METHONTOLOGY: From Ontological Art Towards Ontological Engineering [Internet]. In: AAAI-97 Spring Symposium Series. 24-26 March 1997; Stanford University, EEUU; 1997. Available from: http://oa.upm.es/5484/1/METHONTOLOGY_.pdf [Accessed: October 03, 2018]

[34] Noy NF, Mcguinness DL. Ontology development 101: A Guide to Creating Your First Ontology [Internet]. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880; 2001. Available from: https://protege.stanford.edu/publications/ontology_development/ontology101.pdf [Accessed: October 03, 2018]

[35] Kumar KA. A Scientometric Study of Digital Literacy in Online Library Information Science and Technology Abstracts (LISTA) [Internet]. 2014. Available from: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=2519&context=libphilprac [Accessed: October 03, 2018]

[36] Gullà F, Cavalieri L, Ceccacci S, Papetti A, Germani M. The user-product ontology: A new approach to define an ontological model to manage product searching based on user needs. In: International Conference on Human Interface and the Management of Information (HIMI 2017). 9-14 July 2017; Vancouver. Cham: Springer; 2017. pp. 333-346. DOI: 10.1007/978-3-319-58521-5_27

[37] Song S, Zhang X, Qin G. Multi-domain ontology mapping based on semantics. Cluster Computing. 2017;**20**(4):1-13. DOI: 10.1007/s10586-017-1087-x

[38] Wang S, Schlobach S, Klein M. Concept drift and how to identify it. Web Semantics Science, Services & Agents on the World Wide Web. 2011;**9**(3):247-265. DOI: 10.1016/j.websem.2011.05.003

[39] Fanizzi N, D'Amato C, & Esposito F. Conceptual clustering and its application to concept drift and novelty detection. In: Proceedings of the Semantic Web: Research and Applications, European Semantic Web Conference (ESWC 2008). 1-5 June 2008; Tenerife, Canary Islands. Spain: DBLP, 2008. pp. 318-332

[40] Shotton D. Semantic publishing: The coming revolution in scientific journal publishing. Learned Publishing. 2009;**22**(2):85-94. DOI: 10.1087/2009202

[41] I4OC. Initiative for Open Citations (I4OC) Launches with Early Success [Internet]. 2017. Available from: http://www.arl.org/news/arl-news/4256-initiative-for-open-citations-i4oc-launches-with-early-success [Accessed: October 03, 2018]

Chapter 13

Progress of Studies of Citations and PageRank

Wataru Souma and Mari Jibu

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77389

Abstract

A number of citations have been used to measure the value of paper. However, recently, Google's PageRank is also extensively applied to quantify the worth of papers. In this chapter, we summarize the recent progress of studies on citations and PageRank. We also show our latest investigations of the citation network consisting of 34,666,719 articles and 591,321,826 citations. We propose the generalized beta distribution of the second kind to explain the distribution of citation and introduce the stochastic model with aging effect and super preferential attachment. Furthermore, we clarify the positive linear relation between citations and Google's PageRank. By using this relationship as the benchmark to classify papers, we extract extremely prestigious papers, popular papers, and rising papers.

Keywords: citation, PageRank, SCI-E, fat tail, stochastic model, prestigious papers, popular papers, rising papers

1. Introduction

Citation analysis has a long history. Recently, Hou [1] applied the new method called the reference publication year spectroscopy (RPYS) to 2543 papers including 56,392 references regarding citation analysis in Science Citation Index Expand (SCI-E) and Social Science Citation Index (SSCI) data from 1970 to July 2016. This investigation clarified that the development of citation analysis is divided into five periods: before 1990, 1901–1950, 1951–1970, 1971–2000, and 2001–2016. In this chapter, we focused on the distribution of citations which were introduced by Price [2] and extensively investigated in the third period, that is, 1950s–1970s. In this chapter, we consider that the number of citations expresses the popularity of papers.

The fifth period, that is, 2001–2016, is characterized by a period of rapid expansion and diversified directions. In this period, many conceptions have been introduced, for example, scientific

evaluation indices, citation networks, information visualization, and citing behaviors. A variety of new impact measures has been proposed based on social network analysis in sociology and of network science originated from physics, mathematics, and information science. Bollen [3] summarized 39 impact measures and investigated the correlation between them by using the principal component analysis. Then, Bollen [3] indicated that the notion of scientific impact is a multidimensional construct that cannot be adequately measured by any single indicator, although some measures are more suitable than others.

In this chapter, we focus on the Google's PageRank which is first proposed by Brin and Page [4] to obtain the list of useful web pages for queries by users. Thus, if we define the usefulness of web page as the number of links cited by the other web pages, the search engine should propose the list of portal sites, that is, popular web pages. Hence, this list is useless for web users. To overcome this problem, based on the concept of vote, Brin and Page [4] defined the usefulness of web pages as the number of votes from the linking web pages. In the algorithm of Google's PageRank, the number of ballets is proportional to the usefulness of the web page, that is, the useful web page has many ballets. As a result, the useful web page collects votes from the useful web pages. Thus, the Google's PageRank expresses the prestige of web pages. We consider that this characteristic of Google's PageRank is valid for the case of citation network.

This chapter is organized as follows. In Section 2, we explain characteristics of dataset used in this chapter. The distribution of citation and the stochastic model of citation network are elucidated in Section 3. In Section 4, we introduce Google's PageRank and calculate it. We consider the correlation between citation and PageRank in Section 5. Section 6 is devoted to conclusions.

2. Data

In this chapter, we use Science Citation Index Expand (SCI-E) provided by Clarivate Analytics Co., Ltd. This dataset contains bibliographic information of scientific papers published from 1900 to the present. However, due to limited research budget of authors, we use the dataset from 1981 to 2015 in this chapter. This dataset contains 34,666,719 papers and 591,321,826 citations.

In this chapter, we denote the number of papers published in the year t as $n(t)$. **Figure 1** depicts the change of $n(t)$. In this figure, $n(t)$ almost monotonically increased from 1981 to 2013 and decreased after 2013. However, this behavior of $n(t)$ is fake. This is because the dataset was made at the beginning of 2016 and it partially contains papers published in 2014 and 2015. It takes a few years for all the papers to be included in SCI-E.

