DSI

DATA SCIENCE AND INFORMETRICS

Volume 1
Number 1
February 2021

Editorial

Junping Qiu, Dangzhi Zhao, Fei Shu

Research Articles

DATA SCIENCE AND INFORMETRICS · VOLUME 1 · NUMBER 1 · 2021 FEB.

DATA SCIENCE AND INFORMETRICS

数 据 科 学 与 信 息 计 量 学

2021
Volume 1
Number 1

DATA SCIENCE AND INFORMETRICS (QUARTERLY)

DSI
数据科学与信息计量学
ISSN : 2694-6106 (Online)
ISSN : 2694-6114 (Print)
CEP : 941B0083

Address: No 1158, Blvd. 2, Xiasha Higher Education Zone, Hangzhou, Zhejiang, 310018, P.R. China.
Tel: 86-0571-86873846. Manuscript Submission Email: dsi@hdu.edu.cn.
Published by: Hangzhou Dianzi University, No 1158, Blvd. 2, Xiasha Higher Education Zone, Hangzhou, Zhejiang, 310018, P.R. China.
Edited by: Editorial Office of Data Science and Informetrics (DSI), No 1158, Blvd. 2, Xiasha Higher Education Zone, Hangzhou, Zhejiang, 310018, P.R. China.
Printed by: Hangzhou Yuhang People Printing Co., Ltd, No 68, Rd. Gouyun, Liangzhu Town, Hangzhou, Zhejiang, 311112, P.R. China.

DSI CALL FOR PAPERS

"Data Science and Informetrics" (DSI) is sponsored by Hangzhou Dianzi University (HDU), China Association of Science of Science & Technology Policy Research (CASSSP) and the Institute of Internet Industry, Tsinghua University. This journal is a high-level, international peer review journal in the field of data science and informetrics jointly undertaken by the Academy of Data Science and Informetrics, the Chinese Academy of Science and Education Evaluation and the School of Computer Science at HDU. It is also the association journal of the China Society for Scientometrics and Informetrics.

While based in China, the objective of DSI is to embrace an international vision of data science and informetrics, to innovate and develop the theories, methods and technologies of data and informetrics, and to explore data-driven interdisciplinary development research in data science and informetrics. Contributions to DSI can be submitted as research papers, technical reports, literature reviews, short communications and book reviews. We welcome all manuscripts from the following fields:

- Theoretical foundation and conceptual framework of data science and informetrics;
- Data mining, data analytics, machine learning and knowledge discovery, as well as intelligent processing of various data (including text, image, video, graph and network);
- Big data architecture, infrastructure, computing, matching, indexing, query processing, mapping, search, retrieval, interopera?bility, exchange, and recommendation;
- Data science applications in scientific, business, governmental, cultural, behavioral, social and economic, health and medical, human, natural and artificial (including online/Web, cloud, IoT, mobile and social media) domains;
- Ethics, quality, privacy, safety and security, trust, and risk in data science;
- The convergence of bibliometrics, scientometrics, webometrics, altmetrics, informetrics within the context of data science;
- Informetrics as a research method applied to research in other quantitative fields.

DSI adheres to academic standards, anonymous peer review, highlights the spirit of originality, aligns with its world-class academic counterparts, and strives to become a platform for experts and scholars in the field of data science and informetrics.

DSI is published 4 issues a year, and open to public solicitation and distribution at home and abroad. The ISSN is: ISSN 2694-6106 (Online) and ISSN 2694-6114 (Print). The CEP is: 941B0083.

Manuscript Submission: For now, authors should submit manuscripts in APA 7th style electronically to dsi@hdu.edu.cn. For more information regarding submission guidelines (e. g. manuscript format), please also contact: dsi@hdu.edu.cn. Once we have new submitting channel, we will immediately update it here.

DSI
Data Science and Informetrics

Volume 1
Number 1
February 2021

CONTENTS

Editorial

Research Articles

Editorial

Welcome to the very first issue of the journal Data Science and Informetrics (DSI). The interdisciplinary fields of data science and informetrics have been extremely active in recent years. DSI is created to publish research results in these fields and to promote innovative data-driven interdisciplinary research across these fields.

DSI adheres to international standards for scholarly journals with the goal to become an internationally recognized scholarly journal. We employ a double-blind peer review process, and emphasize originality and quality of research. DSI is truly open access and does not charge any Article Processing Charges (APCs) to authors.

DSI welcomes contributions from around the world reporting original research on theories, methods and technologies in data science and informetrics. This includes all metrics such as bibliometrics, scientometrics, webometrics, and altmetrics, as well as all types of data such as academic data, government data, business and industry data.

We cordially invite you to submit your work for publication in DSI. Your contributions to DSI will not only help DSI succeed, but also get feedback from our expert editorial board members and reviewers as well as become accessible widely and immediately upon publication. We would like to congratulate the authors of papers in this issue on being the first to publish in DSI.

We are grateful to all members of the DSI editorial board for accepting our invitation to join the board and to actively support DSI. We would also like to thank the many current and future expert reviewers for volunteering their time and expertise for the benefits of authors, DSI, and the community and science as a whole.

DSI is the official English journal of the Chinese Society for Scientometrics and Informetrics in collaboration with a number of organizations in China. These include the Chinese Association for Science of Science and S&T Policy, the Institute of Internet Industry at Tsinghua University, and the Institute of Data Science and Informetrics at Hangzhou Dianzi University.

Editors–in–Chief:

Prof. Junping Qiu
Institute of Data Science and Informetrics, Hangzhou Dianzi University, China

Prof. Dangzhi Zhao
School of Library and Information Studies,
University of Alberta, Canada

Prof. Fei Shu
Institute of Data Science and Informetrics,
Hangzhou Dianzi University, China

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

RESEARCH ARTICLES

Theoretical Data Science: bridging the gap between domain–general and domain–specific studies

Chaolemen Borjigin*, Chen Zhang, Zhizong Sun, Ni Yi

Key Laboratory of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China
School of Information Resource Management, Renmin University of China, Beijing, China

ABSTRACT

The entering into big data era gives rise to a novel discipline called Data Science. Data Science is interdisciplinary in its nature, and the existing relevant studies can be categorized into domain-independent studies and domain-dependent studies. The domain-dependent studies and domain-independent ones are evolving into Domain-general Data Science and Domain-specific Data Science. Domain-general Data Science emphasizes Data Science in a general sense, involving concepts, theories, methods, technologies, and tools. Domain-specific Data Science is a variant of Domain-general Data Science and varies from one domain to another. The most popular Domain-specific Data Science includes Data journalism, Industrial Data Science, Business Data Science, Health Data Science, Biological Data Science, Social Data Science, and Agile Data Science.

The difference between Domain-general Data Science and Domain-specific Data Science roots in their thinking paradigms: DGDS conforms to data-centered thinking, while DSDS is in line with knowledge-centered thinking. As a result, DGDS focuses on the theoretical studies, while DSDS is centered on applied ones. However, DSDS and DGDS possess complementary advantages. Theoretical Data Science (TDS) is a new branch of Data Science that employs mathematical models and abstractions of data objects and systems to rationalize, explain and predict big data phenomena. TDS will bridge the gap between DGDS and DSDS. TDS contrasts with DSDS, which uses casual analysis, as well as DGDS, which employs data-centered thinking to deal with big data problems in that it balances the usability and the interpretability of Data Science practices.

The main concerns of TDS are concentrated on integrating the data-centered thinking with the knowledge-centered thinking as well as transforming a correlation analysis into the casual analysis. Hence, TDS can bridge the gaps between DGDS and DSDS, and balance the usability and the interpretability of big data solutions.

The studies of TDS should be focused on the following research purpose: to develop theoretical studies of TDS, to take advantages of active property of big data, to embrace design of experiments, to enhance causality analysis, and to develop data products.

KEYWORDS

Data Science; Big Data; Theoretical Data Science; Domain-general Data Science; Domain-specific Data Science

* Corresponding author: chaolemen@ruc.edu.cn

1 Introduction

The bottlenecks in human being's data capabilities that capture or create, store, manage, compute, analyze as well as utilize data have been eliminated because of widespread applications of novel technologies. For instance, Internet of Things extends makes it possible for us to capture or digitalize the information of total populations instead of samples; Cloud Computing virtualizes computing resources and provides scalable on-demand services so that we can store, manage, compute, and analyze data at a low cost; Mobile Computing records the thoughts, feelings, and behaviors of the individuals as well as the social networks between them. As a result, we are entering into a data enriched-offerings era that is distinct from any previous era in human history.

Big data is shifting today's scientific paradigm, and giving rise to a novel discipline called Data Science. How to take advantages of big data in order to survive in data enriched-offerings era is one of the hot topics for most disciplines from basic science such as Statistics and Computer Science to applied sciences, including Social Sciences. As a result, the research on big data from different disciplines begins to converge on an emerging discipline called Data Science. Data Science deems data-centered thinking as an alternative paradigm for data-related tasks, which is different from the knowledge-centered thinking in traditional research. However, the studies of Data Science are spread across a variety of disciplines, and we need to conduct the in-depth research on its core theories, main methods, typical techniques, and best practices.

The rest of this paper is structured as follows: Section 2 discusses the brief history, interdisciplinarity, and taxonomy of Data Science, and categorizes the existing studies into two basic subgroups: Domain-general DS and Domain-specific DS. Then, Section 3 proposes the current research topics of Domain-general Data Science, including data wrangling, data computing, data management, data analysis, and data product development. In addition, the states of arts of typical Domain-specific Data Science are described in Section 4: Data Journalism, Industrial Data Science, Business Data Science, Health Data Science, Biological Data Science, Social Data Science, and Agile Data Science. Furthermore, Section 5 provides a comparative study between Domain-general Data Science and Domain-specific Data Science, and a comprehensive solution for integrating those two distinct branches of Data Science theories by introducing Theoretical Data Science. Finally, in Section 6, the critical topics for Theoretical Data Science studies are proposed.

2 Data Science: The Science of Big Data

Data Science is a new emerging discipline that is termed to address challenges that we are facing and going to face in data-enriched offerings era. It provides new theories, methods, models, technologies, platforms, tools, applications, and best practices of big data. And one of the main purposes of Data Science research is to reveal the new challenges and opportunities brought by big data.

2.1 A Brief History of Data Science

Peter Naur, the Turing Award winner, coined the term of Data Science in his book entitled *Concise Survey of Computer Methods* in 1974. He defined Data Science as the science of dealing with data, and further proposed that it is different from Datalogy, which is the sci-

ence of data and of data processes and its place in education (Naur, 1974). In 2001, William S. Cleveland published the paper, *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*, proposing that Data Science is an emerging branch of Statistics. In 2013, Nature published the article, *Computing: A Vision for Data Science* (Mattmann, 2013), and Communications of the ACM published the paper, *Data Science and Prediction* (Dhar, 2013). Both of those two articles discussed the Data Science from the perspective of Computer Science. Then, Data Science was also identified as a branch of Computer Science. Data Science has begun to get much more public attention since 2010s. Patil DJ and Davenport T H published the article entitled *Data Scientist: The Sexiest Job of the 21st Century* in Harvard Business Review in 2012. Barack Obama won the presidency by implementing and using big data strategies in the 2012 US presidential election (Kitchin, 2013). The White House announced Patil DJ the first U.S. Chief Data Scientist in 2015.

Data Science was on an one-way trip to the peak of inflated expectations and would enter plateau of productivity in 2 - 5 years according to Gartner's 2014 Hype Cycle for Emerging Technologies. Gartner's 2016 Hype Cycle for Data Science is a growth curve that shows the breadth and depth of excitement about Data Science, with new technologies and some significant movements from last year. Gartner's 2016 Hype Cycle for Data Science shows: R entered the plateau of productivity; Simulation, Ensemble Learning, Video or Image Analytics and Text Analysis were climbing the slope of enlightenment; Hadoop-Based Data Discovery was obsolete before plateau; Speech Analytics, Model Management and Natural-Language Question Answering have passed the peak of inflated expectations and slided into the trough of disillusionment; Citizen Data Science, Model Factory, Algorithm Marketplaces and Prescriptive Analytics have recently come to light.

2.2 The interdisciplinarity of Data Science

In 2010, Drew Conway proposed his Data Science Venn Diagram (Figure 1) to reveal the interdisciplinary of Data Science. The Venn Diagram shows that Data Science is a combination of hacking skills, math & statistics knowledge, and substantive expertise. Now, there are many variations of his Venn diagram such as Jerry Overton's Data Science Venn Diagram (2016), but all of them are less influential than Drew Conway's Venn Diagram.

Data Science is interdisciplinary, it has three basic components: knowledge, expertise, and skills. The knowledge in Data Science is domain-independent, while the expertise and the skills are domain-dependent. The knowledge in Data Science is evolving into Domain-general Data Science and the expertise and the skills from application fields is the source of Domain-specific Data Science.

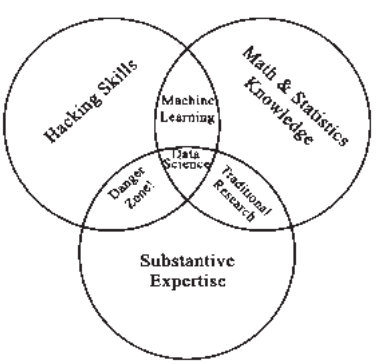


Figure 1 Venn Diagram by Drew Conway (2010)

2.3 Taxonomy of Data Science

Data Science is an emerging discipline that incorporates theories with domain-independent knowledge and domain-dependent business practices and skills. As a result, there are two types of Data Science: Domain-general Data Science and Domain-specific Data Science.

Domain-general Data Science regards Data Science as an independent discipline, while the Domain-specific Data Science argues that Data Science heavily depends on a specific application domain. The main research topics on Domain-specific Data Science include Data Journalism, Materials Data Science, Big Data Finance, Big Data Society, Big Data Ethics, and Big Data Education.

Domain-general Data Science is a theoretical foundation for Domain-specific Data Science. Domain-general Data Science involves the general ideas, theories, methods, concepts and tools of Data Science, Domain-specific Data Science is commonly restricted within a specific application discipline.

3 Domain–general Data Science

Domain-general Data Science is devoted to issues on domain-independent Data Science, involving concepts, theories, methods, technologies, tools and so on. The counterpart is Domain-specific Data Science, which are two different terms. Domain-general Data Science aims to solve theoretical challenges related to Data Science itself, including the core theories of Data Science and Data Wrangling, Data Computing, Data Management, Data Analysis and Data Products Development. It is worth noting that the basic theories are within the research scope of Data Science, while the theoretical bases are outside the scope.

3.1 Core theories

Core theories of DGDS involve new concepts, theories, methods, technologies, and tools applied in Data Science.

Big data and Data Science. Data Science is a science about big data, which covers a whole set of knowledge system of it. Big data is also one of the research objects of Data Science. It is broken by IBM (2013)data scientists into four dimensions: volume, variety, velocity and veracity. Big data analytics is the application of advanced analytic techniques to very big data sets (Russom, 2011). Many organizations have invested in developing products using Big Data Analytics to address the monitoring, experimentation, data analysis, simulations, large and disperse datasets. Data-Driven: Data-Driven refers to the process of doing things based on big data rather than purely on experience or intuition, such as Data-Driven Decision Making and Data-Driven Modelling. Jim Gray (2007)proposed that the scientific paradigm is shifting from Experimental Science, Theoretical Science and Computational Science to the fourth scientific paradigm—Data-Intensive Science. More efficient data-intensive techniques are required, such as cloud computing, social computing, and biological computing. Datafication is the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis(Mayer-Schönberger & Cukier, 2013). Along with the Internet of Things and Sensors, Quantified Self is also a hot topic of datafication.

Life Cycle of Data Science. Theories of Data Science involves potential guidance for the process of Data Wrangling, Data Computing, Data Management, Data Analysis, Data Products Development. In general, models are implementations of theory of Data Science (Das, 2017). But Data Science has grown up rapidly in the model applications of the big data

era. It has achieved successful practice in many fields, but its theoretical research is still lagging far behind its practical application. Theory-guided Data Science was proposed as an emerging paradigm for scientific discovery from data to the effectiveness improvement of Data Science models (Karpatne et al., 2017).

Methodologies of Data Science. Method guides the direction of problem solving and can promote the development of technology. Technology is used to solve the problem by performing an operation. Data Science is an interdisciplinary field that requires methods related to statistics, machine learning, deep learning, data analysis, data visualization, data processing, cloud computing, data engineering and so on. But Data Science is not to cover all aspects of these fields, but to unify some of their methods to gain insights from data. Technologies of Data Science are the specific executions of these methods.

Technologies of Data Science. Data scientists need to master the popular technologies include: Linear Regression, Resampling Methods, Nonlinear Models, Variance Analysis and Time Series Analysis, Decision Tree, Support Vector Machines, Random Forest, Principal Component Analysis, Classification, Clustering, Semi-supervised Learning and Reinforcement Learning, Deep Learning, Data Collection, Data Storage, Data Preprocessing, Data Mining, Natural Language Processing, and computer vision.

Tools of Data Science. Data Science tools that most of the data scientists used include: open-source Data Science programming languages, such as Python, R and Julia; big data computing tools, especially Hadoop MapReduce and Spark; big data storage tools, including HDFS and GFS; big data management systems, such as NoSQL, new SQL, and cloud RDB; Data mining tools, such as RapidMiner, Data Melt; Data visualization tools, including Tableau, D3.js and PowerBI; Machine learning or Deep Learning tools such as TensorFlow, PyTorch and Keras.

3.2 Data Wrangling

Data Wrangling is the initial process of transforming raw data into another form for improving data quality (Kandel et al., 2011). It is the first phase of most data-driven projects and known as data wrangling, data munging or janitorial work (Endel & Piringer, 2015). It also concerns that how to apply data scientists' skills of creative design, critical thinking, and curious questioning to the data wrangling activities and how to avoid Garbage in and Garbage out (GIGO).

Data preparation accounts for about 80% of the time of a Data Science project (Patil, 2012). Data wrangling is the preparation for analysis, which involves the following operations: data auditing, data cleaning, data conversion, data integration, data masking, data reduction, data annotation.

Data Auditing. Data auditing uses pre-determined evaluation methods to check data quality and identify its problems (Abdallah et al., 2017), the problems include: missing data, abnormal data, data contradicting each other, tampered data that cannot be traced back to its source. Taleb et al. (2018) proposed an across-the-board quality management framework describing the key quality evaluation practices to be conducted through the different Big Data stages.

Data Cleaning. Data cleaning is the process of altering messy data to tidy data by filtering duplicate data, identifying incorrect data, and processing missing values. However, big data processing is different from the small-scale data preprocessing in that the former has high robustness with low data quality. Big data cleaning aims to promote the quality of data form, whether the data is Tidy Data. Hadley Wickham (2014) put forward the concepts of Tidy Data

as well as Data Tidying, and proposed that Tidy Data should follow three basic principles: each variable must have its own column; each observation must have its own row; each value must have its own cell. In most cases, completely tidy data cannot be obtained after once data cleaning operation. Therefore, these middle data that may contain messy data need to be audited again and then cleaning is continuing. Müller and Freytag (2005) pointed that data cleaning is an iterative and normally never finished process that consists of the four consecutive steps: data auditing, workflow specification, workflow execution and post-processing and controlling.

Data Conversion. When the form of the original data does not meet the requirements of the analysis algorithm, it is necessary to perform Data Conversion on the original data from one data format or one big data environment to another (Gao et al., 2016). The usual techniques for data conversion include: smoothing data by binning regression and clustering to remove noise from the data (Han et al., 2011), constructing new features based on other features, and performing data standardization such as Min-max normalization and zero-mean normalization.

Data Integration. Data integration is the practice of combining data from different sources, and then providing the user with a unified view of these data (Lenzerini, 2002). Representative tools and techniques for data integration include Manual Integration, Common User Interface, Integration by Applications, Integration by Middleware, Uniform Data Access, and Common Data Storage (Amghar et al., 2019). But there are many challenges for the process of data integration, such as the entity identification problem, some attributes of the data set are redundant, different data sources have different measurement scales.

Data Masking. The main purpose of data masking is to protect sensitive data (Kuacharoen , 2014), such as someone's address or phone number. Specifically, data masking is the process of transforming the individual (or organization) sensitive data to reduce the sensitivity of information on the premise of not affecting the accuracy of data analysis. Typical techniques for data masking include: substitution, shuffling, number and date variance, data encryption, deleting sensitive data and replacing with NULL values (Sarada et al., 2015; Mansfield-Devine, 2014).