If we consider papers as nodes and regard citations from a citing paper to a cited paper as directed links, we can consider the dataset of citations as a directed network. We call such a network as the citation network. The citation network consists of many connected components. We denote the number of nodes contained in connected components as c and represent a frequency of c as $F(c)$. **Figure 2** depicts $F(c)$. We can find that there is the largest connected

http://dx.doi.org/10.5772/intechopen.77389

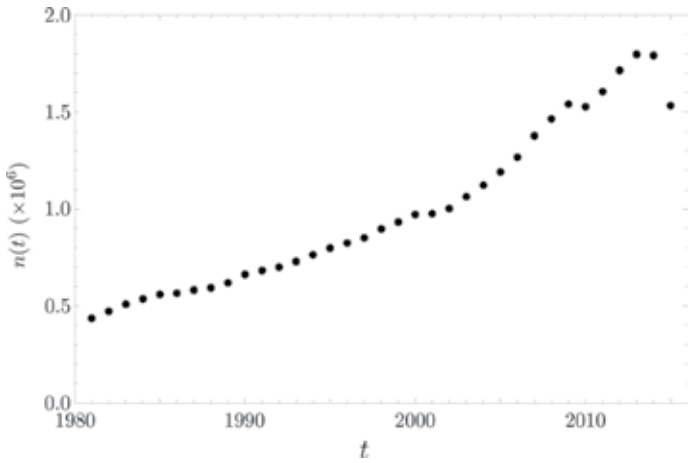


Figure 1. Yearly change of the number of e-articles.

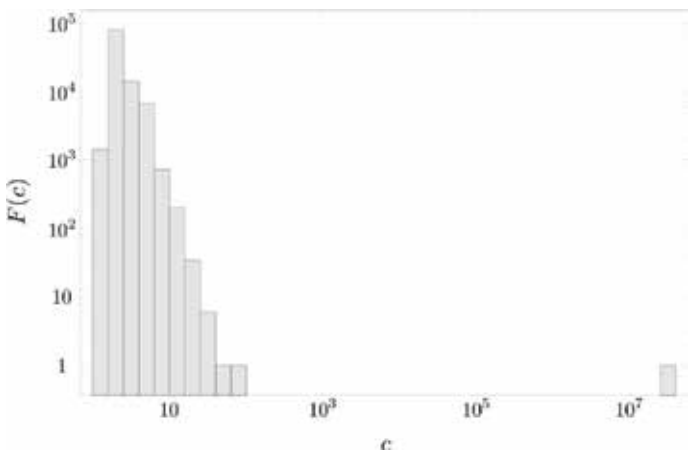


Figure 2. Distribution of the size of connected components.

component. This largest connected component consists of 34,428,322 nodes which are 99.3% of the total number of papers contained in the dataset, and of 591,177,607 links which are 99.98% of the total number of citations contained in the dataset. In the following section, we focus on the largest connected component.

3. Distribution and dynamics of citations

In this chapter, we argue for the distribution of the citations and stochastic models which lead to the citation network.

3.1. Distribution

The number of citations is represented by the number of in-degree, k, of the corresponding nodes. **Figure 3** is a double-logarithmic scale plot of the rank size distribution, $R(k)$, of citations. The right-tail part of the distribution decreases almost monotonically. This means that this part follows a power-law distribution, that is, $R(k) \propto k^{-\mu}$. Here, the exponent μ is called Pareto exponent originated in the name of Italian economist Vilfredo Pareto. The dashed line in **Figure 3** is the reference line which is the power law distribution with $\mu = 2$, that is, $R(\mu) \propto k^{-2}$.

Pareto [5] first investigated the fat-tail behavior of the right-tail part of personal income and wealth distributions. After Pareto, many types of distribution functions have been mainly proposed in the field of economics, especially in the investigation of personal income distribution (e.g., see [6, 7]). On the other hand, in the field of scientometrics, Price [2] first applied the power law distribution to the citation network and found that the distribution of the number of citing (the number of out-going degree in terms of network science) follows the power law distribution with $\mu = 1$ and that of the number of citations (the number of incoming degree in terms of network science) obeys the power law distribution with $\mu = 1.5$ or $\mu = 2$. The latter result is same as the reference line in **Figure 3**.

Rednar [8] investigated papers published in 1981 and cataloged by the Institute for Science Information (783,339 papers) and 20 years of publications in Physical Review D, vols. 11–50 (24,296 papers) and found that the right-tail part of both distributions of citation follows the power law distribution with $\mu = 2$. This result is same as Price [2] and the reference line in **Figure 3**. Rednar [9] investigated 110 years (from July 1893 through June 2003) of publications in Physical Review, the topical journals Physical Review A-E, Physical Review Letters, Review of Modern Physics, and Physical Review Special Topics: Accelerators and Beam (353,268 papers and 3,110,839 citations) and found that the entire distribution of the number of citation follows a log-normal distribution.

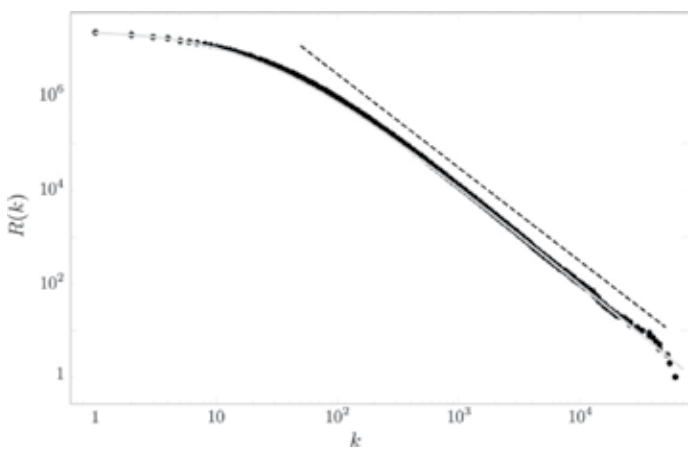


Figure 3. Rank size distribution, $R(k)$, of the number of citations, k.

Albarrán and Ruiz-Castillo [10] studied 5 years (1998–2002) of publications in Web of Science (3.7 million papers) and found that the power law distributions of the right-tail part of the distribution of citation are not rejected for 17 of the 22 scientific fields of Web of Science. Albarrán et al. [11] investigated same dataset of Albarrán and Ruiz-Castillo [10] and found that the power law distributions of the right-tail part of the distribution of citation are not rejected for 140 of the 219 scientific sub-fields of Web of Science. Recently, Brzezinski [12] investigated scientific papers published between 1998 and 2002 drawn from Scopus and found that the power law hypothesis is rejected for half of the Scopus field of science.