Data Reduction. Miles and Huberman (1994) explained that Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that "final" conclusions can be drawn and verified. Data reduction hardly affects the results of data analysis subsequently. There are two methods used commonly for data reduction: dimensionality reduction and numerosity reduction (Ghojogh & Crowley, 2019). The former usually uses linear algebra methods such as PCA, SVD, FLDA and DWT, the typical method of the latter is SOS (Kalegele et al., 2013).

Data Annotation. Data Annotation is the process of adding metadata to the data enabling modelling (Nagowah et al., 2019) which is adding necessary contexts of color, texture, shape, keywords, or semantic information. Data Annotation involves coding, rating, grading, tagging, and labeling of data (Carpenter, 2008).

Data Wrangling tools include Excel, SQL, Python, R, and Trifecta. In traditional data warehousing, Data Wrangling was carried out using Extract-Transform-Load (ETL) platforms, with significant manual involvement in specifying, configuring or tuning many of them (Koehler et al., 2017). It is needed to adopt adaptive, pay-as-you-go solutions that automatically tune the wrangling process, which means that users are able to contribute effort to the process of data wrangling in whatever form and at whatever moment they choose (Furche et al., 2016).

3.3 Data Computing

The next step of data preparation is to gain an insight into the value from big data and solve various problems of big data. Big data computing is an effective way that combines large scale compute, new data intensive techniques and mathematical models to build data analytics (Kune et al., 2016).

The four dimensions (volume, variety, velocity, and veracity) of big data bring challenges to traditional methods of data computing. Firstly, the volume of big data is exploding. Data computing can no longer be done by a single computer, but can only be shared by multiple machines. Secondly, the types of big data is various. Christie Schneider, a Watson marketing lead of IBM, claimed that more than 80% of today's data is unstructured (Schneider, 2016). It is hard to present in columns and rows and calculated in a structured database. The unstructured data is a big hurdle in computing and analysis part as they do not have a common format (Prasad & Agarwal, 2016). Thirdly, the data is growing at a high speed, and about 1.7 megabytes of fresh information will be created every second by 2020 (Dadheech et al., 2019). Fourthly, veracity of the data will directly affect the results of data calculation and data analysis. Veracity is probably the toughest nut to crack, and one of the biggest problems with big data is the tendency for errors to snowball (Tee, 2013). There are a lot of false, noise, or dirty data mixed in with valuable data. How to calculate these data is a big problem for traditional data computing technology.

Therefore, the traditional methods of data computing are not suitable for big data. A solution dividing data into small data units and calculating in the storage place where the data is located was found. Cloud computing is a distributed computing paradigm and differs from traditional ones such as centralized computing and grid computing. The emergence of MapReduce promoted the development of big data computing, and it has quickly become the current mainstream tool of the big data computing models. There are also some representative tools of cloud computing: Google GFS, Google BigTable, Spark and YARN.

Batch computing and stream computing are two important forms of big data computing (Sun et al., 2015). Batch computing is a big data computing method in which data is collected uniformly, stored in a database, and then processed in batches. Streaming computing is to process the data stream, which is a method of requiring real-time computing. Nowadays, there are many researches and applications related to batch computing. The basic MapReduce model and its implementations like Hadoop, is completely focused on batch processing (Shahrivari, 2014). Hadoop provides a distributed file system and off-line batch computing framework (Lin et al., 2013). The traditional batch computing process is carried out after a certain amount of data is accumulated; stream computing can achieve real-time processing and effectively reduce processing delay. While big data is becoming ubiquitous, the demand for large-scale processing of data streams is becoming increasingly urgent, which leads to the sprout of many distributed stream computing systems (Lu et al., 2014). The current widely used stream computing frameworks are Spark Streaming, Flink, Storm, S4 and Kafka.

3.4 Data Management

Data management refers to management activities, including acquiring, validating, storing, and processing required data to ensure the accessibility, reliability, and timeliness of the data to users (Myers, 2019). Data Management Maturity (DMM) model is a comprehensive framework designed in 2014 by CMMI of data management practices in six key categories include: Data Strategy, Data Governance, Data Quality, Data Operations, Platform & Architecture, Supporting Processes. According to the CMMI's introduction (2019), it is used to "provide

the best practices to help organizations build, improve, and measure their enterprise data management capability allowing for timely, accurate and accessible data across your entire organization". While in the age of big data, data management is the practice of organizing and maintaining data processes to meet ongoing big data lifecycle needs. Managing the ubiquitous big data is a challenge for data scientists. A database is a structured collection of data stored in a computer (Bai & Bhalla, 2020). Users store the data of transactions to be managed in the database, which helps to organize, maintain, process, and utilize data more conveniently. Data scientists must be proficient in not only traditional relational database, but also some emerging technologies such as NoSQL, NewSQL and relational cloud for data management.

Database, Data Warehouse and Data Lake are different data storage approaches for data management. Table 1 shows the studies that related to database, data warehouse, data lake.

Table 1 Comparison between database, data warehouse and data lake

Comparing Dimension	Database	Data Warehouse	Data Lake
oriented	application –oriented (Velicanu & Matei, 2007; Bontempo & Zagelow, 1998; Warners & Randriatoamanana, 2016)	subject –oriented (Velicanu & Matei, 2007; Bontempo & Zagelow, 1998; Warners & Randriatoamanana, 2016; Meredith et al., 2008)	operation –oriented (John & Misra, 2017)
data	Structured (Bontempo & Zagelow, 1998)	Structured (Bontempo & Zagelow, 1998)	Structured, Semi –structured, Unstructured (original forms) (Khine & Wang, 2018; John & Misra, 2017)
objective	support the operational system (Velicanu & Matei, 2007; Meredith et al., 2008; Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003)	support the decision –making system (Velicanu & Matei, 2007; Meredith et al., 2008; Vaisman & Esteban,2014; Mallach, 2000; Lechtenbörger & Vossen, 2003)	Support dynamic analytical applications (for query) (Khine & Wang,2018; John & Misra,2017)
integration	Limited integration (Bontempo & Zagelow,1998)	Integrated (Bontempo & Zagelow,1998; Warners & Randriatoamanana,2016; Meredith et al.,2008; Lechtenbörger & Vossen,2003)	Integrated (Miloslavskaya & Tolstoy,2016)
Update frequency	High (Bontempo & Zagelow, 1998; Meredith et al., 2008; Vaisman & Esteban,2014)	Low (Bontempo & Zagelow, 1998; Warners & Randriatoamanana,2016; Meredith et al., 2008; Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003)	High (Khine & Wang,2018; Aftab & Siddiqui,2018)
Usage	Predictable retrieval (Bontempo & Zagelow,1998; Meredith et al., 2008; Vaisman & Esteban,2014)	Ad hoc retrieval (Bontempo & Zagelow,1998; Meredith et al., 2008;Vaisman & Esteban,2014; Lechtenbörger & Vossen,2003)	Real time analytics (Madera & Laurent,2016)

Comparing Dimension	Database	Data Warehouse	Data Lake
Data modeling	UML, ER model (Meredith et al., 2008;Vaisman & Esteban, 2014;Lechtenbörger & Vossen, 2003)	multidimensional model (Meredith et al., 2008;Vaisman & Esteban, 2014; Lechtenbörger & Vossen, 2003; Mallach, 2000)	No specific data model (John & Misra,2017;Aftab & Siddiqui, 2018)
User type	Operators,office employees (Vaisman & Esteban,2014)	Managers,executives (Vaisman & Esteban, 2014), analysts (Meredith et al., 2008)	Data Scientists (especially those familiar with domain) (Madera & Laurent,2016)
Access frequency	High (Meredith et al., 2008; Vaisman & Esteban,2014)	From medium to low (Meredith et al., 2008;Vaisman & Esteban,2014)	accessible as soon as it is created (Khine & Wang,2018; Miloslavskaya & Tolstoy,2016)
Access type	Read,insert,update,delete (Meredith et al., 2008;Vaisman & Esteban,2014)	Read,append only (Vaisman & Esteban,2014),select (Meredith et al., 2008)	Read and write (Khine & Wang,2018;Miloslavskaya & Tolstoy,2016)
Response time	Short (Vaisman & Esteban, 2014;Lechtenbörger & Vossen, 2003)	Can be long (Vaisman & Esteban, 2014)	Short (Madera & Laurent, 2016)
normalized level	normalized tables (Warners & Randriatoamanana,2016; Meredith et al., 2008;Vaisman & Esteban,2014)	non–normalized(Warners & Randriatoamanana,2016; Meredith et al., 2008;Vaisman & Esteban,2014)	non–normalized(Southwick, et. al., 2015)

3.5 Data Analysis

Exploratory Data Analysis (EDA) was proposed by American statistician John Tukey in the 1970s (Tukey, 1977). Any activity of a Data Science project starts with EDA (Putatunda et al., 2019). When data scientists are faced with a variety of messy "dirty data" and do not know how to understand the data immediately. Exploratory data analysis is an effective way to help them achieve the purpose of data understanding and lay the foundation for subsequent data analysis.

According to purposes, the data analysis can be divided into descriptive analysis, predictive analysis and prescriptive analysis (Sivarajah et al., 2017). Descriptive analysis is most used in business analysis and it mainly solves the problem of "what has happened" by analyzing the collected data to obtain various quantitative characteristics reflecting objective phenomena. It includes data dispersion analysis, concentration analysis and frequency analysis. Predictive analysis is based on history and facing the future and it mainly focuses on what will happen in the future by the means of data mining and statistical modeling tools to analyze historical data, so as to predict what will happen in the future or the probability that something will happen. A typical method of predictive analysis is time series analysis. Finally, prescriptive analysis is a practice-oriented method mainly to solve the problem of "what should be done"

by analyzing what has happened, the causes of events and various possibilities in order to help users determine and choose the best actions and measures.

Traditional data analysis is deeply influenced by the formal theories of statistics (Tukey & Wilk, 1966). It mainly uses sampling data to infer the real situation, which means that traditional data analysis needs to extract useful information when the amount of data is limited. With the rapid growth of data volume in modern society, traditional data analysis is gradually turning to big data analysis. Big data analysis is mainly to acquire insights from all data (not sampling data) to support decision making, without considering the distribution status of data and without hypothesis testing. Traditional data analysis tools are not enough to manage big data, so some open-source tools for big data analysis are indispensable to Data Science. The most popular open-source tools for big data analysis are R and Python.

3.6 Data Products Development

Data Product is a kind of product that facilitates an end goal through the use of data (Patil, 2012). Data product development is indispensable for Data Science. Data product development activities are rarely undertaken in a traditional product development sequence that involves identifying the need, developing the product. On the contrary, data product development activities often take place in a continuous, iterative fashion, with the important activities conducted in parallel (Davenport & Kudyba, 2016). And the ability to develop data products is becoming increasingly critical to every business in big data era. Therefore, one of the missions of Data Science project is to develop data products.

Unlike traditional industrial products, data products can be entities or invisible objects. Cao (2017) defined data product as an output of Data Science which is "from data, or is enabled or driven by data, and can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system". Data product refers to anything that can help others use data to achieve goals. For example, Google Glass is a data product enabled by Google big data. Data products include data set products, information products, knowledge products, and intelligence products.

Data product development involves all activities of the Data Science project process, including datafication, data munging, data tidying, exploratory data analysis, data analysis, data product development. Not only is the result of a Data Science project a data product, but the intermediate product created by each activity is also a data product. Data Jujitsu is the art of turning data into products (Patil, 2012). It focuses on that the process of data product development must be highly artistic and centered on target users.

4 Domain–specific Data Science

Researchers from different disciplines have shown their own distinct concerns and perspectives on Data Science. The new term of Data Science and its variant concepts are widely used in Domain-specific Data Science. There are nine hot topics in domain-specific Data Science literature.

4.1 Data Journalism

As one of the new research directions of Journalism, Data Journalism is a way of seeing journalism as interpolated through the conceptual and methodological approaches of computation and quantification in the era of big data (Parasie & Dagiral, 2013; Lewis, 2015). The development of data journalism has roughly gone through the following three stages. At the

beginning, the typical event is the report titled Investigation of the Education System for Juvenile in The Guardian in 1982. This report breaks the narrative mode of traditional news (Timetoast, 2021). However, data journalism in that period lacked the necessary technical means, and there was no systematic theoretical support and was only scattered attempts. In the period of precision journalism, the typical event is Meyer's "Precision journalism: A reporter's introduction to social science methods (Meyer, 2002). Statistical and social survey research methods are introduced into news practice to collect data scientifically and improve the accuracy and objectivity of data in news reports. Although the data at this stage was paid attention to, news reports were still mainly narrative. Figures and charts were also only auxiliary to news reports. In the period of data journalism, the typical event is the establishment of Data Journalism Awards in 2012. This marks the beginning of data journalism as a new form of news that has received widespread attention.

Research hotspots of data journalism include the following aspects: Talent training of data journalism, presentation of data journalism, combination of data journalism and artificial intelligence. The talents required for data journalism have interdisciplinary characteristics. They not only need to understand traditional news acquisition and editing methods, but also need to master the theories and techniques of Data Science. Therefore, how to establish a new talent training model that meets the needs of data journalism has become a hot issue discussed by academia and journalism. For example, "The Data Journalism Handbook" co-written by journalists from various countries provides practical operating procedures and classic cases for talent training (Gray et al., 2012). Computational Journalism courses are available at Columbia Journalism School and Stanford University (Jstray, 2012; Nguyen, 2016). The presentation of data journalism has changed from the original table and visualization to the present storytelling and gamification. This has always been a hotspot in this field. Storytelling of data is different from visualization. It can provide data journalism with stronger narrative and better user experience. Gamification is the use of games to tell data stories, which can capture readers' interest more than traditional news. The combination of data journalism and artificial intelligence has become a new trend in the development of data journalism. For example, Reuters began to use artificial intelligence technology to track social media networks such as Twitter to obtain newsworthy events and discussions in 2017 (Matthews, 2019). This approach can save a lot of labor and time, so it has been imitated by many news media. The Post uses artificial intelligence technology to combine data with story templates to develop software for automatic news writing (Underwood, 2019).

The research challenges of data journalism mainly include two levels: data and technical. The difficulty at the data level lies in how to judge the reliability of the data source and how to ensure that the data is true during data processing and analysis. Common data sources currently include publications, traditional media, the Internet, government public data, and public geospatial data. It is generally believed that government public data and publications have high credibility, while the data authenticity on the internet is low. Data authenticity is the life of data journalism, but it is not easy to achieve data authenticity in the practical applications. For example, operations such as cleaning, conversion, merging or reshaping data during data processing may inadvertently cause data errors. Thus, the data is not true. Technical difficulties, from the perspective of practitioners of data journalism, are mainly the difficulty of mastering new technologies. For example, it is not easy to master the technologies commonly used in the field of data journalism, such as data analysis models, artificial intelligence technologies, Java, HTML, and Python. From the reader's point of view, technology in-

evitably increases the cognitive burden when it provides an intuitive presentation effect. For example, Java-based interactive data journalism often requires the installation of corresponding software to run normally.

There are many typical applications of data news, which mainly focus on health news and financial news. Take health news as an example, "Battling Infectious Diseases in the 20th Century: The Impact of Vaccines" from The Wall Street Journal (DeBold & Friedman, 2015). The background of the work was that many families in the United States believed that the children were too young to have an inoculation, which would affect the health of the children. Using a calendar heat map, based on 70 years of publicly available government data in the United States, the team created the work, which shows that vaccination has led to a significant decline in the number of people falling ill from the epidemic. This work reassured fearful parents with real data.

The main breakthrough of data news is the construction of professional data news teams and the positive usage of new technologies. The production of data news is a process in which multiple departments cooperate and multi-types of work full participate. It is necessary to break the internal departmental restrictions of traditional news organizations and establish a professional data news team that meets the new needs of the development of data news. Team members should cover all types of work of data news production, from journalists and editors to designers and technicians. At present, data news mainly uses data visualization technology to visualize data. However, data visualization is not the only choice. The rise of artificial intelligence technology provides a new opportunity for the development of data news. The automatic generation of data news has become a new trend in data news utilization technology.

4.2 Industrial Data Science

Industrial big data mainly studies how to apply big data in the field of industrial manufacturing, so as to realize the innovation of industrial manufacturing. Different from the previous focus on internally structured data, industrial big data needs to focus on the entire life cycle data of products and services in the industrial field, including structured and unstructured data (Ministry of Industry and Information Technology of the People's Republic of China, 2020).

The research hotspots in the field of industrial big data can be divided into two aspects: How to establish industrial big data and how to use industrial big data. Different countries have put forward different plans on how to establish industrial big data, the most representative of which are German Industries 4.0, Industrial Internet of the United States and Made in China 2025. In the practice of industrial big data, research hotspots can be subdivided into: attribute data extraction (Ma et al., 2014), data management philosophy and standards (Zhou et al., 2016), data-centric business operations, Physical deployment, cloud storage and supporting software platform deployment of the Internet of Things (Raptis et al., 2019). How to use industrial big data? There are four types of common products, respectively is process visualization, process optimization, decision support, fault detection. Process visualization and fault detection are mainly oriented to the production field, and visualization 3D modeling algorithm (Yandun et al., 2020) and fault identification algorithm (Dahbura & Masson, 1984) are the research hotspots at present. Process optimization and decision support are oriented to the activities in multiple fields of production and management. At present, the

research focus is the establishment of process optimization model, decision algorithm and relevant management formulation (Kruzhilko & Maystrenko, 2019; Hollowell et al., 2019).

The research challenges of industrial big data are divided into: in the industrial scenario, the data format of multiple data sources is not uniform. There are structured data, unstructured data, and non-digital data (hand-drawn charts). How to digitize and unify the format of these data is one of the difficult problems faced by industrial big data. In addition, data is generated in real time in industrial scenarios. How to store and analyze these data is also difficult. Moreover, the high utilization of data requires the establishment of supporting data collection, storage, processing, analysis, utilization standards and long-term mechanisms. But the reality is that the establishment of utilization mechanism of industrial big data requires a lot of manpower and financial resources. Finally, it is difficult to find new insights from massive amounts of data. This requires professional domain knowledge, keen insight and Data Science capabilities, but there is currently a lack of such compound talents (Davenport & Patil, 2012).

Typical applications of industrial big data: In the automobile manufacturing industry, automobile manufacturers collect vehicle conditions and operating habits of the owner through the sensors that come with the automobile, and then analyze the returned data to improve services and quality of products. In addition, the use of industrial big data can also help factories quickly discover machine failures and deal with them in time to reduce losses.

The main breakthrough of industrial big data lies in the guidance of government's policy. Because companies that develop industrial big data require high upfront investment, they are currently dominated by large companies. The participation of small and medium-sized enterprises is not enough, and if things go on like this, they will lose their competitiveness. Therefore, the guidance of government's policy can help small and medium-sized enterprises to participate in industrial big data. At the same time, it can also guide sharing and co-construction of data between enterprises to save social costs.

4.3 Business Data Science

Business Big Data was used to support business decision or produce products via precision marketing, user profiling and advertising. Consumption big data comes from the links related to product sales, such as customer registration data, order data, browsing record, purchase record, evaluation, consultation, feedback, complaint, suggestion. Research in this field focuses on how to use consumption big data. According to the analysis purpose, it can be divided into descriptive analysis, predictive analysis, diagnostic analysis and prescriptive analysis. Descriptive analysis mainly uses the descriptive statistical information of the data, such as median, mean value, standard deviation, to understand the distribution and characteristics of the data, so as to help the merchants understand the current situation of the goods as a whole. Predictive analysis mainly studies how to use models to predict unknown situations, such as establishing linear regression models using economic and population variables to predict electricity consumption (Bianco et al., 2009), using multiple random forests to predict urban water consumption (Chen et al., 2017), and using neural network technology to predict the online buying behavior of Indian buyers (Prashar et al., 2016). Diagnostic analysis mainly looks for the reasons that influence buying behavior, such as discussing the influence of advertising, social media (Zhang & Pennacchiotti, 2013) and website functions (Zhao et al. , 2016) on buying behavior. Normative analysis mainly studies how to make plans to increase

product sales, such as bundle sales (Kaserman, 2007), bonus incentive policies (Chung et al., 2014), comprehensive sales strategies (Leigh & Marshall, 2001).