Although there are many researches besides the studies stated above, there are no studies that used vast amounts of data to approach the overall picture of citation distribution, like this chapter. The light gray line in **Figure 3** is the best fit by the generalized Beta distribution of the second kind (GB2) (or called the beta prime distribution) (e.g., see [13, 14]) with the probability density function:

$$f(k; a, b, \mu, \nu) = \frac{ak^{a\mu-1}}{b^{a\mu}B(\mu, \nu)}\left[1 + \left(\frac{k}{b}\right)^{a}\right]^{-(\mu+\nu)}, \quad (1)$$

with $a = 0.7$, $b = 15.2$, $\mu = 2.0$, $\nu = 3.0$. Here, $B_{(\mu, \nu)}$ is the Beta function.

Table 1 depicts the top 20 papers of citation. In this table, r_k is the rank of citation, k is the number of citations at the beginning of 2016, and k', which is enclosed in parentheses, is the number of citations at the beginning of January 2018. The characteristics of this list are that the subjects of papers are almost Biochemistry & Molecular Biology and that the publication years of papers are relatively old.

r_k	$k(k')$	First author	Title	Journal, Year	Subject
1	60,967 (62,404)	P. Chomczynski	Single-step method of RNA isolation by …	Analytical Biochemistry, 1987	Biochemistry & Molecular Biology; Chemistry
2	55,143 (65,452)	A.D. Becke	Density-functional thermochemistry. 3…	Journal of Chemical Physics, 1993	Chemistry; Physics
3	52,035 (61,637)	C.T. Leer	Development of the Colle-Salvetti correlation…	Physical Review B, 1988	Physics
4	45,349 (64,127)	G.M. Sheldrick	A short history of SHELX	Acta Crystallographica Section A, 2008	Chemistry; Crystallography
5	44,915 (64,682)	J.P. Perdew	Generalized gradient approximation…	Physical Review Letters, 1996	Physics
6	42,407 (46,286)	J.D. Thompson	Clustal-W – Improving the sensitivity of …	Nucleic Acids Research, 1994	Biochemistry & Molecular Biology

r_k	$k(k^*)$	First author	Title	Journal, Year	Subject
7	39,281 (44,765)	S.F. Altschul	Gapped BLAST and PSI-BLAST: a new…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
8	37,133 (48,832)	S.F Altschul	Basic local alignment search tool	Journal of Molecular Biology, 1990	Biochemistry & Molecular Biology
9	36,988 (56,581)	K.J. Livak	Analysis of relative gene expression data…	Methods, 2001	Biochemistry & Molecular Biology
10	32,657 (37,653)	N. Saitou	The neighbor-joining method—A new …	Molecular Biology and Evolution, 1987	Biochemistry & Molecular Biology; Evolutionary Biology; Genetics & Heredity
11	30,032 (33,046)	Z. Otwinowski,	Processing of X-ray diffraction data collected…	Macromolecular Crystallography, 1997	Biochemistry & Molecular Biology
12	29,615 (34,235)	A.D. Beckead	Density-functional exchange-energy …	Physical Review A, 1988	Physics
13	25,987 (29,094)	J.D. Thompson,	The CLUSTAL_X windows interface: flexible…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
14	25,880 (33,287)	R.M. Baron	The moderator mediator variable distinction…	Journal of Personality and Social Psychology, 1986	Psychology
15	25,696 (29,809)	J.M. Bland	Statistical methods for assessing agreement…	Lancet, 1986	General & Internal Medicine
16	25,340 (30,673)	T. Mosmann	Rapid colorimetric assay for cellular …	Journal of Immunological Methods, 1983	Biochemistry & Molecular Biology; Immunology
17	24,308 (28,923)	S. Iijima	Helical microtubules of graphitic carbon	Nature, 1991	Science & Technology - Other Topics
18	23,894 (34,400)	G. Kresse	Efficient iterative schemes for ab initio total-energy calculations using …	Physical Review B, 1996	Physics
19	23,294 (27,062)	J. Felsenstein	Confidence-limits on phylogenies – an approach using the bootstrap	Evolution, 1985	Environmental Sciences & Ecology; Evolutionary Biology; Genetics & Heredity
20	21,456 (21,529)	A.P. Feinberg	A technique for radiolabeling DNA restriction endonuclease fragments …	Analytical Biochemistry, 1983	Biochemistry & Molecular Biology; Chemistry

Table 1. Top 20 papers of citation.

3.2. Stochastic models

Simon [15] proposed the stochastic model, the so-called Simon's model, to elucidate the empirical distributions: distribution of words in prose samples by their frequency of occurrence, distributions of scientists by number of papers published, distributions of cities by population, distributions of income by size, and distributions of biological genera by number of species. Although assumptions of Simon's model are written in terms of word frequencies, we can express them in terms of network science as follows: assumption I—The probability that a node gets new link is proportional to the number of its degrees, that is, rich get richer or Matthew effect (e.g., see [16]), and assumption II—We add a new node with a constant probability γ. Simon's model elucidates the fact that the right-tail part of the distribution follows the power law distribution with $\mu = 1/(1-\gamma)$.

Price [17] generalized Simon's model, the so-called Price's model, to explain the growth of the citation networks. Barabáshi and Albert [18] introduced the stochastic model, the so-called BA model, based on two concepts: preferential attachment and growth, which corresponds to assumptions I and II of Simon's model, respectively. BA model is the case of $\gamma = 1/2$ of Simon's model and derives the power law distribution with $\mu = 2$. Jeong et al. [19] extended BA model to include an aging effect and a class of homogeneous connection kernels. Golosovsky and Solomon [20, 21] further extended to include an effect of initial attractivity.

Here, we use the model proposed by Jeong et al. [19] and check the aging effect and homogeneity of the growth of citation network. If we denote the number of degree of node i as k_i, the time evolution of k_i is obtained by

$$\frac{dk_i}{dt} = A_i(t)\, k_i^{\alpha}. \tag{2}$$

Here, $A_i(t)$ is an aging factor and $\alpha > 0$ is an unknown scaling exponent. Krapivsky et al. [22] have shown, for the case without the aging factor, for $\alpha = 1$ (linear preferential attachment) the model is just same as BA model and derives the power law distribution with $\mu = 2$. For $\alpha < 1$, the model derives the stretched exponential distribution, and for $\alpha > 1$ (super preferential attachment) a single node connects to nearly all other nodes, akin to gelation.