The research challenge of consuming big data is not the technology of data collection and utilization, but in the legitimacy of data collection and utilization. In 2018, the California Consumer Privacy Act of 2018 (CCPA) was issued by the California Government of the United States, which restricts some of the rights of enterprises to collect and use information, and increases the right to know and Opt-Out right of users. Cambridge Analytica obtained data of as many as 87 million people from Facebook (including sensitive data such as personal accounts, personality tests and social networks of users), and sold it to the Trump presidential campaign to accurately display customized messages for specific groups of people (Grothaus, 2018). As a result, Facebook was fined $643,000 and Cambridge Analytica went bankrupt (Zialcita, 2019). Therefore, how to reasonably collect and use data under legal circumstances is an important problem faced by the consumption big data.

Recommendation system is a typical application in consumption big data. It predicts users' shopping tendency based on user-related consumption big data, so as to select products similar to users' buying tendency for recommendation. For example, Companies like Taobao, Youtube use recommender systems to help their users to identify the correct product or movies. According to a McKinsey survey, Netflix saves the company about $1 billion a year. Amazon owes 35% of its annual revenue to the recommendation system (Sigmoidal, 2017).

The main breakthrough in consumption big data is how to establish long-term relationship with users, which can be divided into two levels. For the first level, as privacy protection is getting more and more attention, it is needed to obtain user authorization if you want to collect and use data, and establishing trust relationship with users is helpful to obtain user authorization. As for second level, enterprises need to establish user loyalty programs to maintain user loyalty through organizing activities and giving small gifts regularly.

4.4 Health Data Science

With the gradual popularization of cordless medical treatment, electronic medical records, and online consultation, the work process in the medical and health field tends to be digitized, resulting in health big data. It mainly focuses on the wide application of big data in health and medical fields including life logging (Gurrin et al., 2014), medical diagnosis, pharmaceutical production, and health care (Raghupathi & Raghupathi, 2014).

The research hotspots of health big data mainly include precision medicine, disease identification and monitoring, and data privacy. For precision medicine, most of the existing research is to explore its feasibility and racial bias (Gurrin et al., 2014). For disease identification and monitoring, it involves the application of machine learning, natural language processing and other technologies in the health field. For example, Automated Identification of Surveillance Colonoscopy in Inflammatory Bowel Disease Using Natural Language Processing (Hou et al., 2013), Prediction of fatty liver disease using machine learning algorithms (Wu et al., 2019). For data privacy, the focus is on emphasizing the importance of privacy and proposing feasible ways to protect privacy. For example, the privacy issues were explained in health big data from a legal and technical perspective (Mounia & Habiba, 2015). A security life cycle model for health big data was proposed (Abouelmehdi et al., 2018). And a security framework and algorithm were built (Chandra et al., 2017).

The difficulty in health big data research is that health data involves important private data

of patients. Therefore, how to ensure the safety of health data in the process of transmission, storage and analysis is very necessary (Mooney & Pejaver, 2018). But at present, security in health big data has not been paid enough attention. The reason is that there is currently no relevant law clarifying the responsibilities of owners of health data. In addition, the protection of data security requires a lot of manpower and financial resources. For example, Community Health Systems was exploited by hackers to obtain social security numbers, dates of birth, phone numbers and actual addresses of 4.5 million patients in 2014. In 2015, Medical Informatics Engineering, an electronic Medical record software company, leaked the data of 3.9 million patients. The leaked content included names, social security numbers, phone numbers, mailing addresses, dates of birth, diagnosis and other sensitive information (Lord, 2020).

The successful application of Google Flu Trends (GFT) is a typical application that utilizes big data on health. In 2009, Jeremy Ginsberg, Matthew H. Mohebbi and Rajan S. Patel published a paper titled "Detecting Influenza Using Search Engine Query Data Based on Search Engine Data" in Nature. This paper introduces GFT, a flu prediction tool launched by Google in 2008, which can predict the nationwide spread of H1N1 in real time, overcoming the lag of official data release. The successful application of GET plays an important role in promoting the application of health big data.

The main breakthrough of health big data is that data masking can be used to protect the privacy of patients. Common technologies of data masking include data encryption, data randomization and data replacement technologies. Among them, data encryption is a reversible method of data masking. This may be cracked through the ciphertext, and the data needs to be decrypted before being used. Data randomization means that when collecting customer information on the server side, if only interested in the attributes of the information in the overall statistical sense, the client can use random algorithms to interfere with data privacy. For example, some real information is randomly deleted, and some false information is introduced to protect personal privacy. This technology can meet the needs of aggregated attribute while desensitizing data. Data replacement includes data pseudonymization, shuffling, and synthetic data.

4.5 Biological Data Science

Harnessing powerful computers and numerous tools for data analysis is crucial in drug discovery and other areas of big-data biology (Marx, 2013). The principles, theories, methods, technologies, and tools of big data are widely adopted to biology, and biological research paradigm is transferring from knowledge-centered paradigm to data-centered paradigm.

Its research problems include three aspects: (1) the development of gene analysis towards "de-sampling", scientists manage to apply big data technologies to efficiently analyze all data of DNA and RNA, instead of sampling analysis. (2) The traditional methods of biological research are to examine and determine the structure and characteristics of the subject using a variety of experimental techniques, such as NUCLEAR magnetic resonance and X-ray crystallography, and new methods such as cryo-electron microscopy, but these methods rely on a wide range of trial and error. The development of big data makes it possible to make strong predictions about complex structures through deep learning. The AlphaFold (2020) used by CASP14, for example, creates an attention-based neural network system that treats protein residues as nodes, connecting neighboring residues together. (3) The transformation of drug discovery to "precision". The example is the analysis of HIV drug resistance. Stanford

University in the United States established a special database, Hivdb. By sequencing HIV from patients in the database and comparing it with standard sequences, drug-resistant mutations can be found to know which drugs are no longer effective for that particular patient, and the remaining drugs can be combined to suppress HIV (Stanford University, 2021).

Big data research focus in the biology mainly includes: "gene sequencing + artificial intelligence" and "deep learning + medical image" and "big data + health records" (1) "gene sequencing + artificial intelligence" refers to the use of machine learning methods, prediction on the genome will change the characteristics of human body/disease/how to impact on phenotype. The implementation method is divided into two steps. First, identify the gene susceptibility locus associated with a characteristic/disease/phenotype. Second, use machine learning to simulate changes in characteristics/diseases/phenotypes. (2) "Deep learning + Medical imaging" refers to the direct analysis of medical images by deep learning algorithm. The existing image processing method is to treat each layer of 3d medical image as 2d image separately, and there are also methods to directly process 3D image after reducing complexity. The detection methods can also be divided into the method of locating and then classifying and the method of directly predicting the target location. (3) "Big data + Health archive" refers to the information about personal lifelong health status and health care behaviors managed electronically, which involves all the process information of patient information collection, storage, transmission, processing, utilization, and integrates information into a huge database. For example, the data volume of PubMed, the internationally famous biomedical database, reaches nearly 20 million records, increasing at a rate of 600,000 to 700,000 each year. The biomedical and pharmacological literature database Embase has more than 11 million records, with 500,000 more records added every year.

Medical Ethics and data security in the era of Big data. On the one hand, the development of science and technology is increasingly dependent on big data, and open source and data sharing have become an important driving force for biological research. But as concerns grow about privacy, particularly genomic privacy, access to important information, such as personal genome data, may be restricted in the future. On the other hand, the more involved the patient, the more likely the biomedical research project is to succeed. However, how to benefit the patients and how to share the benefits is a problem that people face.

Typical applications of big data biology: (1) clinical effect testing. For example, Germany's RWTH Aachen university (RWTH helmholtz-institute for biomedical engineering), German cancer research center (DKFZ), German cancer research association (DKTK) and Heidelberg (NCT) national center for tumor disease scientists have developed a kind of adaptive algorithm, can be directly according to the tumor HE staining tissue slice image prediction of microsatellite instability (MSI), which helps to identify potential can benefit from the immune therapy of gastrointestinal cancer patients. (2) Establish a library of genomics pre-training models. 23andme, an emerging technology company in Silicon Valley, takes the lead in the commercialization of precise SEQUENCING of DNA sequence to deal with diseases caused by genetic code. Based on Apriori algorithm and linear recursive model, Goran Hrovat utilizes big data visual analysis technology to explore patient data and serve for hospital management and decision-making.

4.6 Social Data Science

Social big data comes from joining the efforts of the two previous domains: social media and big data (Bello-Orgaz et al., 2016). Applications of social big data can be extended to a

wide number of domains such as health and political trending and forecasting, hobbies, e-business, cyber-crime, counterterrorism, time-evolving opinion mining, social network analysis, and human-machine interactions. Its research problems mainly include two aspects: big data based on content and big data based on opportunity network. The former focuses on extracting insight from user-generated content across a variety of social media platforms, while the latter focuses on extracting knowledge from interactions between online users by analyzing the web (Zhang et al., 2019).

The research hotspots of social big data mainly include the development and improvement of data mining and data analysis technologies used in social big data and the research on the methods of applying big data to different social fields, such as e-commerce, marketing, journals and public policies.

The research challenges of social big data include knowledge representation, data management, data processing, data analysis, data visualization and other aspects for mass data (Kaisler et al., 2013). Specific examples include accessing a large amount of unstructured data, determining how much data is sufficient to have a large amount of high-quality data, dealing with dynamically changing data streams, or implementing sufficient privacy (ownership and security). One of the most challenging problems is to identify valuable data from large heterogeneous datasets from social media, and analyze that data to discover useful knowledge and improve decisions for individual users and businesses. In order to correctly analyze social media data, traditional analysis techniques and methods need to adapt to and integrate the new big data paradigm to form structured data processing.

Typical applications of social big data include myriads of applications related to marketing, crime analysis and user experience. Marketing applications include advertising on social platforms. Maurer and Wiegmann (2011) analyzed the effectiveness of advertising on social networks. The experiment found that when the social network ads were placed in front of unscreened subjects alone, most of them thought the Facebook ads were annoying, and that placing the same ads on social interactions generated by Facebook tools and applications increased the number of visits and purchases by Jigar consumers. Applications of crime analysis include identifying patterns of crime through big data, allowing the detection and discovery of crimes and their relationships with criminals. Crime hotspots can be identified using a variety of mapping techniques, such as point mapping, geographic area thematic mapping, spatial ellipse, grid thematic mapping, and kernel density estimation (KDE). For User Experience-based Applications, Big data from social media needs to be visualized for better user experiences and services. For example, large amounts of digital data (usually in tabular form) can be converted to different formats. Thus, user intelligibility can be improved. The ability to visualize such big data to support timely decision-making is critical in areas as diverse as business success, drug therapy, network and national security, and disaster management (Keim et al., 2013). Therefore, user experience-based visualization is recognized as an important tool to support decision making. Visualization is also recognized as an important data analysis tool for social media (Kotval & Burns, 2013). It is important to understand what users want from social networking services. There are many visual ways to gather (and validate) the user experience. One of the most famous ways is interactive activity data analysis.

The main breakthrough of social big data lies in the increasingly advanced analysis technology and the increasing risk of privacy leakage. Therefore, many researches on privacy protection have been put forward to solve the problems related to privacy. We can note that there are two well-known methods. The first is to take advantage of "k-anonymity," which is

an attribute of some anonymous data (Sweeney, 2002). Given private data and a specific set of fields, the system (or service) must make the data useful without identifying the body of the data. The second approach is "differential privacy,, which can provide an effective way to maximize the accuracy of statistical database queries while minimizing the opportunity to identify their records (Dwork, 2008).

4.7 Agile Data Science

Agile Big Data is a development methodology that copes with the unpredictable realities of creating analytics applications from data at scale (Jurney, 2017). It is helpful to develop agile software, manage agile projects and establish agile organizations. The philosophy and principles of agile big data include four aspects: componentization and platformization, unification and openness, standardization and interface, self-service and intelligence, and engine driving. Componentization and platformization refer to the modularization abstraction of big data processing links to form a componentized platform with high cohesion of multiple functions. Componentized platforms can be used independently with existing platform components or combined to solve more problems on different links. Unification and openness refers to achieving a balance between simplifying system complexity, improving management and control ability and enhancing fitness, and improving flexibility. Standardization and interface refers to the formation of a series of standardized protocols in big data processing links, including data namespace protocol/metadata and data type specification protocol/data Access Interface protocol/Query language protocol/data transmission protocol/data security protocol. Intersystem interactions are provided in the form of service interfaces and queue interfaces. Self-service and intelligent routine operations including self-service can be better supported in an automated manner; Self-service insight analysis can be better supported in an intelligent way. Engine-driven includes the introduction of advanced engine-driven capabilities to enable agile big data applications to reach external audiences more quickly and actively. At this time, big data applications themselves have become a powerful business-driven engine. Operations including self-service can be better supported in an automated manner; Self-service insight analysis can be better supported in an intelligent way.

The main research content of Agile big data includes three aspects: feature extraction, fusion encapsulation and service interface. 1) Feature extraction: the structured and unstructured data and semi-structured data for data integration and feature extraction, extracted the data of all kinds of different characteristics, including time, space, characteristics or other global features, implementation of data related to the location of the associated attributes, time, space and other observation attribute such as the characteristics of the description. 2) Fusion encapsulation: All kinds of extracted data features or preliminarily preprocessed data are encapsulated into data processing units with unified structure and format according to data processing characteristics and requirements of different computing models, forming standard analysis data sets and providing fast data adaptation for the upper mining and computing services. Metadata definition method and XML/JSON and other technologies can be used to realize the unified definition of different types of data units, and basic information and various attributes definition and description can be carried out for each type of unified data unit, including identification ID, basic attribute, semantic attribute, structural attribute, and other contents. 3) Service interface: encapsulated unified data unit data sets, according to different computing service model to realize fast data adapter, uniform data unit call interface design, through the interface definition and parameter setting unit to encapsu-

late data parsing, and the data sets of various attributes, such as structure information are extracted (Bello-Orgaz et al., 2016).

The key breakthrough of Agile Big Data is how to achieve a unified, standardized, modular and configurable big data architecture to solve the problem of difficult integration between different types of heterogeneous subsystems. Application functions can be combined with existing functional components, and the cost can be reduced through service reuse. The form of data exchanged between components should be standardized and interlaced. Components can be combined with minimal programming or configuration, standardized integration of common models and tools, and simplified usage to provide out-of-the-box data mining and analysis capabilities to non-programmers; Big data application whole process (collection, storage, analysis, management) visualization operation. Based on the iterative nature of the scientific data and using efficient componentized tools, for big data each function subsystem (modules) modular, standardized design model, and according to the actual demand fast quick selection, configuration, structures, large data prototype system, the rapid iteration big data analysis results, and adapt to changing needs, the prototype as soon as possible into a production system. In the process of rapid iteration, rapid feedback and closed-loop verification, customers can gradually complete the reform of system thinking and management thinking of big data analysis, and the principle of quick proof and lean design is the core goal of agile big data application.

5 Integrative studies of Domain–general Data Science and Domain–specific Data Science

There is a wide range of disciplines which have developed Domain-specific Data Science as discussed in Section 4. However, DSDS varies from one domain to another, and different DSDS domains have their own unique research perspectives and interests on Data Science. At the same time, there are also some studies, which focus on Data Science itself and are intent on building Domain-general Data Science.

5.1 Nexus between DGDS and DSDS

There are subtle nexuses between Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS).

First, their fundamental difference roots in the thinking paradigms: DGDS conforms to data-centered thinking, while DSDS is in line with knowledge-centered thinking. Knowledge-centered thinking pattern believes that data will be utilized effectively only when the causality in data is identified. Hence, data analysis in the past dedicates to find, validate and take advantage of causalities. That conventional thinking pattern is very effective in DSDS but inefficient in DGDS since it is hard to identify and validate a causality from large-scale data sets. As a result, the aims of data analysis have shifted from causal analysis to correlation analysis, which put more emphasis on the correlation analysis. In contrast with causality analysis, correlation analysis is time-saving and easy to put into practices. This separation of causality analytics and correlation analytics also triggers collaboration between data scientists and domain experts, and provides a new analysis pattern for DSDS data analysis. For example, applying data analysis to the banking industry can make it more agile. Bank of America has developed a virtual assistant called Erica, which uses predictive analysis and natural language processing to showcase information about bank transaction history or upcoming

bills for customers.

Second, DGDS focuses on the theoretical studies, while DSDS is centered on applied ones. DGDS involves basic theories and activities of Data Science, while DSDS focuses on the applications of Data Science in a specific domain. What DGDS has in common with DSDS is that Statistics, Machine Learning, Data Visualization, and Domain Knowledge are their theoretical bases, and the research in DGDS and DSDS drives the development of the Data Science. Applying DGDS to other specific domains is one of the popular topics in recent studies. Those specific domains include life science, medical care, social governance, education, and business management. As a result, some new research topics such as quantitative self, data journalism and big data analysis gained widely attention of data scientists.

Third, DSDS is domain-dependent, but DGDS is domain-independent. DSDS incorporates theories with domain knowledge and business practice, which was termed to address challenges that we are facing in data enriched offerings era. DGDS provides theoretical guidance for the practical application of DSDS, which involves knowledge that data scientists in every field should master. The core theories of DGDS include concepts, theories, methods, technologies, and tools focus on solving the problem of discipline construction. DGDS puts forward some methods, techniques, and tools at the macro level, which can be used in a few DSDS. It means that DSDS aims to use the core theories of Data Science to solve other disciplines' own problems.

Fourth, DSDS and DGDS possess complementary advantages. Table 2 shows the gaps between DGDS and DSDS in some dimensions of Data Science. Data Science projects are completed via collaborative efforts of domain-general data scientists as well as domain-specific data scientists. Data wrangling, for instance, is a value-added process that needs efforts from not only domain-general data scientists who are good at data-related tasks but also domain-specific data scientists who are familiar with specific business of application domains.

Table 2 the Comparisons of DGDS and DSDS

	DGDS	DSDS
Theory	*	
Practice		*
Data Wrangling		*
Data Computing	*	
Data Management	*	
Data Analysis		*
Data Products Development		*

5.2 Integrating DGDS and DSDS

Theoretical Data Science (TDS) is supposed to bridge the gap between Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS). Theoretical Data Science is a branch of Data Science that employs mathematical models and abstractions of data objects and systems to rationalize, explain and predict big data phenomena. This contrasts with DSDS, which uses casual analysis, as well as DGDS, which employs data-centered thinking to deal with big data problems in that it balances the usability and the interpretability of Data Science practices (Figure 2).

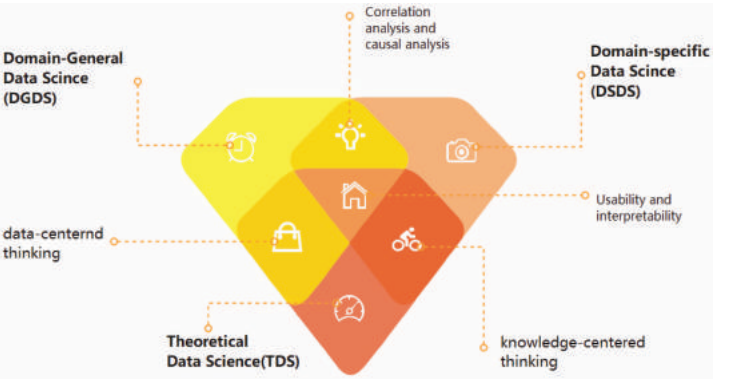


Figure 2 Three Types of Data Science

The main concerns of TDS are concentrated on the following topics:

(1) To integrate the data-centered thinking with the knowledge-centered thinking. Data-centered thinking is the unique thinking pattern of DSDS, while knowledge-centered thinking is the typical thinking pattern of DGDS. TDS integrates them by two different ways: data-centered thinking triggers knowledge-centered thinking, or vice versa. In practical Data Science projects, the integration of DSDS and DSDG is mainly implemented via the collaboration between professional data scientists and experts from other specific business domains.

(2) To transform correlation analysis into casual analysis. TGS believes that correlation analysis is insufficient to address big data problems, and the Data Science projects should convert correlation analysis into casual analysis. Further, TDS regards correlation analysis as the pre-requirements of causal analysis. Correlation analysis is conducted by employing machine learning or statistical methods. However, the causal analysis heavily depends on the related domain knowledge.

(3) To balance the usability and the interpretability. Contradictions between the usability and the interpretability of big data solutions are the trickiest challenges in Data Science studies. TGS balances them by introducing interpretable Machine Learning or explainable Artificial Intelligence. Interpretable methods of TGS can be classified into global interpretation and local interpretation. Global interpretability implies knowing what patterns are present in general, while local interpretability implies knowing the reasons for a specific decision (Doshi-Velez & Kim, 2017).

6 Conclusions

Theoretical Data Science (TDS) is an integrated study of Domain-general Data Science (DGDS) and Domain-specific Data Science (DSDS) in order to bridge the gaps between them. In contrast with DSDS as well as DGDS, TDS adopts the data-centered thinking pattern, recognizes that the property of data is more active than passive, manages to convert data into intelligence, solves data-intensive tasks, conducts data wrangling or munging, enhances user experiences of big data systems, introduces data intensive scientific discovery, as well as educates data scientists. TDS is unique in its scientific objectives as well as research paradigm, and does not replicate directly the experiences from DGDS and DSDS. The following topics are essential for further research on TDS.

1) To conduct in-depth theoretical research on Data Science. There are no shared understandings on Data Science yet. Some of the researchers insist that Data Science is merely interdisciplinary applications of Statistics and Machine Learning, and it does not need its own new theories. They argue that application of Statistics and Machine Learning is crucial for Data Science. They fail to admit the unique theories of Data Science. In fact, Statistics and Machine Learning are the theoretical foundation of Data Science, not its core components. Data Science is an independent discipline like Statistics and Machine Learning. TDS is unique in its scientific mission, research perspective, thinking pattern, underlying principles, and theoretical framework, which are distinct from other disciplines.

2) To take advantage of active property of big data. One of the main contributions of Data Science is that it shifts our thinking pattern and views big data as active beings. People have seen data as passive or dead thing to date, and how to input human intelligence into data is the main concern of the related studies. For instance, traditional data prepressing theories try to convert complex data into simple data through defining schema, data cleansing, and filling missing values. However, TDS highlights the active property of data and begins to discuss how to take advantage of data. As a result, some novel terms, such as data-driven applications, data-centric design, data insights, and big data ecosystem, are widely accepted. TDS regards complexity as a natural attribute of big data and does not conduct traditional data preprocessing. Admitting that data is active rather than passive is the basic starting point of studying TDS.

3) To introduce Design of Experiments into Data Science studies. Design of Experiments (DOE) is one of the essential activities of TDS projects. Data scientists should creatively propose research hypotheses according to the objectives of TDS projects, design corresponding experiments, conduct the data experiments and test the hypothesis. Taking the student programs of Data Science majors in the University of Washington as well as the University of California, Berkeley as examples, courses titled Applied Statistics & Experimental Design or Experiments and Causality are provided, respectively. The both courses focus on improving students' ability in DOE as well as hypothesis testing.

4) To shift Data Science´s research focus from correlation analysis into causality inference. There is a misconception that Data Science only concentrates on correlation analysis, and causality inference is outside the scope of it. However, correlation inference can only be used to identify the correlations in big data, but cannot guide how to optimize and intervene in the identified correlations. Where the correlation changes, the causation relation in big data is required to be analyzed. Hence, to shift the research focus from correlation analysis into causality inference is one of the unique purposes of TDS. In a TDS project, the data scientists are responsible not only to discover possible correlations in big data, but also to reveal the causality behind the correlations with the collaboration of domain experts. To embrace causality analysis is becoming one of the most discussed topics in Data Science. For instance, the course titled Experiments and Causality Analysis or the Causal Inference for Data Science are listed in DS courses at University of California, Berkeley, and Columbia University as well.

5) To take data product development as one of the main tasks of Data Science Projects. Developing data products is one of the distinct objectives of TDS studies. Data products in TDS are not limited to products in data form. All products that utilize data to provide new services should be regarded as a data product. Data can be used to promote product innovation, and traditional products will be transformed into data products by application of DS theories. Google Glasses, for instance, is a data product in that its novel features are derived

from data. Data-centered thinking is the fundamental difference between data products and traditional ones. Data products will be the most common applications of TDS.

Acknowledgements

This work was supported by the Ministry of education of Humanities and Social Science project (Project No.20YJA870003).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Reference

Abdallah, Z. S., Du, L. & Webb, G. I.(2017). "Data preparation" *in Encyclopedia of Machine Learning and Data Mining*. Springer Publishing Company, Boston.

Abouelmehdi, K., Beni–Hessane, A., & Khaloufi, H.(2018). Big healthcare data: preserving security and privacy. *Journal of Big Data, 5* (1), 1. doi: https://doi.org/10.1186/s40537–017–0110–7

Aftab, U., & Siddiqui, G. F.(2018). Big data augmentation with data ware–house: A survey. In *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, 10 –13 December, 2785 –2794. doi: https://doi.org/10.1109/BigData.2018.8622206

Amghar, S., Cherdal, S., & Mouline, S.(2019). Data Integration and NoSQL Systems: A State of the Art. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, New York, October 23–24, 1–6. doi: https://doi.org/10.1145/3372938.3372954

Bai, Y., & Bhalla S.(2020). Introduction to Databases. In Practical database *programming with Visual Basic*.NET, John Wiley & Sons. Hoboken.

Bello–Orgaz, G., Jung, J. J., & Camacho, D.(2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59. doi: https://doi.org/10.1016/j.inffus.2015.08.005

Bianco, V., Manca, O., & Nardini, S.(2009). Electricity consumption forecasting in Italy using linear regression models. *Energy, 34*(9), 1413–1421. doi: https://doi.org/10.1016/j.energy.2009.06.034

Bontempo, C., & Zagelow, G.(1998). The IBM data warehouse architecture. *Communications of the ACM, 41*(9), 38–48. doi: https://doi.org/10.1145/285070.285078

Cao, L.(2017). Data science: a comprehensive overview. *ACM Computing Sur–veys (CSUR), 50* (3), 1–42. doi: https://doi.org/10.1145/3076253

Chandra, S., Ray, S., & Goswami, R. T.(2017, January). Big data security in healthcare: survey on frameworks and algorithms. In *2017 IEEE 7th International Advance Computing Conference(IACC)*(pp. 89–94). IEEE. doi: https://doi.org/10.1109 / IACC.2017.0033

Chen, G., Long, T., Xiong, J., & Bai, Y.(2017). Multiple random forests modelling for urban water consumption forecasting. *Water Resources Management, 31* (15), 4715–4729. doi: https://doi.org/10.1007/s11269–017–1774–7

Chung, D. J., Steenburgh, T., & Sudhir, K.(2014). Do bonuses enhance sales productivity? A dynamic structural analysis of bonus –based compensation plans. *Marketing Science, 33* (2), 165 –187. doi: https://doi.org/10.1287/mksc.2013.0815

Cleveland, W. S.(2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review, 69* (1), 21–26. doi: https://doi.org/10.1111/j.1751–5823.2001.tb00477.x

CMMI.(2019). Data Management Maturity (DMM). Retrieved from https://cmmiinstitute.com/data–management–maturity

Dadheech, P., Goyal, D., & Srivastava, S.(2019). Information Management and Ma–chine Intelligence. *Proceedings of International Conference on Information Management& Machine Intelligence. Jaipur*, 14–15 December, 85–100. doi: https://doi.org/10.1007/978–981–15–4936–6_9

Dahbura, A. T., & Masson, G. M.(1984). An 0(n2. 5) fault identification algorithm for diagnosable systems. *IEEE Computer Architecture Letters,33* (06),486–492.doi:https://doi.org/486–492.10.1109/TC.1984.1676472

Das, S.(2021). *Data Science: Theories, Models, Algorithms, and Analytics*. Srdas.github.io. Retrieved 1 March 2021, from https://srdas.github.io/MLBook/index.html

Davenport, T. H., &Patil, D. J.(2012). Data scientist. *Harvard Business Review, 90*(5), 70–76. doi: https://doi.org/10.1007/s11213–012–9233–0

Davenport, T.H., &Kudyba, S.(2016). Designing and Developing Analytics–Based Data Products. *MIT Sloan Management Review, 58*, 83.

DeBold, T., & Friedman, D.(2015). *Battling Infectious Diseases in the 20th Century: The Impact of Vaccines*. WSJ. Retrieved 1 March 2021, from http://graphics.wsj.com/infectious–diseases–and–vaccines/

Dhar, V.(2013). Data science and prediction. *Communications of the ACM, 56* (12), 64–73. doi: https://doi.org/10.1145/2500499

Doshi–Velez, F., & Kim, B.(2017).Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Dwork, C.(2008, April). Differential privacy: A survey of results. In *Interna–tional conference on theory and applications of models of computation* (pp. 1–19). Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978–3–540–79228–4_1

Earley, S.(2014). Agile analytics in the age of big data. *IT Professional,16*(4), 18–20. doi: https://doi.org/10.1109/MITP.2014.44

Endel, F., & Piringer, H.(2015). Data Wrangling: Making data useful again. *IFAC–PapersOnLine, 48* (1), 111–112. doi: https://doi.org/10.1016/j.ifacol.2015.05.197

Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W.(2016). Data Wrangling for Big Data: Challenges and Opportunities. *19th International Conference on Extending Database Technology (EDBT)*. Bordeaux, 15–18 March, 473–478. doi: https://doi.org/10.5441/002/edbt.2016.44

Gao, J., Xie, C., & Tao, C.(2016). Big Data Validation and Quality As–sur–ance—Issuses, Challenges, and Needs. *2016 IEEE symposium on ser–vice–oriented system engineering (SOSE)*, Oxford, March 29–April 2, 433–441. doi: https://doi.org/10.1109/SOSE.2016.63

Ghojogh B., & Crowley M.(2019) Instance Ranking and Numerosity Reduction Using Matrix Decomposition and Subspace Learning. In *Canadian Conference on Artificial Intelligence*, Kingston, ON, 28–31 May, 160–172. doi: https://doi.org/10.1007/978–3–030–18305–9_13

Gray, J., Chambers, L., & Bounegru, L.(2012). *The data journalism handbook: How journalists can use data to improve the news*. O´Reilly Media, Inc.

Grothaus, M.(2018). *How our data got hacked, scandalized, and abused in 2018*. Fast Company. Retrieved 1 March 2021, from https://www.fastcompany.com/90272858/how–our–data–got–hacked–scandalized–and–abused–in–2018.

Gurrin, C., Smeaton, A. F., & Doherty, A. R.(2014). Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval, 8* (1), 1–125. doi: http://dx.doi.org/10.1561/1500000033

Han, J., Pei, J., & Kamber, M.(2011). *Data mining: concepts and techniques*. Elsevier, Morgan Kaufmann, Waltham.

Hollowell, J. C., Kollar, B., Vrbka, J., & Kovalova, E.(2019). Cognitive decision–making algorithms for sustainable manufacturing processes in Industry 4.0: Networked, smart, and responsive devices. *Economics, Management and Financial Markets, 14* (4), 9–15. doi: https://doi.org/10.22381/EMFM14420191

Hou, J. K., Chang, M., Nguyen, T., Kramer, J. R., Richardson, P., Sansgiry, S., ...& El–Serag, H. B.(2013). Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Digestive dis–eases and sciences, 58*(4), 936–941. doi: https://doi.org/10.1007/s10620–012–2433–8

John, T., & Misra, P.(2017). *Data lake for enterprises*. Packt Publishing Ltd, Birmingham.

Jurney, R. (2017) . *Agile data science 2.0: Building full–stack data analytics ap–plications with spark*. O´Reilly Media, Inc.

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W.(2013, January). Big data: Issues and challenges moving forward. In *2013 46th Hawaii international con–ference on system sciences* (pp. 995–1004). IEEE. doi: https:

//doi.org/10.1109/HICSS.2013.645

Kalegele, K., Takahashi, H., Sveholm, J., Sasai, K., Kitagata, G., & Kinoshita, T.(2013). Numerosity reduction for resource constrained learning. *Journal of Information Processing, 21* (2), 329–341. doi: https://doi.org/10.2197/ipsjjip.21.329

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ...& Buono, P.(2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization, 10* (4), 271–288. doi: https://doi.org/10.1177/1473871611415994

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ...& Kumar, V.(2017). Theory–guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering, 29* (10), 2318–2331. doi: https://doi.org/10.1109/TKDE.2017.2720168

Kaserman, D. L.(2007). Efficient Durable Good Pricing And Aftermarket Tie - In Sales. *Economic Inquiry, 45*(3), 533–537. doi: https://doi.org/10.1111/j.1465–7295.2007.00022.x

Keim, D., Qu, H., & Ma, K. L.(2013). Big–data visualization. *IEEE Computer Graphics and Applications, 33*(4), 20–21. doi: https://doi.org/10.1109/MCG.2013.54

Khine, P. P., & Wang, Z. S.(2018). Data lake: a new ideology in big data era. *4th Annual International Conference on Wireless Communication and Sensor Net–work (WCSN 2017)*, Wuhan, 15–17 December, 2017, 17. doi: https://doi.org/10.1051/itmconf/20181703025

Kitchin, R.(2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in human geography, 3*(3), 262–267. doi: http://www.wired.com/science/discoveries/magazine/16–07/pb_theory

Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A. A., ...& Paton, N. W.(2017). Data context informed data wrangling. In *2017 IEEE Inter–national Conference on Big Data. Boston*,11–14 December, 956–963. doi: https://doi.org/10.1109/BigData.2017.8258015

Konkel, L. (2020, March). *Who Will Benefit From Precision Medicine?. Who Will Bene–fit from Precision Medicine?* UC San Francisco. Retrieved from https://www.ucsf.edu/magazine/benefit–precision–medicine

Kotval, X. P., & Burns, M. J.(2013). Visualization of entities within social media: Toward understanding users´ needs. *Bell Labs Technical Journal, 17*(4), 77–102. doi: https://doi.org/10.1002/bltj.21576

Kruzhilko, O., & Maystrenko, V.(2019). Management decision–making algorithm development for planning activities that reduce the production risk level. *Journal of Achievements in Materials and Manufacturing Engineering, 93* (1–2). doi: https://doi.org/10.5604/01.3001.0013.4141

Kuacharoen, P.(2014). Combination of data masking and data encryption for cloud database. *Applied Mechanics and Materials*, 571–572, 617–620. doi: https://doi.org/10.4028/www.scientific.net/amm.571–572.617

Kune, R., Konugurthi, P., Agarwal, A., Chillarige, R., & Buyya, R.(2015). The anatomy of big data computing. *Software: Practice and Experience, 46* (1), 79–105. doi: https://doi.org/10.1002/spe.2374

Lechtenbörger, J., & Vossen, G.(2003). Multidimensional normal forms for data warehouse design. *Information Systems, 28* (5), 415–434. doi: https://doi.org/10.1016/S0306–4379(02)00024–8

Leigh, T. W., & Marshall, G. W.(2001). Research priorities in sales strategy and performance. *Journal of Personal Selling & Sales Management, 21*(2), 83–93. doi: https://doi.org/10.2307/20832582

Lenzerini, M.(2002). Data integration: a theoretical perspective. *Proceedings of the twenty–first ACM SIGMOD–SIGACT–SIGART symposium on Principles of data–base systems*. New York, 3–5 June, 233–246. doi: https://doi.org/10.1145/543613.543644

Lewis, S. C.(2015). Journalism in an era of big data: Cases, concepts, and critiques. *Digital Journalism, 3*(3), 321–330. doi: https://doi.org/10.1080/21670811.2014.976399

Lin, X., Wang, P., & Wu, B.(2013). Log analysis in cloud computing environment with Hadoop and Spark. *5th IEEE International Conference on Broad–band Network & Multimedia Technology, Guilin*, 17–19 November, 273–276. doi: https://doi.org/10.1109/ICBNMT.2013.6823956.

Lord, N.(2020, March). *Top 10 Biggest Healthcare Data Breaches of All Time*. Digital Guardian. Retrieved from https://digitalguardian.com/blog/top–10–biggest–healthcare–data–breaches–all–time

Lu, R., Wu, G., Xie, B., & Hu, J.(2014). Stream bench: Towards benchmarking modern distributed stream computing frameworks. In *20z14 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, London, 8–11 December, 69–78, doi: https://doi.org/10.1109/UCC.2014.15

Ma, C., Zhang, H. H., & Wang, X.(2014). Machine learning for Big Data analytics in plants. *Trends in plant science, 19* (12), 798–808. doi: https://doi.org/10.1016/j.tplants.2014.08.004

Madera, C., & Laurent, A.(2016, November). The next information architecture evolution: the data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, New York,1 –4 November, 174–180. doi: https://doi.org/10.1145/3012071.3012077

Mallach, E.(2000). *Decision support and data warehouse systems*. Irwin/McGraw–Hill.

Mansfield–Devine, S.(2014). Masking sensitive data. *Network Security, 10*, 17–20. doi: https://doi.org/10.1016/S1353–4858(14)70104–7

Marx, V.(2013). The big challenges of big data. *Nature, 498* (7453), 255–260. doi: https://doi.org/10.1038/498255a

Matthews, K.(2019). *AI in Data Journalism: Pros and Cons*. Datafloq.com. Retrieved 1 March 2021, from https://datafloq.com/read/ai–data–journalism–pros–cons/7116.

Mattmann, C. A.(2013). A vision for data science. *Nature, 493* (7433), 473–475. doi: https://doi.org/10.1038/493473a

Maurer, C., & Wiegmann, R.(2011). Effectiveness of Advertising on Social Network Sites: A Case Study on Facebook. In: Law R., Fuchs M., Ricci F.(eds) *Information and Communication Technologies in Tourism 2011*. Springer, Vienna.

Mayer–Schönberger, V., & Cukier, K.(2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Meredith, R., O´Donnell, P., & Arnott, D.(2008). Databases and data ware–houses for decision support. In *Handbook on Decision Support Systems 1*(pp. 207–230). Springer, Berlin, Heidelberg.

Meyer, P.(2002). *Precision journalism: A reporter´s introduction to social science methods*. Rowman& Littlefield Publishers.

Miles, M. B., & Huberman, A. M.(1994). *Qualitative data analysis: An expanded sourcebook*. SAGE Publications.

Miloslavskaya, N., & Tolstoy, A.(2016). Application of big data, fast data, and data lake concepts to information security issues. In *2016 IEEE 4th Internation–al Conference on Future Internet of Things and Cloud Workshops(FiCloudW)*, Vienna, 22–24 August, 148–153. doi: https://doi.org/10.1109/W–FiCloud.2016.41

Ministry of Industry and Information Technology of the People´s Republic of China.(2020, May). *Guiding opinions of the development of industrial big data from the Ministry of Industry and Information Technology*. Retrieved from https://www.miit.gov.cn/xwdt/gxdt/sjdt/art/2020/art_a61849ebec144ebdb91fa9bc5474554c.html

Mooney, S. J., & Pejaver, V.(2018). Big data in public health: terminology, machine learning, and privacy. *Annual Review of Public Health, 39*, 95–112. doi: https://doi.org/10.1146/annurev–publhealth–040617–014208

Mounia, B., & Habiba, C.(2015). Big data privacy in healthcare Moroccan context. *Procedia Computer Science, 63*, 575–580. doi: https://doi.org/10.1016/j.procs.2015.08.387

Müller, H., & Freytag, J. C.(2005). *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. FürInformatik.

Myers, R.(2019). D*ata Management and Statistical Analysis Techniques.* Scientific e–Resources. Waltham.

Nagowah, S. D., Sta, H. B., & Gobin–Rahimbux, B. A.(2019). Towards Achieving Semantic Interoperability in an IoT–enabled Smart Campus. In *2019 IEEE International Smart Cities Conference (ISC2)*, Casablanca, Morocco, 14–17 October, 593–598. doi: https://doi.org/10.1109/ISC246665.2019.9071694

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C.(2019). Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment, 12* (12), 1986–1989. doi: https://doi.org/10.14778/3352063.3352116

Naur, P.(1974). *Concise survey of computer methods*. Petrocelli Books.

Nguyen, D.(2016, March). *Computational Journalism at Stanford University | Computational Journalism, Spring 2016*. Retrieved from http://www.compjour.org/

Overton, J.(2016). *Going Pro in Data Science: What it Takes to Succeed as a Professional Data Scientist*. O´Reilly Media.