If we discretize the model and consider $\Delta t = 1$ year, Eq. (2) is written by

$$\Delta k_i = A_i k_i^{\alpha}, \tag{3}$$

We investigate the dynamics of growth for 44,932 papers published in 1985. The left panel of **Figure 4** depicts the double-logarithmic scale scatter plot of the number of citations, k_i ($i = 1, 2, \ldots, 44932$), as of 1988 and the change of the number of citations, Δk_i, from 1988 to 1999. If we divide k_i into bins with logarithmically equal separation, $\bar{k}$ and calculate the average value of Δk_i for each bin, $\bar{k}$, we obtain the red dots which are depicted in the right pane of **Figure 4**. By these manipulations, Eq. (3) is written by

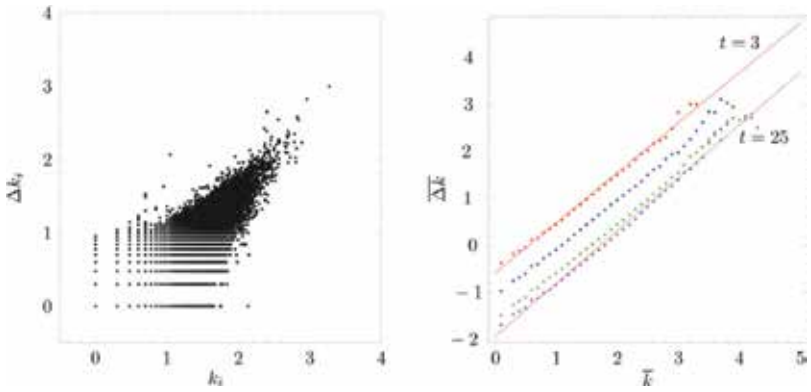


Figure 4. Left: Correlation between the number of citations and increase of the number of citations. Right: Change of the relation between mean citation and mean difference of citation.

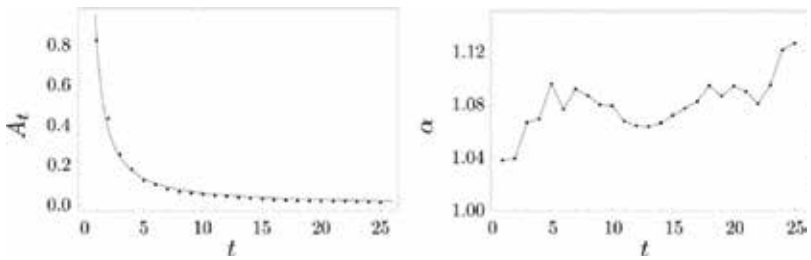


Figure 5. Left: Change of the aging effect. Right: Change of homogeneous factor.

$$\overline{\Delta k} = A_t \overline{k}^{\alpha}. \tag{4}$$

The red and solid line in the right panel of **Figure 4** corresponds to the linear regression of red dots by Eq. (4). The slope of this line corresponds to α and the intercept of it corresponds to A_t. In **Figure 4**, blue, green, and magenta dots are analysis for the year 1993, 2003, and 2010, respectively.

The left panel of **Figure 5** depicts the change of A_t. The solid line in this figure corresponds to the regression by the power law function given by $A_t \propto t^{-1.15}$. The right panel of **Figure 5** depicts the change of α. This figure shows that $\alpha > 1$ for the entire period in which we investigated. From this analysis, we realize that the citation network has the characteristics of super preferential attachment; therefore, it is expected that a single node connects to nearly all other nodes. However, the aging effect prevents the citation network from an oligopolistic network.

4. Distribution of PageRank

Google's PageRank is proposed by Brin and Page [4]. The Google number, G_i, of paper i is defined by the recursion formula (from Chen et al. [23]):

http://dx.doi.org/10.5772/intechopen.77389

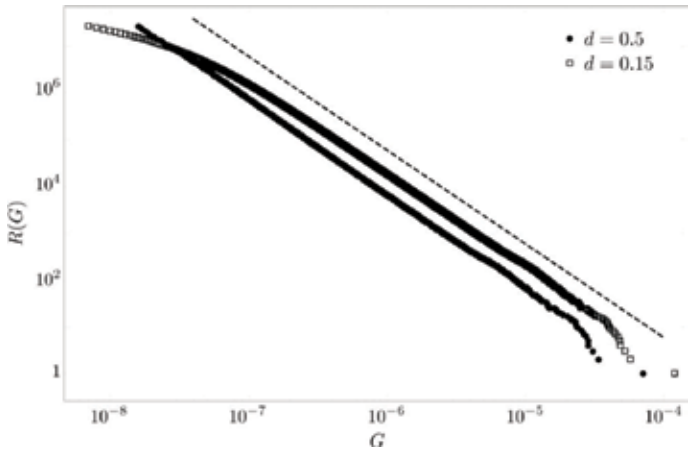


Figure 6. Rank size distribution, $R(G)$, of the Google number, G.

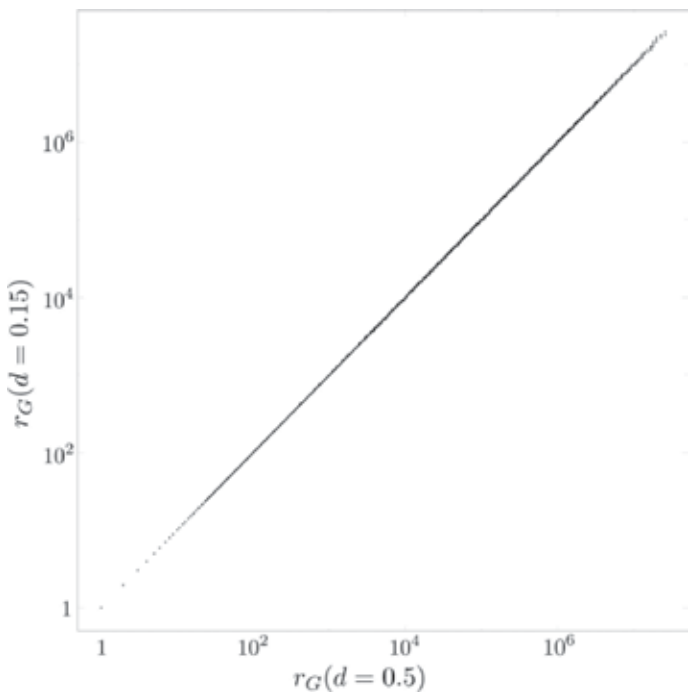


Figure 7. Correlation between the PageRank, r_G, in the case of $d = 0.5$ and $d = 0.15$.