Parasie, S., & Dagiral, E.(2013). Data–driven journalism and the public good:"Computer–assisted–reporters" and

"programmer–journalists" in Chicago. *New Media & Society, 15* (6), 853–871.doi: https://doi.org/10.1177/1461444812463345

Patil D. J.(2012). *Data jujitsu: The art of turning data into data product.* O´Reilly Media, Inc., Sebastopol.

Prasad, B. R., & Agarwal, S.(2016). Comparative Study of Big Data Computing and Storage Tools. *International Journal of Database Theory and Application, 9*(1), 45–66. doi: https://doi.org/10.14257/ijdta.2016.9.1.05

Prashar, S., Vijay, T. S., & Parsad, C.(2016). Predicting online buying behavior among Indian shoppers using a neural network technique. *International Journal of Business and Information, 11* (2), 175. doi: https://doi.org/10.6702/ijbi.2016.11.2.3

Putatunda, S., Rama, K., Ubrangala, D., & Kondapalli, R.(2019). *SmartEDA: An R Package for Automated Exploratory Data Analysis*. arXiv preprint arXiv:1903.04754.

Raghupathi, W., & Raghupathi, V.(2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems, 2* (1), 3. doi: https://doi.org/10.1186/2047–2501–2–3

Raptis, T. P., Passarella, A., & Conti, M.(2019). Data management in industry 4.0: State of the art and open challenges. *IEEE Access, 7*, 97052–97093. doi: https://doi.org/10.1109/ACCESS.2019.2929296

Russom, P.(2011). Big data analytics. *TDWI best practices report, fourth quarter, 19* (4), 1–34.

Sarada, G., Abitha, N., Manikandan, G., &Sairam, N.(2015). A few new approaches for data masking. In *2015 International Conference on Circuits, Power and Computing Technologies*, Nagercoil, 19–20 March, 1–4. doi: https://doi.org/10.1109/ICCPCT.2015.7159301

Schneider, C.(2016). *The biggest data challenges that you might not even know you have–Watson Blog. Watson Blog.* Retrieved 1 March 2021, from https://www.ibm.com/blogs/watson/2016/05/biggest –data –challenges–might–not–even–know

Shahrivari, S.(2014). Beyond batch processing: towards real–time and streaming big data. *Computers, 3*(4), 117–129. doi: https://doi.org/10.3390/computers3040117

Sigmoidal.(2017). *Recommendation Systems – How Companies are Making Money – Sigmoidal*. Sigmoidal. Retrieved 1 March 2021, from https://sigmoidal.io/recommender–systems–recommendation–engine

Sivarajah, U., Kamal, M. M., Irani, Z., &Weerakkody, V.(2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263 –286. doi: https://doi.org/10.1016/j.jbusres.2016.08.001

Southwick, S. B., Lampert, C. K., & Southwick, R.(2015). Preparing controlled vocabularies for linked data: benefits and challenges. *Journal of Library Metadata, 15* (3 –4), 177 –190. doi: https://doi.org/10.1080/19386389.2015.1099983

Stanford University.(2021, March). *HIV drug resistance database*. Retrieved from https://hivdb.stanford.edu.

Sun, D., Zhang, G., Zheng, W., & Li, K.(2015). Key Technologies for Big Data Stream Computing. In *Big Data: Algorithms, Analytics, and Applications,* CRC Press, Boca Raton.

Sweeney, L.(2002). k–anonymity: A model for protecting privacy. International *Journal of Uncertainty, Fuzziness and Knowledge–Based Systems, 10*(05), 557–570. doi: https://doi.org/10.1142/S0218488502001648

Taleb, I., Serhani, M. A., & Dssouli, R.(2018, July). Big data quality: A survey. In *2018 IEEE International Congress on Big Data(BigData Congress)* (pp. 166–173). IEEE.doi: https://doi.org/10.1109/BigDataCongress.2018.00029

Tansley, S., & Tolle, K.(2009). *The fourth paradigm: data–intensive scientific discovery (Vol. 1)*. T. Hey(Ed.). Redmond, WA: Microsoft research.

Tee, J.(2013). *Four ´V´s of big data: volume velocity variety veracity*. TheServ–erSide.com. Retrieved 1 March 2021, from https://www.theserverside.com/feature/Handling–the–four–Vs–of–big–data–volume–velocity–variety–and–veracity

The AlphaFold team.(2020). AlphaFold: a solution to a 50–year–old grand challenge in biology. Deepmind. Retrieved 1 March 2021, from https://deepmind.com/blog/article/alphafold–a–solution–to–a–50–year–old–grand–challenge–in–biology

The Four V´s of Big Data. *IBM Big Data & Analytics Hub.*(2021, March). Retrieved from http://www.ibmbigdatahub.com/infographic/four–vs–big–data

Timetoast.(2021, March). *The history of data journalism timeline*. Retrieved from https://www.timetoast.com/time-

lines/the–history–of–data–journalism

Tukey, J. W.(1977). *Exploratory Data Analysis*. Pearson.

Tukey, J. W., & Wilk, M. B.(1966). Data analysis and statistics: an expository overview. In *Proceedings of the November 7–10, 1966, fall joint computer conference*, New York, 7–10 November, pp. 695–709. doi: https://doi.org/10.1145/1464291.1464366

Underwood, C.(2019). *Automated Journalism – AI Applications at New York Times, Reuters, and Other Media Giants | Emerj.* Emerj. Retrieved 1 March 2021, from https://emerj.com/ai–sector–overviews/automated–journalism–applications

Vaisman, A, & Esteban Z.(2014). *Data Warehouse Concepts in Data Ware–house Systems* (pp.75). Springer, Berlin Heidelberg.

Velicanu, M., & Matei, G.(2007). Database versus Data Warehouse. *Revista Informatica Economică*, 91–95.

Warners, H. L. H. S., & Randriatoamanana, R.(2016). Datawarehouser: A Data Warehouse artist who have ability to understand data warehouse schema pictures. In *2016 IEEE Region 10 Conference (TENCON),* Singapore, 22–25 November, 2205–2208. doi: https://doi.org/10.1109/TENCON.2016.7848419

Wickham, H.(2014). Tidy data. *Journal of Statistical Software, 59* (10), 1–23. doi: https://doi.org/10.18637/jss.v059.i10

Wu, C. C., Yeh, W. C., Hsu, W. D., Islam, M. M., Nguyen, P. A. A., Poly, T. N., ... & Li, Y. C. J.(2019). Prediction of fatty liver disease using machine learning algorithms.*Computer Methods and Programs in Biomedicine, 170*, 23–29. doi: https://doi.org/10.1016/j.cmpb.2018.12.032

Yandun, F., Silwal, A., & Kantor, G.(2020). Visual 3D Reconstruction and Dynamic Simulation of Fruit Trees for Robotic Manipulation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. doi: https://doi.org/10.1109/cvprw50498.2020.00035

Zhang, X., Wang, S., Cong, G., & Cuzzocrea, A.(2019). Social Big Data: Mining, Applications, and Beyond. *Hindaw, 12*, 3, 1749–1772. https://doi.org/10.1155/2019/2059075

Zhang, Y., & Pennacchiotti, M.(2013, May). Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web* (pp.1521 –1532). doi: https://doi.org/10.1145/2488388.2488521

Zhao, Y., Yao, L., & Zhang, Y.(2016). Purchase prediction using Tmall - specific features. *Concurrency and Computation: Practice and Experience, 28* (14), 3879–3894. doi: https://doi.org/10.1002/cpe.3720

Zhou, K., Fu, C., & Yang, S.(2016). Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews, 56*, 215–225. doi: https://doi.org/10.1016/j.rser.2015.11.050

Zialcita, P.(2019, October). Facebook pays $643,000 fine for role in Cambridge Analytica Scandal. Retrieved from https://www.npr.org/2019/10/30/774749376/facebook–pays–643–000–fine–for–role–in–cambridge–analytica–scandal

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Visual Analytics of Large-scale E-government Text Data via Simplified Word Cloud

Yanan Liu[a], Fang He[a], Jin Wen[a], Zhiguang Zhou[a, c], Jinchang Li [b*]

a. School of Information, Zhejiang University of Finance and Economics, Hangzhou, China
b. School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China
c. State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China

ABSTRACT

With the rapid development of Internet technology, a rich set of e-government data are collected by the government departments. For example, a variety of feedback text data can be obtained quickly and efficiently through various channels such as the mayor's mailbox. It is an effective way to improve the working efficiency of the government to extract hot topics from large-scale e-government text data, establish the correlation between topics and geographic space, and interactively explore the sources of public feedback problems. However, it is a difficult task to explore the large-scale e-government text data with traditional visualization methods such as word cloud, because too many words are hardly distributed in a limited space which will largely disturb the visual perception. In this paper, we propose a visual analytics system for large-scale e-government data exploration by means of simplified word cloud. Firstly, a representation learning model is used to embed the text data into high-dimensional space to quantitatively represent the semantic structure features of e-government text data. Then, the high-dimensional vectors are projected into a two-dimensional space where the coordinate distribution of points effectively expresses the semantic similarity of original words, which also presents geographic features that can be quantized by means of a similarity computing model. In order to simplify the understanding of large-scale e-government data and improve the cognitive efficiency of word could, we adopt the adaptive blue noise method to sample the topic words, which can simplify the visual expression of word cloud and improve the understanding efficiency of e-government data without losing the semantic structure features. Furthermore, an abstraction and visual analysis system for large-scale e-government text data is designed and implemented by integrating the above representation learning model, sampling-based abstraction model of word cloud, and topic and geographic correlation analysis model. This system provides convenient human-computer interaction modes and supports users to explore the analysis and extraction of the characteristics hidden in large-scale e-government data. It also helps government departments quickly locate the hot topics of public concern and their related regional distribution, and provides decision support to further improve the work efficiency of the government. Case studies based on real-world datasets further verify the effectiveness and practicability of our system.

KEYWORDS

E-government; Text mining; Text visualization; Visual analytics

* Corresponding Author

1 Introduction

E-government refers to the use of internet technology as a platform for exchanging information, providing services and transacting with citizens, businesses, and other arms of government (Scholta et al., 2019) . With the development and maturity of e-government mode, government departments pay more attention to public participation in e-government and provide efficient communication channels for the public to express their wishes(Shareef et al., 2012). Thus, a rich set of e-government data are produced, which represent the will of the public and are often collected in the form of text. Traditional e-government data analysis methods often extract key information from the text data in a manual way, and then calculate and predict the public's satisfaction with the government departments (Metaxas et al., 2017; Song & Meier, 2018) . However, the process of traditional exploration methods is always cumbersome and complex, which usually requires repeated loading of summary and statistical analysis, resulting in strong uncertainty of the results. Moreover, with the accumulation of data volume and the increasingly complex data structure, the limitations of traditional processing and analysis methods are more prominent.

In the field of visualization, we often utilize bag-of-word models to mine the text topics, and use word cloud to display the topic information. With the increase of data scale, e-government data mining and visualization face the following challenges: C1. Text topics are difficult to mine. E-government data is mostly in the term of short text, and the topic-mining model based on bag-of-word is difficult to accurately extract the semantic information. C2. Text data scale is large, and word cloud visualization method is not effective. Large-scale text data are presented in a limited word cloud space, with serious crowding and overlap, resulting in visual redundancy and difficulty in accurate topic exploration. C3. The geographical distribution is hard to be discovered intuitively in the word cloud. It is difficult to find the geospatial distribution of the topic in the word cloud because of the separation between them.

To tackle the above challenges, we design a visual analytics framework based on the simplified word cloud to explore the large-scale e-government data. Firstly, a representation-learning model Word2vec is employed to extract topic features from e-government short text data. The extracted features have realistic topic meaning and are helpful to understand people's resource preferences. Then, an adaptive blue noise sampling model is conducted in the word embedding space to extract keywords that can effectively express the semantics of original data, which are further utilized to generate simplified word clouds with semantic features preserved. Furthermore, the semantic similarity is calculated to establish the relation between extracted topics and geographical space, and help users to explore the spatial distribution of topics of interest. Finally, a rich set of convenient interaction modes are integrated into the visualization system, enabling users to explore e-government related topics and their spatiotemporal relationships. Case studies based on real-world datasets further verify the effectiveness and practicability of our system, which can provide decision-making basis for the work evaluation and follow-up reform and innovation of the relevant government departments. The main contributions of this paper are as follows:

(1) We design a semantic region-partitioning algorithm to recognize semantic topics in the vectorized space obtained through representation learning, by means of which more correct semantics will be extracted based on the essential characteristics of natural language.

(2) We propose a simplified word cloud generation method based on blue noise sampling

to present the semantic topics of the original e-government data, by means of which the overdrawn problem of words is tackled while the semantic topics are all preserved.

(3) The association between topics and geographic space is constructed in virtue of semantic similarity, which is calculated with semantic distance of words and visualized on the map. It really supports the visual analysis of spatiotemporal changes of semantic topics.

(4) A web-based visual analysis framework is implemented to integrate above models and visual designs, by means of which users can explore the topics and geographic correlation features of e-government data, and select the topics or regions of interest for specific analysis.

The organization of this paper is structured as follows. Section 2 discusses the related work of e-government data analysis. Section 3 introduces analysis tasks and workflow of the system. Section 4 describes the innovation and realization of the algorithms in detail. The visual analysis system and the intention of visual design are described in Section 5. Section 6 evaluates the effectiveness of our system with case studies and expert interviews, and discusses the shortcomings of the system. The last section summarizes the paper and looks forward to the future work.

2 Related Work

In this section, we review the related work, including e-government data analysis, text mining and visualization, and spatiotemporal data visualization

2.1 E–government data analysis

With the development of information technology, more people participate in the evaluation of government work through e-government platforms. They express their views and suggestions on the work of government departments, or consult their concerns to government departments, etc., forming e-government data (Linders, 2012) . These data provide good conditions for government departments to understand the hot issues of public concern. In recent years, it is through the e-government platform that government departments make public opinions play an increasingly important role in government performance evaluation (Bai, 2013), public decision-making support (Nabatchi et al., 2015) , etc. The work of government departments is more inclined to reflect the public value, thus reducing the phenomenon of government failure (Huang, 2004) and improving the governance ability of the government. However, the basis for the public opinion to play its real value in government governance is that government departments can correctly perceive and accept the opinions. Therefore, how to accurately mine the key characteristics of e-government data, and correctly perceive the main content of public opinion expression is extremely important.

Many scholars have conducted research on e-government data. For example, Stylios et al. (2010) use sentiment analysis method to extract public opinions automatically and emotions in online posts, to facilitate the government departments in the future work reference. Mayasari et al. carry out sentiment analysis on tweets based on machine learning method, and study the variation rule of public sentiment on government performance evaluation in Surabaya, Malaysia (Mayasari et al., 2020) . Baojun et al. (2013), aiming at the content analysis of public opinions in the context of smart cities, propose a methodological framework based on LDA topic model to extract potential topics that the government or policy makers may pay attention to and analyze the time series of discussion heat from large-scale opinion information text. Yi et al. (2019) adopt LDA model to mine e-government data for gover-

nance of bike-sharing policies, hoping to provide theoretical basis and decision-making suggestions for the government to make policies more scientifically. Yimin (2018) believes that whether urban planning and construction are well done or not is ultimately measured by the satisfaction of the masses. The public opinions of the draft of urban master plan can reflect the citizens' satisfaction with various areas of urban development in a specific period. They utilize text-mining technology to analyze the e-government data of Beijing urban planning. Zhengrong (2019) takes advantage of big data analysis technology to excavate the topic features of public concern, so as to facilitate government departments to understand the key information of public concern and better respond to public demand.

It can be seen from the above literature that some studies have focused on the mining and analysis of e-government data. However, there are still three deficiencies in this field. First, the number of literatures in this research field of e-government data analysis is still small, so the study of this paper has a certain contribution nonetheless. Second, related studies have not used professional visualization technology to make visual analysis of domain data, which makes these studies fail to directly reveal the hidden features of public opinion data. Third, the relevant research did not simplify large-scale data, did not relate the topic with the geographic space, and failed to adopt the depth mining and visual analysis of data features. In view of the above three shortcomings, this paper conducts interactive visual analysis of e-government data based on text mining and abstraction, which is of great significance to explore the semantic characteristics of topics and the correlation characteristics with geographic space.

2.2 Text mining and visualization

With the popularity of social network, the scale of social network text data is getting larger and larger. How to find valuable information quickly and accurately from these massive data has become a major challenge in the field of information science and technology. Text mining is a text processing technology that extracts meaningful information from unstructured data and discovers the potential value of large-scale text information (He et al., 2013).

The essence of text is natural language. One way is to construct semantic embedding space from the context of language to carry out text mining. Many previous works are based on representation learning to mine language features. For example, Hotho et al. (2003) use WordNet to convert word vectors into concept vectors, and measure the affinity between documents by calculating the similarity between concept vectors. Kim et al. (2015) present a hierarchical similarity measurement method based on search fragments on short texts to calculate the similarity between short texts. Other scholars focus on the exploration and analysis of linguistic models in textual data.BengioHolger et al. (2006) propose to learn the distributed representation of words and the probability function of word sequences simultaneously to counter the curse of dimensionality. Dauphin et al. (2017) develop a finite context recognition method through stacked convolution, which allows parallelization on sequential tags and can improve the processing efficiency of text data. Ghanbarpour and Naderi (2020) propose a language model based attribute specific ranking method, which sort candidate answers according to their semantic information until they reach the corresponding attribute level. Collins et al. (2009) produce the concept of multi-model semantic interaction, in which semantic interaction can be used to guide multiple models at multiple data scale levels to enable users to solve larger data problems.Angus et al. (2012) introduce Conceptual Recursive Graph to process text, which is a tool for drawing recursive graphs based on similarity of

concepts rather than terms. They build a part of speech model and apply the algorithm to measure the similarity between two sessions.

Clustering analysis of text data is another method of text feature mining.Beil et al. (2002) propose two text-clustering algorithms: FTC plane clustering based on frequent sets and HFTC hierarchical clustering.Yin and Wang (2014) present a folding Gibbs sampling algorithm based on Dirichlet Polynomial Mixed Model Short Text Clustering (GSDMM) to solve the problems caused by short text's sparseness, high dimension and large volume. The method achieves a good balance between the completeness and uniformity of clustering results. In order to understand the attention of bioinformatics community to different sub-fields, Janssens et al. (2007) deeply merge the text content with the structure of citation graph, and improve the unsupervised clustering performance of text based on Fisher reverse chi-square hybrid clustering method.

Text visualization refers to the process of transforming abstract data into visual graphics. By extracting, transforming and mapping eigenvalues of data, the data is finally displayed in the form of images, which is the basic technology of data visual analysis in this paper (Card et al., 1999). As the saying goes, a picture is worth a thousand words, and more than 80% of the information obtained by human beings from the outside world comes from the visual system (Lei et al., 2014). The presentation of text data in a visual and intuitive form is conducive to the analysis of the hidden information and knowledge behind the data. When confronted with massive texts, people need to browse the main contents of each text or the whole text set quickly, so it is necessary to display the text visually.

Word cloud is a commonly used text visualization method, which maps the size of words in two-dimensional space by taking the frequency of occurrence of words as the correlation measure. For instance, Wordle creates a presentation similar to the word cloud and uses a heuristic method to optimize the use efficiency of the visual area (Viegas et al., 2009) Seifert et al. (2008) introduce an algorithm for rendering compact visualization, which takes any convex polygon as the boundary to obtain higher space utilization. Wang et al. (2018) design a consistency preserving word cloud generation method, namely Edwordle, which allows users to move and edit words while preserving the neighborhood of their words. Paulovich et al. (2008) propose a kind of least square projection (LSP) to represent documents by arranging graphic marks in the visual space, and the distance of documents in the projection space reflect the content similarity. Andrews et al. (2002) propose a method of hierarchical organization of document sets to optimize the design of Voronoi diagrams and use boundary polygons to visualize document sets of specific levels in the hierarchy.

2.3 Spatiotemporal data visualization

The spatiotemporal attribute is an important feature of text data, which refers to the time attribute and the geographic attribute. Time attribute refers to the generation time of index data, while geographic attribute refers to the specific place where behaviors and events occur or belong. Visualization of data with spatiotemporal attributes is conducive to exploring data characteristics under different spatiotemporal conditions, to assist decision-making and management.

In the field of data visualization, many scholars have discussed how to conduct efficient visual analysis of spatiotemporal data. Wang et al. (2014) explore the characteristics of vehicle operation data at traffic checkpoints in Nanjing. They use dots on the map to describe the geographic location of traffic checkpoints, and design attributes such as color, number of ar-

rows and direction to represent the speed, direction and volume of traffic flow at different checkpoints. Users can intuitively analyze and discover important traffic hubs and traffic flow information in Nanjing. Pu et al. (2013) propose a visual analysis system T-Watcher. According to GPS data, the map is divided into raster points, which are clustered to form a regional view, and the color brightness represents the traffic flow. The dense area of taxi passengers can effectively present the distribution of hot spots in the city, thus helping the traffic department to monitor and analyze the complex traffic situation in big cities. Wu et al. (2016) design a visual analysis system, TELCOVIS, which can effectively analyze urban crowd movement behavior for recording the telecommunication data exchanged between mobile phones and base stations in Guangzhou. This system focuses on the behavioral characteristics of co-occurrence, and studies its feature extraction and association analysis. In addition, visual effects such as contour tree diagram and parallel coordinate diagram of geographic view are designed to help users quickly identify the common behavioral characteristics of the crowd, and provide assistance and support for relevant departments to study and analyze urban crowd activities and various kinds of derived social problems. Cao et al. (2012) make use of Twitter data to develop a visual analysis system Whisper, which is able to analyze social network public opinion effectively in real time by combining geographic information.

3 Task Analysis and System Overview

3.1 Data introduction

The data to verify the visual analysis system comes from the "Topic-Overview" section of a city's network political platform, which published 28,357 e-government opinion data for the city's three districts and four county-level governments between 2014 and 2020. The text data contains geographic attributes, which is suitable for the research objective of analyzing the correlation characteristics between public concerns and geographic topics in this paper. With this experimental data set, it provides a new perspective for the research on opinion data of government work, and a convenient interactive visual analysis way for government departments to understand the correlations between public concern topics and geographical space.

3.2 Requirements analysis

Through close communication and in-depth exploration with domain experts, we have a detailed understanding of the practical problems and interested directions of domain experts in e-government data analysis, and finally summarize four visual analysis tasks.

T1. Visualization of semantic structure representation

The government is very concerned about the topics that people in different geographical regions are most interested and how the level of concern varies. Data with geographical attributes integrate the geospatial features of the topics, which lead to the distribution of the topics covered by the data in different semantic regions. It is difficult for the classical topic mining methods to obtain accurate semantics when extracting topics. Therefore, how to characterize the text, construct semantic regions, and describe semantic correlations so that the extracted topics contain correct semantics is very important.

T2. Visualization of topic semantic features

Displaying all the keywords in the layout space visually will cause serious overlap and occlusion, which is not conducive to intuitive analysis of the meaning of the topics for users. How

to sample representative words from a large number of keywords to minimize the loss of semantic information and present the topic semantics clearly in a visual way is critical.

T3. Visualization of geographical distribution of topics

Topics that people are concerned not only share common characteristics, but also are affected by the particularity of geographical regions, which leads to geographical differences in topics. It is worth studying that how to establish the correlation between the topic and geographical space, and visually express the distribution of topics in geographic space, to show the hot issues that the government departments in different regions need to focus on intuitively.

T4. Visual analysis system for geographical distribution characteristics of topics

How to integrate text mining algorithm, visualization and interaction to design a system used by government departments for interactive visual analysis of topic and geographic correlation features of e-government data is of great significance. It can provide convenient data analysis tools for government departments, realize the one-stop transformation and analysis of data into visual interface, interactively choose interested regions or topics according to their own interests, and detect the changes in the topics of public concern in different regions.

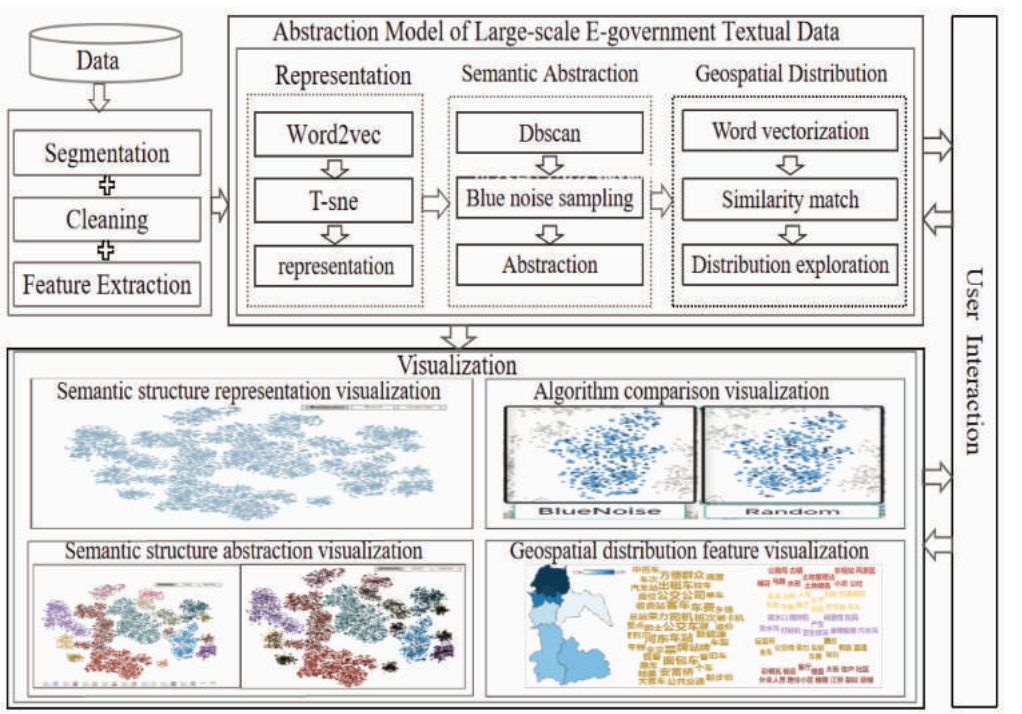


Figure 1 Flow chart of visual analysis system for large-scale e-government text data

3.3 System overview

The system flow chart of this paper is shown in Figure 1. Firstly, the data is segmented and cleaned, and the geographical features of the data are extracted. Secondly, Word2vec, a classical model of word representation learning, is used to construct the semantic space of the text data. t-SNE is employed to project the high-dimensional semantic vector into a two-dimensional plane, so as to facilitate the exploration of the semantic structure features of the

data, and prepare for mining topics by dividing semantic regions according to semantic structure. Then the blue noise sampling technique is adopted to extract the key semantic features of each topic in order to determine which topics the public discusses. After that, based on the results of representation learning, the semantic similarities between topics and geographic opinions are matched by the semantic similarity calculation, to establish their correlations. At last, by means of visualization technology, word representation learning, blue noise sampling and semantic similarity calculation methods along with convenient interaction are integrated into a visual analysis system effectively. It helps the government to explore the topics of public concern for government work and the correlations between topic and geographic space deeply.

4 E–government Data Visualization

4.1 Visualization of semantic structure representation

Large-scale e-government data targeted at different regional governments integrate the working characteristics of different functional departments and the special concerns of people in many geospatial regions, making the data characteristics complex and diverse. Compared with probability-based topic mining, it is more advantageous to extract topics from the perspective of semantic structure. In this paper, the text data is represented as word vectors, embedded into high dimensional semantic space, and the semantic similarity of words is judged from the spatial distance, to obtain the hot topic information of e-government data. The specific process is as follows:

(1) Word embedding

Word2vec is a natural language processing method that converts words into vectors through representation learning to express semantic information of text. By embedding words into a semantic space where semantically similar words are close(Shen et al., 2014), it is easy to judge the similarity between words according to geometric distance. Compared with the classical methods of learning text representation through bag-of-word model, Word2vec model takes full account of contextual semantic information and produces a higher learning quality(Tang et al., 2015). Therefore, in this paper we use Word2vec to represent text data and construct semantic space to reveal semantic structure of words. Each word is defined as a vector composed of word and word ID, and the corpus is generated by a series of words, as shown in Equation 1.

$$D=(w_1, w_2, w_3, ..., w_N) \quad (1)$$

where N is the count of words and D is the corpus generated by Word2vec training. In the process of word representation learning, the setting of text window size has an important influence on the result. Formula 2 is used in this paper to optimize the learning result of the model.

$$\frac{1}{T}\sum_{i=k}^{T-k} \log p(w_t | w_{t-k}, ..., w_{t+k}) \quad (2)$$

where T is the size of text window, and the corresponding context words of a given word is represented by $p(w_t | w_{t-k}..., w_{t+k})$.

(2) Dimensionality reduction

Words are converted to high-dimensional vectors by means of Word2vec model. However, it is difficult to visually explore the semantic structure of data and calculate semantic distance in a high-dimensional vector space with hundreds of dimensions. Therefore, to solve

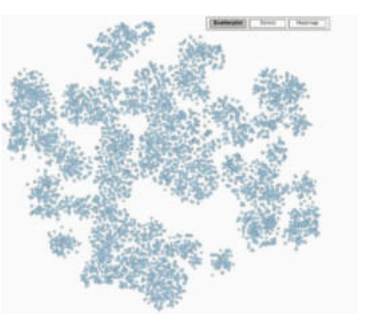
the problem of visual occlusion and computation difficulty of high-dimensional vectors, it is necessary to project high-dimensional vectors into two-dimensional space. t-SNE (t-distributed Stochastic Neighbor Embedding) (van der Maaten & Hinton, 2008) is an effective way for dimensionality reduction of high-dimensional vectors. It is able to capture both local and global features of data, and effectively retain the original features of data during dimension compression (Wattenberg et al., 2016). Due to the good performance in dimension reduction, t-SNE is used to project the representation learning results of Word2vec, to better reveal the semantic similarity of Word2vec in two-dimensional semantic space through the connection and closeness of words (Xia et al., 2018; Zhao et al., 2019).

(3) Semantic region division of topics

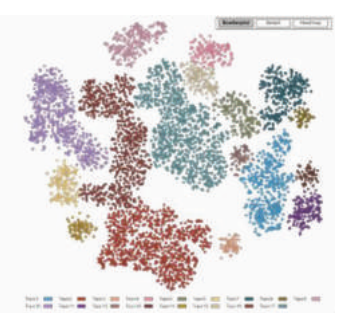
In order to further extract semantic features from the semantic space and obtain semantic categories, so as to extract topics of government work opinion data, in this paper we use DBSCAN (van der Maaten & Hinton, 2008) density clustering algorithm to classify words in the semantic space into meaningful categories according to their densities. The core of DBSCAN clustering is the density of clustering objects, which defines the cluster as the maximum set of density connection points. It can divide regions with high enough density into clusters and treat data points with low-density values as outliers. Since it is impossible to predict the number of topics concerned by the public, it is obviously not feasible to control the learning process by applying supervision conditions, such as setting independent variables and the number of target clusters. Compared with supervised clustering methods such as K-means (Hartigan & Wong, 1979) and GMM (Ebeida et al., 2014), DBSCAN has good performance under unsupervised conditions, so we choose DBSCAN clustering method to mine topics.

(4) Structure representation

In order to capture the representation results more conveniently and verify the effectiveness of the proposed method, we further design the representation projection view to explore the representation results quickly. Figure 2(a) shows the results of public opinion representation. Each data point in the figure represents a word, and the closer the distance between data points is, the more semantically similar the words are. It can be seen that the compactness of semantic structure is different among different semantic sections. Figure 2(b) displays the topics that people are concerned, which are represented by different colors.



(a) Word2vec semantic structure representation



(b) Topic of DBSCAN for clustering

Figure 2 Semantic structure representation and clustering

4.2 Visualization of topic features

(1) Blue noise sampling

Based on large-scale corpus, if all the words in a topic are arranged in the word cloud,

they will overlap and block each other in the limited screen space, resulting in a lot of visual confusion. Thus, it is impossible to perceive the semantics of the topic of interest clearly, which affects the effective analysis of data. Therefore, sampling representative words from large-scale data is an effective method to solve the serious occlusion of large-scale data layout in limited space. In order to simplify the words in original topic and maintain the semantic features, so that the sampled words can represent the original semantics of the topic to the maximum extent, we use the blue noise sampling algorithm based on Poisson disc to perform adaptive sampling for each topic. Blue noise sampling is a commonly used sampling algorithm in the field of graphics, which simultaneously satisfies the randomness and uniformity of the distribution of sampling point sets and can maximize the retention of the original semantics of data. It has a wide application in point cloud sampling, texture rendering, geometric processing and other aspects (Yan et al., 2015). In the sampling process, an active point is selected randomly as the center by throwing dart, and the radius is set by the sampling rate to generate the sampling disk. The generated Poisson disk must meet the minimum distance characteristic, that is, only if the distance between the centers of any two Poisson disks is greater than the sampling radius, the generated sampling point is valid. If the generated disk is inconsistent with the previous one, it will be rejected (Godwin et al., 2017). To maintain the semantic structure of the original data, we use kernel density estimation to evaluate the semantic structure density of the samples, with the formula as follows:

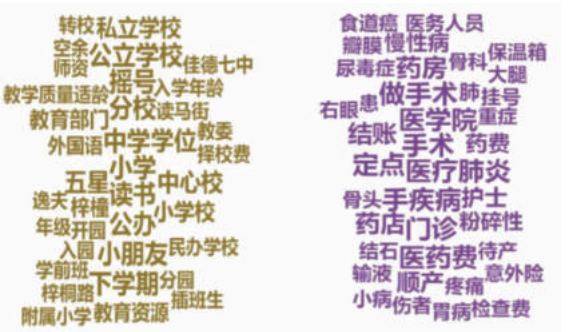
$$f(\mathrm{p}) = \sum_{i}^{n} k_h (p - p_j) \qquad (3)$$

where P = $\{p_1, p_2, \ldots, p_m\}$ is the coordinate position of a sequence of words, kh represents the Gaussian kernel function, h represents the bandwidth used to control the smoothness of the constructed density domain, and n is the total number of points in the local region. R= r_a/f (p) is used to obtain the sampling radius, where ra is the sampling rate, which is the number of subject words that the analyst needs to display.

In addition, the comparative experiment between random sampling and blue noise sampling algorithm is added to further prove the superiority of blue noise sampling in the sampling of topic words. The basic principle of random sampling is to select one sampling data point randomly at a time, and the algorithm will automatically traverse all the data until the total number of sampling data points meets the preset sampling rate requirements. Random sampling is also one of the commonly used sampling methods in data abstraction.

(2) Word cloud

As mentioned above, topics with different semantics are obtained based on representation learning, and then expressed by sampled words through blue noise sampling. In order to display the semantic features of each topic visually, we use word cloud to show the topic keywords that people are concerned. Word cloud is a very convenient and effective text visualization technology. Words displayed in the word cloud are the topic words obtained by blue noise sampling. The size of the word represents the occurrence probability of the word in the document. The greater the occurrence probability, the larger the size of the word in the word cloud. Figure 3 shows the word cloud view of topic 8 and topic 11, in which we can directly see that topic 8 focuses on education, while topic 11 focuses on medical care. Both of them have clear semantics.



(a)Topic 8 (b)Topic 11

Figure 3 Word cloud of topic

Moreover, we also design a multi-topic semantically preserved word cloud to display the top five hot topic words of each region simultaneously. Firstly, the layout space of the word cloud is divided, and the same topic words are placed in the same area. Secondly, according to the attribution of words to a topic, the keywords of different topics are rendered with the corresponding topic colors in the representation space.

4.3 Visualization of geographical distribution features of topics

(1) Exploration of geographical distribution features of topics

In order to further explore the distribution characteristics of topics in geographical space, we introduce the concept of semantic similarity calculation to analyze the correlation between topics and geographical space. Semantic similarity calculation (Palangi et al., 2014) refers to a method to calculate semantic association of text by calculating semantic distance between two texts to carry out similarity matching. Semantic similarity computing has been widely used in intelligent search and matching, machine translation, etc. Drawing on the concept of semantic similarity computing, we match the topic with the semantics of public opinion in different geographical regions to obtain their associations. The vector coordinates of the words in the topic in the two-dimensional semantic space are defined as:

$$L_i = ((x_1, y_1)(x_2, y_2)\dots(x_n, y_n)) \quad (4)$$

where n is the number of words in the topic corpus, (x_i, y_i) is the vector coordinate of word i in semantic space, and L_i represents the coordinate set of word vector. The vector of words in the semantic space of the geographic regional corpus is defined as:

$$L_j = ((x_1, y_1)(x_2, y_2)\dots(x_m, y_m)) \quad (5)$$

where m is the number of words in the geographic corpus, (x_j, y_j) is the vector coordinate of word j in semantic space, and L_j represents the coordinate set of word vector. Considering the difference of word vectors in semantic direction, we use cosine distance to match the text, and the formula is shown in 6:

$$D = \frac{\vec{L_i} \cdot \vec{L_j}}{||L_i|| \cdot ||L_j||} = \frac{\sum_i^n (x_i, y_i) \cdot \sum_j^m (x_i, y_i)}{\sqrt{\sum_i^n (x_i^2 + y_i^2)} \cdot \sqrt{\sum_j^m (x_j^2 + y_j^2)}} \quad (6)$$

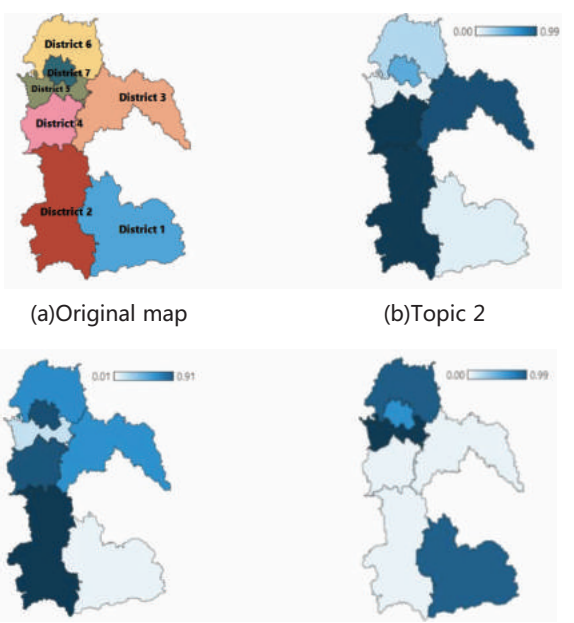
In our system, the semantic similarity between topic corpus and geographic corpus is used to represent the association between topic and geographic information. The more similar the semantics is, the greater the degree of correlation is, indicating that the attention of that

topic in the geographic area is higher.

(2) Visualization of geographic distribution features of topics

This section designs the visualization scheme of the geographic distribution features of topics, including the visualization of the geographic distribution features of a certain topic and the visualization of the hot topics of a certain region. Firstly, a visual display of the distribution of attention heat in different regions is designed for a certain topic. We use map view to facilitate government departments to analyze the distribution characteristics of the topic in different districts and counties. Figure 5(a) is a map of the administrative divisions of a city, including 7 districts. Different colors are used to fill different districts according to how much attention each county pays to the topic. The darker the color is, the higher the attention of the topic in a certain district. In Figure 5, (b) shows topic 2, which deals with the management and planning of land and resources in rural construction; (c) displays topic 3, which is about the renovation and demolition of dilapidated houses; (d) is topic 15, which is related to construction, contract signing and tax payment. From the distribution differences of different topics on the map, it can be seen that topic 2 has a higher attention heat in districts 2, 3 and 4; topic 3 has the highest attention heat in districts 2, 4 and7; and topic 15 has a higher attention heat in districts 1, 5, and 6.

Secondly, in order to support the government departments to further select specific regions of interest and analyze the hot topics that people are concerned about in a certain region, we further design the visualization of hot topics in a region. By clicking on a region in Figure 4(e), the corresponding top 5 hot topics in that region will be highlighted in Figure 4 (b), and the topic word cloud in Figure 4(c) will be replaced with the corresponding topic word cloud for a region of concern.