$$G_i = (1-d)\sum_{i\,\mathrm{nn}\,j}\frac{G_j}{k_j}+\frac{d}{N}. \tag{5}$$

Here, $N = 34428322$ is the total number of articles contained in the largest connected component of the citation network. The sum is over the neighboring nodes j in which a link points to node i. In Eq. (5), d is a free parameter that controls the convergence and effectiveness

of the recursion calculation. In the original Google's PageRank [4], $d = 0.15$ is adopted and appropriate for the case of world wide web. On the other hand, $d = 0.5$ is adopted in [23] and appropriate for the case of citation network.

Figure 6 depicts the double-logarithmic scale plot of the rank size distribution of Google number, $R(G)$. In this figure, filled circles correspond to the case of $d = 0.5$ and open squares correspond to that of $d = 0.15$. The dashed line in this figure is the reference line and represents the power law distribution with $\mu = 2$. This value of exponent is same as the case of distribution of citation as depicted in **Figure 3**. Although the rank size distribution of Google number depends on d, the Google's PageRank, r_G, is almost the same as depicted in **Figure 7**. This figure is the double-logarithmic scale plot of r_G, and the abscissa is r_G in the case of $d = 0.5$, and the ordinate is r_G in the case of $d = 0.15$.

Table 2 depicts the top 20 lists of the Google's PageRank. The characteristics of this list are that papers belong to many subjects and that the publication years of papers are relatively old.

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
1	7.1314	4	45,349 (64,127)	4	G.M. Sheldrick	A short history of SHELX	Acta Crystallographica Section A, 2008	Chemistry; Crystallography
2	3.4074	1	60,967 (62,404)	0.5	P. Chomczynski	Single-step method of RNA isolation by acid…	Analytical Biochemistry, 1987	Biochemistry & Molecular Biology; Chemistry
3	3.1210	26	18,109 (18,789)	8.67	G.M. Sheldrick	Phase annealing in SHELX-90 – direct methods for…	Acta Crystallographica Section A, 1990	Chemistry; Crystallography
4	2.8852	2	55,143 (65,452)	0.5	A.D. Becke	Density-functional thermochemistry. 3…	Journal of Chemical Physics, 1993	Chemistry; Physics
5	2.8578	64	12,824 (14,640)	12.8	J. Kennedy	Particle swarm optimization	IEEE International Conference, 1995	Computer Science
6	2.7879	15	25,696 (29,809)	2.5	J.M. Bland	Statistical methods for assessing agreement…	Lancet, 1986	General & Internal Medicine
7	2.6547	3	52,035 (61,637)	0.43	C.T. Lee	Development of the Colle-Salvetti correlation…	Physical Review B, 1988	Physics
8	2.5745	76	11,685 (18,640)	9.5	D.G. Lowe	Distinctive image features from scale-invariant…	International Journal of computer Vision, 2004	Computer Science
9	2.4425	5	44,915 (64,682)	0.56	J.P Perdew	Generalized gradient approximation made…	Physical Review Letters, 1996	Physics

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
10	2.3890	46	14,128 (17,990)	4.6	S. Kirkpatrick	Optimization by simulated annealing	Science, 1983	Science & Technology - Other Topics
11	2.3430	11	30,032 (33,046)	1	Z. Otwinowski	Processing of X-ray diffraction data…	Macromolecular Crystallography, 1997	Biochemistry & Molecular Biology
12	2.3236	97	10,368 (11,590)	8.08	F.H. Allen	Table of bond lengths determined by X-RAY…	Journal of the Chemical Society-Perkin Transactions 2, 1987	Chemistry
13	2.2868	6	42,407 (46,286)	0.56	J.D. Thompson	Clustal-W - improving the sensitivity of…	Nucleic Acids Research, 1994	Biochemistry & Molecular Biology
14	2.1787	8	37,133 (48,832)	0.57	S.F. Altschul	Basic local alignment search tool	Journal of Molecular Biology, 1990	Biochemistry & Molecular Biology
15	2.1481	7	39,281 (44,765)	0.47	S.F. Altschul	Gapped BLAST and PSI-BLAST: a new…	Nucleic Acids Research, 1997	Biochemistry & Molecular Biology
16	2.0319	10	32,657 (37,653)	0.63	N. Saitou	The neighbor-joining method – a new method…	Molecular Biology and Evolution, 1987	Biochemistry & Molecular Biology; Evolutionary Biology; Genetics & Heredity
17	1.9081	17	24,308 (28,923)	1	S. Iijima	Helical microtubules of graphitic carbon	Nature, 1991	Science & Technology - Other Topics
18	1.8685	107	9775 (10,827)	5.94	H.D. Flack	On enantiomorph-polarity estimation	Acta Crystallographica Section A, 1983	Chemistry; Crystallography
19	1.8001	82	11,242 (12,850)	4.32	A.L. Spek	Single-crystal structure validation with the…	Journal of Applied Crystallography, 2003	Chemistry; Crystallography
20	1.7796	129	8818 (8849)	6 s.45	N. Walker	An empirical-method for correcting…	Acta Crystallographica Section A, 1983	Chemistry; Crystallography

Table 2. Top 20 papers of Google's PageRank.

5. Correlation between citation and PageRank

Bollen and Rodriquez [24] described that the Institute for Scientific Information (ISI) Impact factor (IF) which is defined as the mean number of citations a journal receives over a two-year

period is a metric of popularity and that the Google's PageRank is a metric of prestige. This concept is also proposed by Chen et al. [23] and Maslov and Redner [25] which investigated all publications in the Physical Review family of journals from 1893 to 2003 and found the linear relation between the Google number and the number of citations. Furthermore, [23, 25] found that some outliers from this linear relation, especially the papers of which the ranking of PageRank is remarkably high and that of citation is slightly high, are universally familiar to physicists [23, 25] called such papers scientific "gems." Ma et al. [26] applied the concept of [23–25] to the field of biochemistry and molecular biology from 2000 to 2005. Though these studies investigated the citation network of some selected scientific field, this chapter investigates the citation network consisting of all scientific fields.