(a)Original map (b)Topic 2

(c)Topic 3 (d)Topic 15

Figure 4 Geographic distribution features of topics

4.4 Visual analysis system

In order to facilitate government departments to understand the differences in topics of

public concern in different regions and adjust the direction of work more accurately, a visual analysis system of topic and geographic correlation characteristics is designed to help government departments conduct convenient interactive visual analysis of public opinions. The interface of the system is shown in Figure 5. Its main views include: (a) Control panel, which helps users adjust and control relevant parameters. (b) Representation visualization of semantic structure of opinions, which is used to show the semantic structure of public opinions and the topic of concern. (c) Topic semantic feature cloud visualization, describing the semantic features of the topic obtained based on Word2vec, DBSCAN and blue noise sampling, and showing the semantic features of the top five hot topics in each district. (d) Data overview window, showing topics obtained after multiple processing and the number of words on topics. (e) Visualization of the geographic distribution characteristics of topics, which describes the distribution characteristics of each topic in different districts and counties. (f) Opinions display window for displaying the corresponding original opinions of different districts and counties, and the topic words contained in the original opinions.

In order to facilitate the government departments to choose the corresponding topics or regions according to their own interests for analysis, the system provides a large number of convenient man-machine interaction window linkage operation. Users can quickly analyze and explore the hot topics of public concern and the geographical spatial distribution characteristics of the topics. The text mining algorithm encapsulates the required features in the back-end database after mining. By loading data for visual display in the front-end system, government departments can conduct visual analysis of data according to their own interests.

Combined with the algorithm function and according to the research objectives of this paper, the interaction design of the system is as follows. After the system loads the data, first, when clicking the Word2Vec button in Figure 5(a), the system displays the Word2Vec semantic structure representation projection diagram in Figure 5(b). If the DBSCAN button is clicked, Figure 5(b) shows 17 topics, each represented by a cluster of different colors. Secondly, click on any one in the topic bar chart in Figure 5(d) to highlight the position of the topic in Figure 5(b) and the sampled words of the topic synchronically. All other topics are diluted. At the same time, the word cloud of this topic is shown in Figure 5(c), and the heat distribution of this topic is interactively shown in Figure 5(e) as well. Finally, when clicking on a district in Figure 5(e), the top 5 hot topics and sampled words of this region are highlighted in Figure 5(b), and word clouds of the top 5 hot topics of this region are shown in Figure 5(c). The original opinion data of the district and county are displayed in the opinion display window in Figure 5(f), and the topic words of each district and county are highlighted in the opinion display, too. Interaction design will be demonstrated in case studies more realistically.

5 Evaluation

In this paper, we adopt React, Python, D3.js and other technologies to implement a Web-based visual analysis system. It supports users to explore topic and geographic associations for large-scale e-government text data. The effectiveness and usefulness of the algorithm and visual analysis system are verified through a series of cases on real data sets.

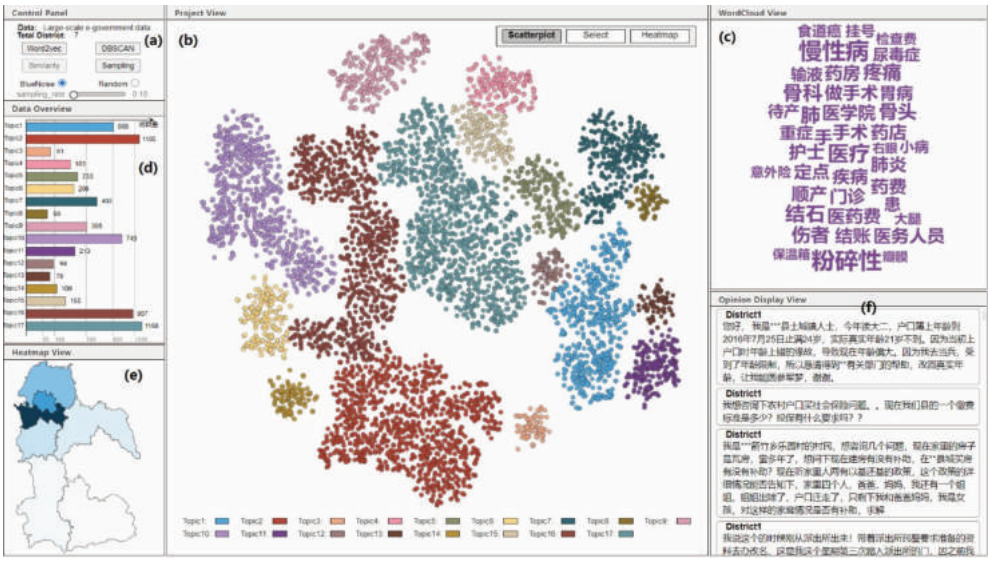


Figure 5 Visual interface of visual analysis system

5.1 Case Study

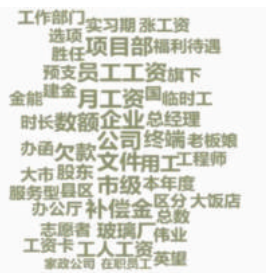
In the process of specific case study, we invite users with experience and needs of large-scale e-government text data analysis to use the designed system. The data analysis process and feedback of users are summarized, recorded and analyzed from the visual analysis perspectives of topic representation structure and semantic features, algorithm comparison and geographic spatial distribution features of topics.

Case1 Visual analysis of topic representation structure and semantic features

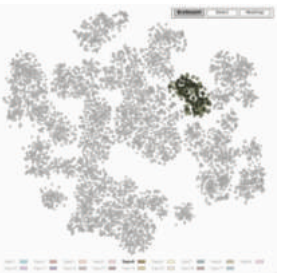
This section makes a visual analysis of the representation structure and semantic features of the topics of public concern through specific case data. Users select topics 5, 6, 9 and 14, whose word cloud diagrams and the corresponding semantic structure representation views are shown in each sub-graph in Figure 6. The highlighted cluster in Fig. 6(b) is the semantic structure representation view of topic 5 in semantic space, in which the highlights with the cross symbol are the sampled data points that correspond to the words in Fig. 6(a). These sampled points are evenly distributed in the topic cluster. It can be seen that the representative words of topic 5 in the word cloud keep the original semantics of topic 5 well, without serious loss of original semantics due to sampling. Further analysis of the topic semantic representation structure diagrams of topics 6, 9 and 14 selected by users, namely 6 (d), 6(f) and 6(h), shows that the distribution of sampling highlights is also relatively uniform. Therefore, it can be inferred that blue noise sampling can keep the semantic information of the original topic to the maximum extent, and optimize the spatial distribution of the topic words in the word cloud to avoid the problem of incomplete perception of the semantic information of the topic caused by visual clutter and occlusion.

Using the four topics selected by users, the semantic features expressed by the topics are further analyzed visually, to judge which topics the public concerns. As shown in Figure 6(a), the representative topic words are salary increase, employee salary, in-service employee welfare, treatment, internship, enterprise, company, etc. It can be seen that topic 5 is about wages, labor relations, labor security and other aspects that people pay close attention. The government should formulate more comprehensive labor law and related regulations, and

strictly enforce the law to ensure the benefits of workers. In Figure 6(c), the topic words are urban roads, vehicles, tickets, street vendors, no parking, traffic flow, roadblocks, streetlights, speed bumps, etc. It can be seen that topic 6 is about urban management, which reminds the city to improve in the management of traffic and street vendors.

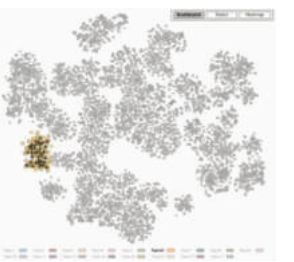


(a)Topic 5

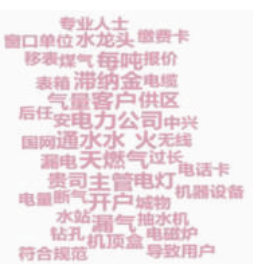


(b)Semantic structure representation for topic 5

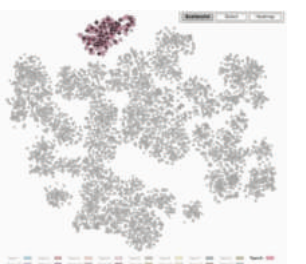
(c)Topic 6



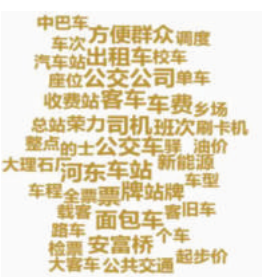
(d)Semantic structure representation for topic 6



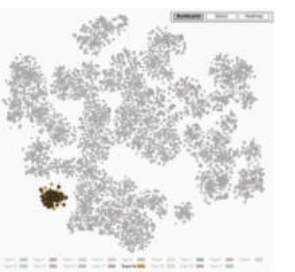
(e)Topic 9



(f)Semantic structure representation for topic 9



(g)Topic 14



(h)Semantic structure representation for topic 14

Figure 6 Visual analysis of topic representation structure and semantic features

In Figure 6 (e), the subject words are electric quantity, natural gas, power company, leakage, meter box, faucet, etc. It can be seen that topic 9 is about water supply, power supply and gas supply. People have a strong demand for basic conveniences of life, and

relevant departments should listen to public opinions, do a good job in providing services and adjust prices reasonably. In Figure 6(g), the topic words are bus, bus company, driver, public transportation, train number, station, ticket, toll booth, etc., and users can easily perceive that topic 14 is about transportation. This shows that the public's demand for convenient transportation is very strong, and the government departments should actively provide more public transportation services.

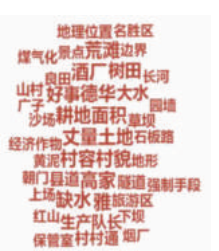
Case2 Visual analysis of algorithm comparison

This section further verifies the validity of blue noise sampling applied to semantic feature preservation of the topic by comparing it with random sampling algorithm. As shown in Figure 7, we selected three topics to compare the differences between blue noise sampling algorithm and random sampling algorithm in semantic feature preservation of topics.

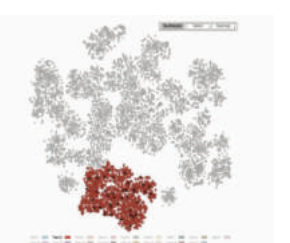
Figure 7(a) is the word cloud of topic 2, and the red highlighted part in figure 7(b) is the position of topic 2 in the semantic space, where the data points marked by crosses are the words sampled by the blue noise sampling algorithm, corresponding to the topic words in the word cloud. The red highlighted part in Figure 7(c) is also the position of topic 2 in the semantic space, where the data points marked by the cross are the words sampled by the random sampling algorithm. According to the distribution of sampled data points in Fig. 7(b) and Fig. 7(c), it can be perceived that the blue noise sampled data points have a relatively uniform distribution in the semantic space of the topic. While the results of random sampling show an irregular distribution, with either too many local sampling points or too few local sampling points. In Figure 7(c), the three unsampled local regions are further framed. That is to say, the semantics of these three local regions will be missing when the random sampling algorithm is used to sample representative topic words, resulting in incomplete semantics when the semantic features of the topic are perceived in the word cloud.

Next, for topic 7, whose word cloud is shown in 7(d), from which we can know that topic 7 is about teaching qualifications and campus life. From Figure 7(e) and 7(f), it can be seen that the distribution of sampling points of highlighted topics is still relatively uniform in the case of blue noise sampling, and relatively uneven in the case of random sampling whose data points in two local semantic regions are not sampled.

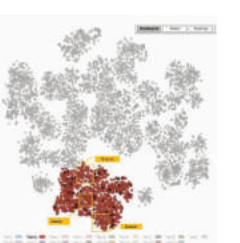
Finally, through the analysis of topic 10, it can be seen that the blue noise sampling points in Fig. 7(h) present a relatively uniform distribution. However, the random sampling points in 7(i) are not evenly distributed. In the upper left and lower right parts of the purple highlighted cluster, there is a large semantic area that does not involve any sampling points, respectively. It can be seen that the result of random sampling leads to serious loss of semantic information.



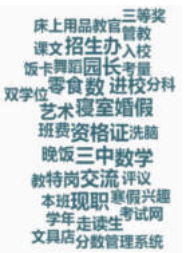
(a)Topic 2



(b)Distribution of blue noise sampling for topic2



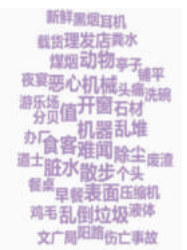
(c)Distribution of random sampling for topic2



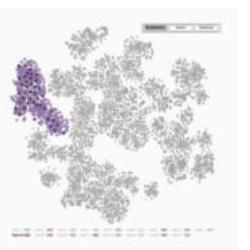
(d)Topic 7

(e)Distribution of blue noise sampling for topic7

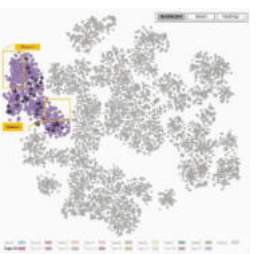
(f)Distribution of random sampling for topic7



(g)Topic 10



(h)Distribution of blue noise sampling for topic10



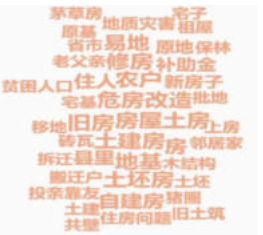
(i)Distribution of random sampling for topic10

Figure 7 Comparison between blue noise sampling and other sampling algorithms

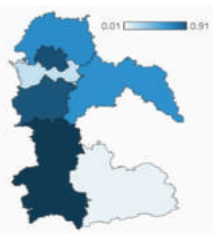
Case3 Visual analysis of geographic distribution characteristics of topics

This section further provides the visual design of the geographical spatial distribution features of topics, to facilitate the government departments to perceive the geographical spatial distribution of the topics of public concern visually. The specific visual design includes selecting a topic of interest, exploring the distribution characteristics of the topic in the geographic space and selecting a region of interest to explore the top 5 topics most relevant to the region.

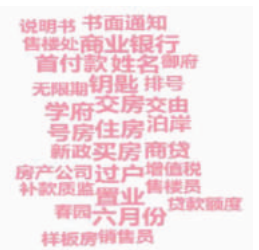
As shown in Figure 8, the geographic distribution characteristics of the topics are discussed through the user's selection of 4 topics. First, from the word cloud in Figure 8(a), it can be seen that topic 3 is about related to the renovation of dilapidated houses and the demolition of houses, which attracts the highest attention in districts 2,4 and 7. Therefore, people in these three districts pay more attention to the topic related to the reconstruction and demolition of houses than those in others. Then, the user clicks topic 4, and it can be seen from Fig. 8(c) that topic 4 focuses on issues related to house purchase, mortgage and deed tax. Fig. 8(d) shows that the topic has the highest attention in districts 5 and 16, and relatively low attention in districts 2, 3 and 4. When the user chooses topic 13, it can be found from Fig. 8(e) that the semantics of topic 13 is about social security, and as shown in Fig. 8(f), this topic has the highest attention in districts 6 and 7. Finally, the user chooses topic 15, and it can be found from the word cloud in Figure 8(g) that topic 15 is related to project construction, project tax payment and project contract signing. Topic 15 attracts more attention in districts 2 and 1 than other counties, as shown in Figure 8(h).



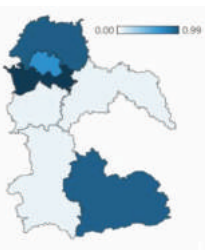
(a)Topic 3



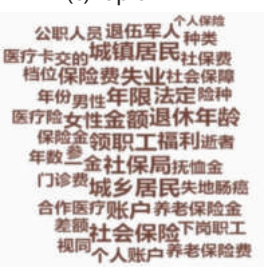
(b)Geographic distribution features for topic 3



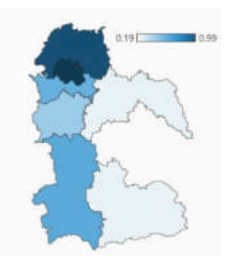
(c)Topic 4



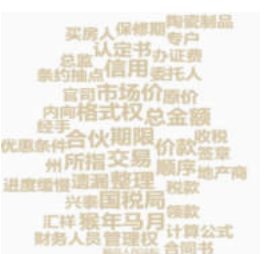
(d)Geographic distribution features for topic 4



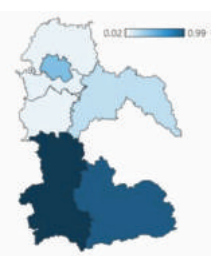
(e)Topic 13



(f)Geographic distribution features for topic 13



(g)Topic 15

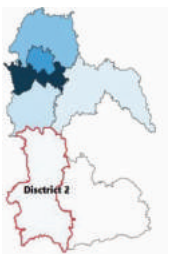


(h)Geographic distribution features for topic 15

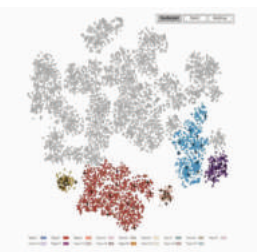
Figure 8 Visual analysis of geographic distribution features for topics

Furthermore, visual analysis of hot topics in a certain region is provided. By selecting a district or county, users can visually see the top 5 topics of concern in the region. As can be seen from Figure 9(a)-(c), the top five topics that district 2 pay attention to are Topic 1: household registration, family planning and marriage registration, etc., Topic 2: rural construction and land management, Topic 3: housing demolition and reconstruction, etc., Topic 11: medical treatment, and Topic 14: public transportation. When the user selects district 3 in Figure 9(d), Figure 9(e) shows the top five topics of concern for the district, and Figure 9(f) synchronization shows the corresponding word cloud. Users can perceive that the top five topics of concern in district 3 include topic 2: rural construction and land resource management, topic 6: city traffic management, topic 10: the urban environment and pollution con-

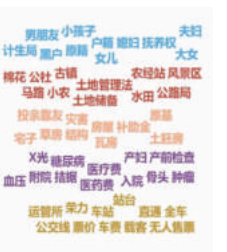
trol, topic 14: public transportation, and topic 16: community management. The user further clicks district 4 in Figure 9(g). Figure 9(h) shows the top 5 topics most relevant to district 4, and Figure 9(i) shows the word cloud of the 5 topics. The users can easily perceive the five most relevant topics of district 4 as topic 2: rural construction and land resource management, topic 3: housing demolition and reconstruction, etc., topic 6: urban traffic management, topic 14: public transportation, and topic 16: community management.



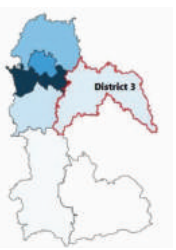
(a)District 2



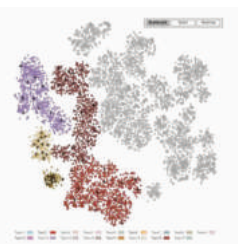
(b)Top 5 hot topics for district 2



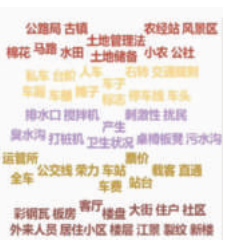
(c) Word cloud of top 5 hot topics for district 2



(d)District 3

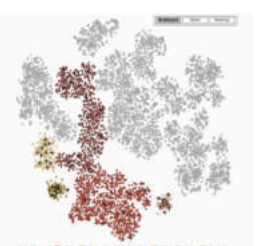


(e)Top 5 hot topics for district 2

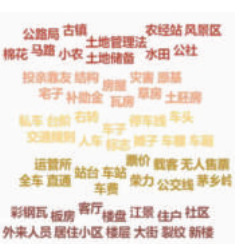


(f)Word cloud of top 5 hot topics for district 2

(g)District 4



(h)Top 5 hot topics for district 2



(i)Word cloud of top 5 hot topics for district 2

Figure 9 Visual analysis of hot topics

From the analysis of the above figure, it can be seen that although there are some differences in the hot topics concerned by different districts, there are also some common points. For example, from the point of view of traffic development, the three counties are more concerned about traffic problems. In combination with the reality, these three districts and counties are relatively deficient in economic and transportation development. People rely heavily on public transportation for travel, so they may pay more attention to the development of transportation and road construction. In addition, the three hot topics of regional discussion are all related to rural construction, land resource management and crop cultiva-

tion. Combined with the actual situation, we found that these three areas belong to the suburban counties, which have natural advantages in the development of crop planting and animal husbandry. Moreover, compared with the southern mountains, these three areas belong to the plain terrain and the terrain is relatively gentle. Therefore, when addressing livelihood issues, government departments should not only pay attention to the topics of common concern, but also take measures according to local conditions and plan public service and resource allocation policies according to the actual problems in each region.