Figure 8 depicts the double-logarithmic scale plot of the correlation between the number of citations, k, and the Google number, G. In this figure, the solid gray line represents the mean value $\langle G \rangle$ calculated for bins of k with logarithmically equal width. This figure shows that $\langle G \rangle$ versus k is smooth and increases linearly with k for $k \geq 500$. Thus, the Google number and citations are almost similar measures characterizing the importance of papers. This result means that prestige (Google number) is proportional to popularity (citations) in many cases.

However, there are outliers which have high prestige comparing to popularity. These papers are located above the solid gray line in **Figure 8** and are regarded as extremely prestigious papers. If we denote the citation rank as r_k and the Google's PageRank as r_G, these extremely prestigious papers are extracted by the order of Google's PageRank with the constraint given by the ratio r_k/r_G. **Table 3** depicts the top 20 extremely prestigious papers selected by using the constraint $r_k/r_G > 10$. The characteristic of this list is that the subjects of papers are almost information science.

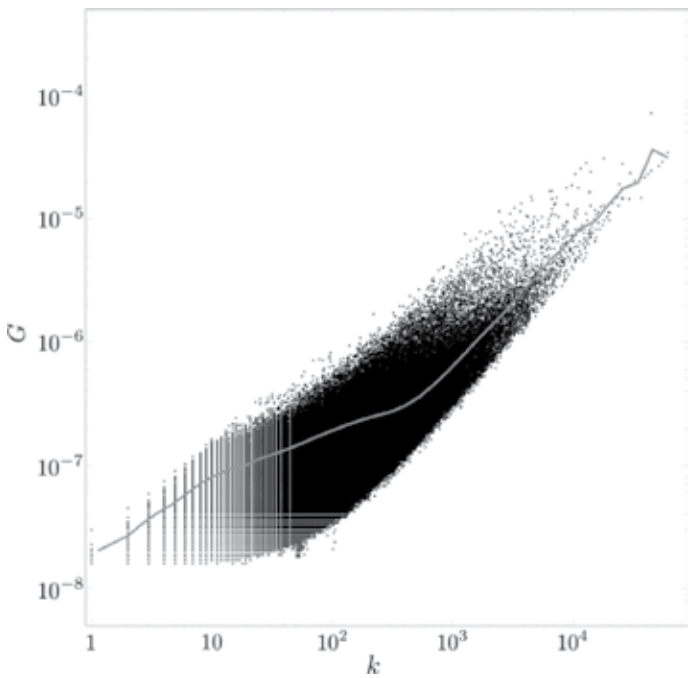


Figure 8. Correlation between the number of citations, k, and the Google number, G.

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
5	2.8578	64	12,824 (14,640)	12.8	J. Kennedy	Particle swarm optimization	Proceedings of IEEE International Conference, 1995	Computer Science
22	1.6861	240	6500 (7458)	10.91	S.M. Alamouti	A simple transmit diversity technique for wireless…	IEEE Journal on Selected Areas in Communications, 1998	Engineering; Telecommunications
25	1.5103	516	4465 (6605)	20.64	I.F. Akyildiz	Wireless sensor networks: a survey	Computer Networks, 2002	Computer Science; Engineering; Telecommunications
33	1.4160	481	4611 (6276)	14.58	Z. Pawlak	Rough sets	International Journal of Computer & Information Sciences, 1982	Information Science & Library Science
36	1.3169	784	3740 (5402)	21.78	I.F. Akyildiz	A survey on sensor networks	IEEE Communications Magazine, 2002	Engineering; Telecommunications
43	1.2155	998	3309 (4707)	23.21	T.R. Gruber	A translation approach to portable ontology…	Knowledge Acquisition, 1993	Computer Science; Information Science & Library Science
48	1.1432	828	3656 (4463)	17.25	P. Gupta	The capacity of wireless networks	IEEE Transactions on Information Theory, 2000	Computer Science; Engineering
49	1.1387	1916	2441 (2839)	39.10	S. Floyd	Random early detection gateways for congestion…	IEEE-ACM Transactions on Networking, 1993	Computer Science; Engineering; Telecommunications
53	1.1102	1247	2991 (3879)	23.53	G. Bianchi	Performance analysis of the IEEE 802.11 distributed…	IEEE Journal on Selected Areas in Communications, 2000	Engineering; Telecommunications
60	1.0626	608	4149 (5968)	10.13	S. Haykin	Cognitive radio: Brain-empowered wireless…	IEEE Journal on Selected Areas in Communications, 2005	Engineering; Telecommunications
76	0.9431	967	3360 (3961)	12.72	T. Murata	Petri nets - properties, analysis and applications	Proceedings of the IEEE, 1989	Engineering
79	0.9388	1758	2535 (3702)	22.25	W.B. Heinzel-man	An application-specific protocol architecture for…	IEEE Transactions on Wireless Communications, 2002	Engineering; Telecommunications
90	0.8884	1190	3048 (4075)	13.22	R. Ahlswede	Network information flow	IEEE Transactions on Information Theory, 2000	Computer Science; Engineering

r_G	$G(10^{-5})$	r_k	$k(k')$	r_k/r_G	First author	Title	Journal, Year	Subject
93	0.8767	1565	2691 (3401)	16.83	T. Wiegand	Overview of the H.264/AVC video coding standard	IEEE Transactions on Circuits and Systems for Video Technology, 2003	Engineering
97	0.8598	1045	3245 (4674)	10.77	M. Dorigo	Ant system: Optimization by a colony of…	IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 1996	Automation & Control Systems; Computer Science
116	0.7923	2736	2052 (2426)	23.59	D. HAREL	Statecharts - a visual formalism for…	Science of Computer Programming, 1987	Computer Science
120	0.7838	4059	1705 (2982)	33.83	M. WEISER	The Computer for the 21st-century	Scientific American, 1991	Science & Technology - Other Topics
121	0.7796	1406	2840 (4011)	11.62	S. Deerwester;	Indexing by latent semantic analysis	Journal of the American Society for Information Science, 1990	Computer Science; Information Science & Library Science
128	0.7584	3165	1914 (1948)	24.73	A.E. Leviton	Standards in herpetology and ichthyology…	Copeia, 1985	Zoology
129	0.7582	7409	1274 (1478)	57.43	X.Y. Wang	Room-temperature all-semiconducting…	Physical Review Letters, 2008	Physics

Table 3. Top 20 extremely prestigious papers.

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
125	8890 (17,192)	627	0.3250	5.02	D. Hanahan	Hallmarks of Cancer: The Next Generation	Cell, 2011	Biochemistry & Molecular Biology; Cell Biology
297	5817 (10,877)	1580	0.2042	5.32	D.W. Huang	Systematic and integrative analysis of large gene list…	Nature Protocols, 2008	Biochemistry & Molecular Biology
304	5747 (9681)	1608	0.2023	5.29	Y. Zhao	The M06 suite of density functionals for main…	Theoretical Chemistry Accounts, 2008	Chemistry
327	5533 (8874)	1810	0.1897	5.54	D.P. Bartel	MicroRNAs: Target Recognition and…	Cell, 2009	Biochemistry & Molecular Biology; Cell Biology

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
375	5147 (6894)	2128	0.1757	5.67	B.P. Lewis	Conserved seed pairing, often flanked by…	Cell, 2005	Biochemistry & Molecular Biology; Cell Biology
414	4912 (5825)	2506	0.1619	6.05	T. Jenuwein	Translating the histone code	Science 2001	Science & Technology - Other Topics
419	4895 (5350)	2123	0.1759	5.07	P. Li	Cytochrome c and dATP-dependent formation…	Cell, 1997	Biochemistry & Molecular Biology; Cell Biology
535	4382 (4604)	2802	0.1534	5.24	Z.G. XIA	Opposing effects of ERK and JNK-P38 map…	Science, 1995	Science & Technology - Other Topics
543	4343 (5864)	3120	0.1447	5.75	R.C. LEE	The C. elegans heterochronic geneG…	Cell, 1993	Biochemistry & Molecular Biology; Cell Biology
547	4327 (4633)	2865	0.1517	5.24	A. Hall	Rho GTPases and the actin cytoskeleton	Science, 1998	Science & Technology - Other Topics
600	4164 (5479)	3269	0.1411	5.45	S. Akira	Pathogen recognition and innate immunity	Cell, 2006	Biochemistry & Molecular Biology; Cell Biology
611	4144 (4888)	3585	0.1348	5.87	B.D. Strahl	The language of covalent histone modifications	Nature, 2000	Science & Technology - Other Topics
640	4063 (5604)	3359	0.1390	5.25	M.E. Raichle	A default mode of brain function	PNAS, 2001	Science & Technology - Other Topics
645	4054 (5303)	3326	0.1398	5.16	E.K. Miller	An integrative theory of prefrontal cortex function	Annual Review of Neuroscience, 2001	Neurosciences & Neurology
657	4026 (4967)	3572	0.1351	5.44	R.O. Hynes	Integrins: Bidirectional, allosteric signaling…	Cell, 2002	Biochemistry & Molecular Biology; Cell Biology
661	4005 (4335)	4096	0.1262	6.20	S.R. Datta	Akt phosphorylation of BAD couples survival…	Cell, 1997	Biochemistry & Molecular Biology; Cell Biology
706	3912 (5288)	4825	0.1166	6.83	T. Kouzarides	Chromatin modifications and their function	Cell, 2007	Biochemistry & Molecular Biology; Cell Biology

r_k	$k(k')$	r_G	$G(10^{-5})$	r_G/r_k	First author	Title	Journal, Year	Subject
751	3806 (5109)	4096	0.1262	5.45	M. Corbetta	Control of goal-directed and stimulus-driven…	Nature Reviews Neuroscience, 2002	Neurosciences & Neurology
752	3805 (4313)	4446	0.1213	5.91	A. Brunet	Akt promotes cell survival by phosphorylating and…	Cell, 1999	Biochemistry & Molecular Biology; Cell Biology
785	3739 (4363)	3972	0.1280	5.06	J.D. Fontenot	Foxp3 programs the development and…	Nature Immunology, 2003	Immunology

Table 4. Top 20 extremely popular papers.

On the other hand, there are also outliers which have low prestige comparing to popularity. These articles are located below the solid gray line in **Figure 8** and are regarded as extremely popular papers. These articles are extracted by the order of citation rank with the constraint given by the ratio r_G/r_k. **Table 4** depicts the top 20 extremely popular papers selected by using the constraint $r_G/r_k > 5$. These articles are divided into two groups. One group contains papers which are published in Nature, Science, and the Proceedings of the National Academy of Science of the United State of America (PNAS). Besides, publication year of these papers are approximately over 10 years ago. Furthermore, the growth rate of citations, k'/k, of those papers are low. The other group includes papers which are mainly published in Cell and are published relatively recently. What is more, the growth rate of citations, k'/k, of those papers are extremely high. Thus, we can regard these papers as rising papers.

6. Conclusions

We investigated papers published from 1981 to 2015 and contained in SCI-E. The total number of papers is 34,666,719 and that of citations is 591,321,826. We extracted the largest connected component from this dataset. The obtained citation network consists of 34,428,322 nodes (articles) and 591,177,607 links (citations).

The right-tail part of the rank size distribution of citations follows the power law distribution with exponent $\mu = 2$, that is, $R(k) \propto k^{-2}$. Furthermore, we introduced the generalized beta distribution of the second kind (GB2) as the best-fit function to the whole range of citation distribution. We introduced the stochastic model with growth, preferential attachment, and aging effect. Through the numerical analysis, we obtained the value of the parameter set.

Although the number of citations represent the popularity of papers, Google's PageRank reflects the prestige of papers. We evaluated Google's PageRank for the largest connected component which consists of 34,428,322 articles and 591,177,607 link citations. We found that the citations and Google numbers have a positive linear relation. We consider this positive

linear relation as a benchmark and selected extremely prestigious and extremely popular papers. We found that the subject of extremely prestigious papers is almost information science. Furthermore, we found that extremely popular papers are divided into popular papers and rising papers.

We conclude this chapter by describing two remaining issues. One concerns the stochastic model. Though we introduce GB2 as the best-fit function to the whole range of citation distribution, there is no stochastic model that explains GB2. The other concerns the weight of links in the citation network. Almost all studies have investigated citation networks as unweighted networks. However, it is possible to define weight of links, for example, similarity between papers.

Acknowledgements

This work is supported by Nihon University College of Science and Technology Grants-in Aid 2012 and 2016. The authors thank the Yukawa Institute of Theoretical Physics at Kyoto University. Discussions during the YITP workshop YITP-W-17-14 on "Econophysics 2017" were useful to complete this work.

Author details

Wataru Souma[1]* and Mari Jibu[2]

*Address all correspondence to: souma.wataru@nihon-u.ac.jp

1 College of Science and Technology, Nihon University, Japan

2 Japan Science and Technology Agency, Japan

References

[1] Hou J. Exploration into the evolution and historical roots of citation analysis by referenced publication year spectroscopy. Scientometrics. 2017;**110**:1437-1452. DOI: 10.1007/s11192-016-2206-9

[2] de Solla Price DJ. Networks of Scientific Papers. Science. 1965;**149**:510-515. DOI: 10.2307/1716232

[3] Bollen J, Van de Sompel H, Hagberg A, Chute R. A principal component analysis of 39 scientific impact measures. PLoS One. 2009;**4**:e6022. DOI: 10.1371/journal.pone.0006022

[4] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems. 1998;**30**:107-117. DOI: 10.1016/S0169-7552(98)00110-X

[5] Pareto V. Cours d'économie politique: professé a l'université de lausanne - tone second. Rouge: Lausanne F; 1897

[6] Arnold BC. Pareto Distributions. 2nd ed. US: CRC Press; 2015. p. 456. ISBN: 9781466584846

[7] Aoyama H, Fujiwara Y, Ikeda Y, Iyetomi H, Souma W, Yoshikawa H. Macro-Econophysics: New Studies on Economics Networks and Synchronization. UK: Cambridge University Press; 2017. pp. 53-96. ISBN: 9781107198951

[8] Rednar S. How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B. 1998;**4**:131-134. DOI: 10.1007/s100510050359

[9] Redner S. Citation statistics from 110 years of physical review. Physics Today. 2005;**58**: 49-54. DOI: 10.1063/1.1996475

[10] Albarrán P, Ruiz-Castillo J. References made and citations received by scientific articles. Journal of the American Society for Information Science and Technology. 2011;**62**:40-49. DOI: 10.1002/asi.21448

[11] Albarrán P, Crespo JA, Ortuño I, Ruiz-Castillo J. The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics. 2011;**88**:385-397. DOI: 10.1007/s11192-011-0407-9

[12] Brzezinski M. Power laws in citation distributions: Evidence from Scopus. Scientometrics. 2015;**103**:213-228. DOI: 10.1007/s11192-014-1524-z

[13] McDonald JB. Some generalized functions for the size distribution of income. Econometrica. 1984;**52**:647-663. DOI: 10.2307/1913469

[14] Kleiber C, Kotz S. Macro-Econophysics: Statistical Size Distributions in Economics and Actuarial Sciences. John Wiley and Sons; 2003. DOI: 10.1002/0471457175.ch2

[15] Simon HA. On a class of skew distribution functions. Biometrika. 1955;**42**:425-440. DOI: 10.2307/2333389

[16] Merton RK. The Matthew effect in science. Science. 1968;**159**:56-63. DOI: 10.1126/science.159.3810.56

[17] de Solla Price DJ. A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science. 1976;**27**:292-306. DOI: 10.1002/asi.4630270505

[18] Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;**286**: 509-512. DOI: 10.1126/science.286.5439.509

[19] Jeong H, Néda Z, Barabási AL. Measuring preferential attachment in evolving networks. Europhysics Letters. 2003;**61**:567-572. DOI: 10.1209/epl/i2003-00166-9

[20] Golosovsky M, Solomon S. Stochastic dynamical model of a growing citation network Based on a Self-Exciting Point Process. Physical Review Letters. 2012;**109**:098701. DOI: 10.1103/PhysRevLett.109.098701

[21] Golosovsky M, Solomon S. Growing complex network of citations of scientific papers: Modeling and measurements. Physical Review E. 2017;**95**:012324. DOI: 10.1103/PhysRevE.95.012324

[22] Krapivsky P, Redner S, Leyvraz F. Connectivity of Growing Random Networks. Physical Review Letters. 2000;**85**:4629-4632. DOI: 10.1103/PhysRevLett.85.4629

[23] Chen P, Xie H, Maslov S, Redner S. Google PageRank algorithm, scientific gems, physical review Citations. Journal of Informetrics. 2007;**1**:8-15. DOI: 10.1016/j.joi.2006.06.001

[24] Bollen J, Rodriquez MA, Van de Sompel H. Journal status. Scientometrics. 2006;**69**: 669-687. DOI: 10.1007/s11192-006-0176-z

[25] Maslov S, Redner S. Promise and pitfalls of extending Google's PageRank algorithm to citation networks. Society for Neuroscience. 2008;**28**:11103-11105. DOI: 10.1523/JNEUROSCI.0002-08.2008

[26] Ma N, Guan J, Zhao Y. Bringing PageRank to the citation analysis. Information Processing & Management. 2008;**44**:800-810. DOI: 10.1016/j.ipm.2007.06.006

Edited by Mari Jibu and Yoshiyuki Osabe

Technological change is one of the greatest issues in the modern world. As the world faces societal challenges, e.g., climate challenges, aging problem, and energy security, technology will contribute to new or better solutions for those problems. New technologies take time to develop and mature; moreover, they tend to be born in the gaps of multiple technology fields; therefore, early detection of emerging technological concepts across multiple disciplines will be a very important issue. Our goal seeks to develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technological emergence using one of the major foresight exercises, scientometrics. There is now a huge flood of scientific and technical information, especially scientific publications and patent information. Using the information patterns of emergence for technological concepts has been discovered and theories of technical emergence have been also developed in several years. We have been developing visualization tools in which thousands of technical areas have been interacted with each other and evolved in time. Several indicators of technical emergence have been improved by universities, international organizations, and funding agencies. This book intends to provide readers with a comprehensive overview of the current state of the art in scientometrics that focuses on the systematic, continuous, and comprehensive assessment of technological emergence.

Published in London, UK

IntechOpen

ISBN 978-1-83881-692-6