5.2 Discussion

In this paper, the method of large-scale e-government text data exploration is to use the word representation learning method to obtain the semantic structure features of large-scale text data, and then use the density clustering method to explore the semantic region in the representation space to obtain the topic of government work opinion data. Then the topic words are obtained based on blue noise sampling. This method can divide topics based on semantic structure density, and effectively retain semantic information of original data while reducing visual clutter of large-scale network, so that users can better perceive topics. In addition, the association between topic and geography is constructed based on semantic similarity calculation, which draws on the method of calculating semantic similarity between two text datasets in the field of natural language processing. This method can well measure the degree of association between topic and geographic information. However, there are still some problems in this paper that have not been well solved and need to be further studied.

Firstly, by embedding words into vectorized space, geometric distance can be used to effectively represent semantic similarity of words. Nevertheless, due to the randomness of word representation learning and the approximation of t-SNE dimension reduction, some errors will inevitably occur. In the future work, this study will try to design a better topic mining model, so as to mine the topics of public concern more accurately and reduce the errors caused by the randomness of the model. Secondly, in addition to the deviation caused by representation learning and dimensionality reduction, blue noise sampling will also lead to the loss of original information. In future research, we will try our best to design a better algorithm to optimize blue noise sampling, or avoid using this method, but still optimize the layout of the subject word in the visual space. Finally, when visual analysis of the topic and geographic association features is conducted, due to the limitation of data acquisition, only the data of different districts and counties in a certain city is used. Therefore, in the research on the association between people's concern topics and geographical space, the difference analysis of people's concern topics in different regions is only limited to the comparative analysis of small geographical space areas. Even though different districts and counties have differences in economic conditions, terrain conditions, political status, social culture and other aspects, which makes the focus of people in different regions will be different to some extent. But we hope that future work will be able to take data from more places over a larger geographic area, so that we can study whether people's concerns are more markedly different in places that are geographically distant. This will also have a greater reference value for the government's more macro policy control.

6 Conclusion

In this paper, we explore the abstraction of large-scale e-government text data, and design

a visual analysis system to analyze large-scale text data in a convenient visual way. Firstly, the word representation learning method is used to embed the text data into the high dimensional vector semantic space. The two-dimensional semantic space is constructed by dimensionality reduction. According to the semantic structure of words, the density clustering algorithm based on the semantic structure is selected to divide the semantic region of the constructed two-dimensional plane, so as to mine the topic of public concern. Then the blue noise sampling algorithm is used to mine the representative topic words. After that, the association between topic and geographical space is established based on semantic similarity calculation. Secondly, we integrate the algorithms t0 design the flexible interactive visual analysis system of large-scale e-government text data. Finally, the validity of the algorithm and the system is evaluated through the real data set to further verify the practical value of the system.

Acknowledgements

This work was supported by the National Natural Science Foundation of China(No.61872314, No.61802339), the Natural Science Foundation of Zhejiang Province (No.LY18F020024), the Humanities and Social Sciences Foundation of Ministry of Education in China (No. 18YJC910017), the Major Humanities and Social Sciences Research Project in Zhejiang Province (2018QN021), and the Open Project Program of the State Key Lab of CAD&CG of Zhejiang University(No.A2001).

References

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., & Tochtermann, K. (2002). The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. *Information Visualization, 1*, 166–181. https://doi.org/10.1057/palgrave.ivs.9500023

Angus, D., Smith, A., & Wiles, J.(2012). Conceptual Recurrence Plots: Revealing Patterns in Human Discourse. *IEEE Transactions on Visualization and Computer Graphics, 18* (6), 988–997. https://doi.org/10.1109/tvcg.2011.100

Bai, W.(2013). A Public Value Based Framework for Evaluating the Performance of e-Government in China. *iBusiness, 5*(3), 26–29.

Baojun, M., Nan, Z., & Tao, S.(2013). Big data analysis of public feedback in the context of smart cities: the perspective of probabilistic topic modeling. *E-Government*(12), 9–15.

Beil, F., Ester, M., & Xu, X.(2002). Frequent Term-Based Text Clustering. *KDD ´02 :Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 436 –442. https://doi.org/10.1145/775047.775110

BengioHolger, Y., SchwenkJean-Sébastien, SenécalFréderic, & Gauvain, M.-L.(2006). A Neural Probabilistic Language Models. In D. E. Holmes & L. C. Jain (Eds.), *Innovations in Machine Learning. Studies in Fuzziness and Soft Computing*(Vol. 194, pp. 137–186). Springer.

Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., & Qu, H.(2012). Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. *IEEE Transactions on Visualization and Computer Graphics,18* (12), 2649–2658. https://doi.org/10.1109/tvcg.2012.291

Card, S., Mackinlay, J., & Shneiderman, B.(1999). *Readings in Information Visualization: Using Vision To Think.* Morgan Kaufmann Publishers.

Collins, C., Carpendale, S., & Penn, G. (2009). DocuBurst: Visualizing Document Content using Language Structure. *Computer Graphics Forum, 28*(3), 1039–1046. https://doi.org/10.1111/j.1467-8659.2009.01439.x

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D.(2017). *Language Modeling with Gated Convolutional Networks* Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning

Research. http://proceedings.mlr.press

Ebeida, M., Mitchell, S., Awad, M., Park, C., Swiler, L., Manocha, D., & Wei, L.–Y.(2014). Spoke Darts for Efficient High Dimensional Blue Noise Sampling. *ACM Transactions on Graphics, 37*. https://doi.org/10.1145/3194657

Ghanbarpour, A., & Naderi, H.(2020). An Attribute–Specific Ranking Method Based on Language Models for Keyword Search over Graphs. *Ieee Transactions on Knowledge and Data Engineering, 32*(1), 12–25. https://doi.org/10.1109/tkde.2018.2879863

Godwin, A., Wang, Y., & Stasko, J. T.(2017). TypoTweet Maps: Characterizing Urban Areas through Typographic Social Media Visualization.

Hartigan, J. A., & Wong, M. A.(1979). A K–Means Clustering Algorithm. *Journal of the Royal Statistical Society: Series C(Applied Statistics), 28*(1), 100–108. https://doi.org/https://doi.org/10.2307/2346830

He, W., Zha, S., & Li, L.(2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33* (3), 464–472. https://doi.org/10.1016/j.ijinfomgt.2013.01.001

Hotho, A., Staab, S., & Stumme, G.(2003, NOV 19–22, 2003). *Ontologies improve text document clustering* Third Ieee International Conference on Data Mining, Proceedings, MELBOURNE, FL. <Go to ISI>://WOS:000188999400077

Huang, D.(2004). *History of Western Administrative Doctrine(Revised Edition)*. Wuhan University Press.

Janssens, F., Glanzel, W., & De Moor, B.(2007, AUG 12–15, 2007). *Dynamic Hybrid Clustering of Bioinformatics by Incorporating Text Mining and Citation Analysis* Kdd–2007 Proceedings of the Thirteenth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, San Jose, CA <Go to ISI>://WOS:000266628300037

Kim, Y., Han, J., Yuan, C., & Assoc Comp, M.(2015, AUG 10–13, 2015). *TOPTRAC: Topical Trajectory Pattern Mining* 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Univ Technol Sydney, Adv Analyt Inst, Sydney, AUSTRALIA. <Go to ISI>://WOS:000485312900063

Lei, R., Yi, D., Shuai, M., Xiao–Long, Z., & Guo–Zhong, D.(2014). Visual Analytics Towards Big Data. *Journal of Software, 25*(9), 1909–1936.

Linders, D.(2012). From e–government to we–government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly, 29* (4), 446–454. https://doi.org/10.1016/j.giq.2012.06.003

Mayasari, R., Fithriasari, K., Iriawan, N., & Winahju, W.(2020). Surabaya Government Performance Evaluation Using Tweet Analysis. *MATEMATIKA*, 36, 31–42. https://doi.org/10.11113/matematika.v36.n1.1176

Metaxas, T., Makaratzi, E., & Terzidis, K.(2017). Improving service quality to local communities via a citizen satisfaction measurement in Greece: The ´MUSA´ approach. *The Journal of Developing Areas*, 51 (3), 77–101.

Nabatchi, T., Becker, J. A., & Leighninger, M.(2015). Using Public Participation to Enhance Citizen Voice and Promote Accountability. In J. L. Perry & R. Christensen(Eds.), *Handbook of Public Administration*(3rd ed., pp. 137–151). Wiley.

Palangi, H., Deng, L., Shen, Y., Gao, J., He, X., Chen, J., Song, X., & Ward, R.(2014). Semantic Modelling with Long–Short–Term Memory for Information Retrieval.

Paulovich, F. V., Nonato, L. G., Minghim, R., & Levkowitz, H.(2008). Least square projection: A fast high–precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics, 14*(3), 564–575. https://doi.org/10.1109/tvcg.2007.70443

Pu, J., Liu, S., Ding, Y., Qu, H., Ni, L., & Ieee.(2013). *T–Watcher: A New Visual Analytic System for Effective Traffic Surveillance.* https://doi.org/10.1109/mdm.2013.23

Scholta, H., Mertens, W., Kowalkiewicz, M., & Becker, J.(2019). From one–stop shop to no–stop shop: An e–government stage model. *Government Information Quarterly, 36* (1), 11–26. https://doi.org/10.1016/j.giq.2018.11.010

Seifert, C., Kump, B., Kienreich, W., Granitzer, G., & Granitzer, M.(2008). On the beauty and usability of tag clouds. In E. Banissi, L. Stuart, M. Jern, G. Andrienko, F. T. Marchese, N. Memon, R. Alhajj, T. G. Wyeld, R.

A. Burkhard, G. Grinstein, D. Groth, A. Ursyn, C. Maple, A. Faiola, & B. Craft(Eds.), *Proceedings of the 12th International Information Visualisation*(pp. 17–+). https://doi.org/10.1109/iv.2008.89

Shareef, S. M., Jahankhani, H., & Dastbaz, M.(2012). E–Government Stage Model: Based On Citizen–Centric Approach in Regional Government in Developing Countries. *International Journal of Electronic Commerce Studies, 3*(1), 145–164.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G.(2014). A Latent Semantic Model with Convolutional–Pooling Structure for Information Retrieval. *CIKM 2014 – Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, 101–110. https://doi.org/10.1145/2661829.2661935

Song, M., & Meier, K. J.(2018). Citizen Satisfaction and the Kaleidoscope of Government Performance: How Multiple Stakeholders See Government Performance. *Journal of Public Administration Research and Theory, 28*(4), 489–505. https://doi.org/10.1093/jopart/muy006

Stylios, G., Christodoulakis, D., Besharat, J., Vonitsanou, M.–A., Kotrotsos, I., Koumpouri, A., & Stamou, S.(2010). Public Opinion Mining for Governmental Decisions. *Electronic Journal of e–Government, 8* (2), 202–213.

Tang, D., Qin, B., & Liu, T.(2015). Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews–Data Mining and Knowledge Discovery, 5* (6), 292–303. https://doi.org/10.1002/widm.1171

van der Maaten, L., & Hinton, G.(2008). Visualizing Data using t–SNE. *Journal of Machine Learning Research, 9*, 2579–2605. <Go to ISI>://WOS:000262637600007

Viegas, F. B., Wattenberg, M., & Feinberg, J.(2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics, 15*(6), 1137–1144. https://doi.org/10.1109/tvcg.2009.171

Wang, Y., Chu, X., Bao, C., Zhu, L., Deussen, O., Chen, B., & Sedlmair, M.(2018). EdWordle: Consistency–preserving Word Cloud Editing. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 647–656. https://doi.org/10.1109/tvcg.2017.2745859

Wang, Z., Ye, T., Lu, M., Yuan, X., Qu, H., Yuan, J., & Wu, Q.(2014). Visual Exploration of Sparse Traffic Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1813–1822. https://doi.org/10.1109/tvcg.2014.2346746

Wattenberg, M., Viégas, F., & Johnson, I.(2016). How to Use t–SNE Effectively. *Distill, 1*. https://doi.org/10.23915/distill.00002

Wu, W., Xu, J., Zeng, H., Zheng, Y., Qu, H., Ni, B., Yuan, M., & Ni, L. M.(2016). TelCoVis: Visual Exploration of Co–occurrence in Urban Human Mobility Based on Telco Data. *IEEE Transactions on Visualization and Computer Graphics, 22*(1), 935–944. https://doi.org/10.1109/tvcg.2015.2467194

Xia, J., Ye, F., Chen, W., Wang, Y., Chen, W., Ma, Y., & Tung, A. K. H.(2018). LDSScanner: Exploratory Analysis of Low–Dimensional Structures in High–Dimensional Datasets. *IEEE Transactions on Visualization and Computer Graphics, 24*(1), 236–245. https://doi.org/10.1109/tvcg.2017.2744098

Yan, D.–M., Guo, J.–W., Wang, B., Zhang, X.–P., & Wonka, P.(2015). A Survey of Blue–Noise Sampling and Its Applications. *Journal of Computer Science and Technology, 30* (3), 439–452. https://doi.org/10.1007/s11390–015–1535–0

Yi, Y., Yi, Z., Mei, L., & Wen, D.(2019). Research on the adoption of public feedback opinions based on the LDA model: a comparative analysis of the revision of shared bicycle policy and data mining. *Information Science, 37*(1), 86–93.

Yimin, W.(2018). Comparative analysis of the public opinions of the two editions of Beijing´s master plan based on text mining. *Beijing Planning and Construction, 2*(1), 87–94.

Yin, J., & Wang, J.(2014). A Dirichlet multinomial mixture model–based approach for short text clustering. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2623330.2623715

Zhao, Y., Luo, F., Chen, M., Wang, Y., Xia, J., Zhou, F., Wang, Y., Chen, Y., & Chen, W.(2019). Evaluating Multi–Dimensional Visualizations for Understanding Fuzzy Clusters. *IEEE Transactions on Visualization and Computer Graphics, 25*(1), 12–21. https://doi.org/10.1109/tvcg.2018.2865020

Zhengrong, W.(2019). *Research on the effectiveness of online political–civilian interaction based on public perception* Lanzhou University]. Lanzhou.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Spatial Clustering and Epidemiological Trends of Hand, Foot and Mouth Disease in Mainland China, 2009–2015

Jinguo Xin[a], Chen Yang[b]

a. Center of Information and Economy Social Development, Hangzhou Dianzi University, Hangzhou, China;
b. Hangzhou Vocational & Technical College, Hangzhou, China

ABSTRACT

HFMD can be caused by a variety of enteroviruses, including Coxsackievirus A16 and enterovirus 71. There are no effective therapeutic measures to cure HFMD at present. So, this study aimed to analyze the spatial relativity and the local accumulation type based on the theory of spatial analysis and the spatial autocorrelation analysis module of ArcGIS and GeoDa. We found that there was a seasonal trend in HFMD. The lowest incidence appeared in February, and the peak of the reported incidence was occurred during the period from May to June. However, in most cases, another peak appeared from September to November. The trend of incidence was related to age, too. The overall trend of the reported incidence was a U-shape in north-south orientation and exposed an inverted U-shape in east-west. The correlation between the spatial distribution of HFMD was positive. Hunan, Guangxi and Guangdong were the hot areas, while the cold spots were Jilin, Inner Mongolia, Xinjiang, Gansu and Qinghai.

KEYWORDS

HFMD; Spatial clustering; Epidemiological trends; China

1 Introduction

HFMD is an intestinal disease caused by a variety of enteroviruses. Infants and children are more susceptible to the HFMD (Cardosa et al., 1999; Chen et al., 2004; Kobayashi et al., 2013; Wu et al., 2010). In infected people, enteroviruses cause severe illness characterized by fever, painful mouth sores, herpes and so on. Generally, there are three modes of transmission patterns: spray, air and direct contact with a patient (Sun et al., 2016). Signs of HFMD include outbreak on a large scale in a short time, high infectiousness and mortality (Xiao et al., 2016). It also has a profound negative effect on the physical and mental development of children. With increasing age, the symptoms will be milder and less.

HFMD has been categorized as a Class C legal infectious disease on May 2, 2008. From January to June in 2018, the case load of HFMD in China was 1 004 056, of which 22 cases resulted in death. Studies showed that HFMD occurred throughout the mainland of China. The incidence rate of HFMD was from 37.01/100,000 to 205.06/100,000. In recent years, the reported mortality was between 6.46/100,000 and 51.00/100,000(National Health and Family Planning Commission of PRC,2018). Therefore, identifying the clustering areas and the period of HFMD outbreak provided evidence and scientific base for prevention and control measures.

2 Material and Methods

The Data of HFMD: The data of patients and incidence data in 31 provinces in the mainland of China (excluding Taiwan, Hong Kong and Macau) from 2009-2015 was collected from The Data-center of China Public Health Science (CPC). The database collected all data since the country started to implement network report of infectious disease in 2008.

Administrative vector maps of 31 provinces in the mainland of China were downloaded from Diva GIS (http://www.diva-gis.org/Data). We can match the epidemiological data of HFMD with the national administrative vector map by using the province name as the matching variable in the GIS10.3.

Descriptive analysis: The data of HFMD cases were collected by month and year. We draw the tendency chart of the incidence and mortality of HFMD reported from 2009 to 2015. The graphic location was aggregated to analyze the statistical characteristic of HFMD, too.

Global trend analysis: In this paper, we draw the spatial distribution chart of incidence of HFMD reported. This approach can well stimulate the development trend of incidence data.

Global spatial autocorrelation analysis: This paper used the Moran's I index to determine whether there was a global spatial autocorrelation relationship in 31 provinces. The value of the Global Moran's I is between -1 and 1; if I is more than 0 and P is less than 0.05, the study variables in adjacent regions are clustered; if Global Moran's I is equal or close to 0, there is no autocorrelation (Getis & Ord, 2010; Liu et al., 2018), meaning that the data is random. If Global Moran's I is less than 0 and P is less than 0.05, indicating that the research variables in adjacent regions have different aggregation. It means that the data is discrete. The Global Moran's I coefficient formula is:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where n means the number of study space regions. The W_{ij} is the weight coefficient of the i and j region, reflecting the spatial relationship between the i and j region. If i area is contiguous with j region, W_{ij} =1; otherwise W_{ij} =0. x_i and x_j are the research variables(eg, morbidity) in the i and j region, and $\bar{x}$ is the mean of the study variables (eg, average morbidity). Generally, the coefficient test of Global Moran's I is based on the standardized Moran's I or z value, which test whether a property of each area relevant or not. The formula is as follows:

$$Z(I) = \frac{I - E(I)}{\sqrt{V(I)}}$$

$$E(I) = \frac{-1}{n-1}$$

$$V(I) = E(I^2) - [E(I)]^2$$

Where E(I) and V(I) represent the expected value and the sample variance, respectively. the probability of spatial autocorrelation is 90% when 1.96 >|Z(I)|>1.65; If the value of |Z(I)| is less than 2.58 and greater than 1.96, the probability of spatial autocorrelation is 95%; when the value of |Z(I)|is more than 2.58, the probability of spatial autocorrelation is 99%.

Local spatial autocorrelation analysis: The local spatial autocorrelation analysis is per-

formed by the Local Moran's I, and the formula is as follows:

$$I = \frac{n(x_i - \bar{x})\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

If I>0, it indicates that the area has similar spatial aggregation with the adjacent area; if the value of I<0, it means that the spatial aggregation of the area is not similar to that of the adjacent area. In this paper, Moran scatter plot was used to analyze the incidence of HFMD in mainland of China (Anselin, 1995; Ord & Getis, 2010).

Moran scatter plot analysis: The Moran scatter plot describes the correlation between the variables and spatial lagged variable by observing the value of the variables in each position (Chen, 2017; Ord & Getis, 2010). In this paper, we used the GeoDa to export the Moran scatter plot of the reported incidence of HFMD in main land of China from 2009 to 2015 (Li & Chen, 2008; Tang et al., 2014). And the evolutionary path of the reported incidence of HFMD in provinces was summarized.

3 Results

Overall Distribution: As we can see from Figure1, the reported incidence and mortality of HFMD fluctuated slightly in each year. It reached its peak in 2012. Although the reported incidence of HFMD in 2014 was higher than that in 2013, it was still not higher than that in 2012. At the same time, the reported mortality of HFMD in 2009-2015 showed the same tendency.

The monthly time series of the case load of HFMD in mainland of China showed a significant seasonal variation (Figure 1). There was a distinct seasonal tendency. The number of registered HFMD cases increased to maximum at the end of spring and the beginning of summer in each year, while decreased in the following summer. The case load was low in autumn and winter.

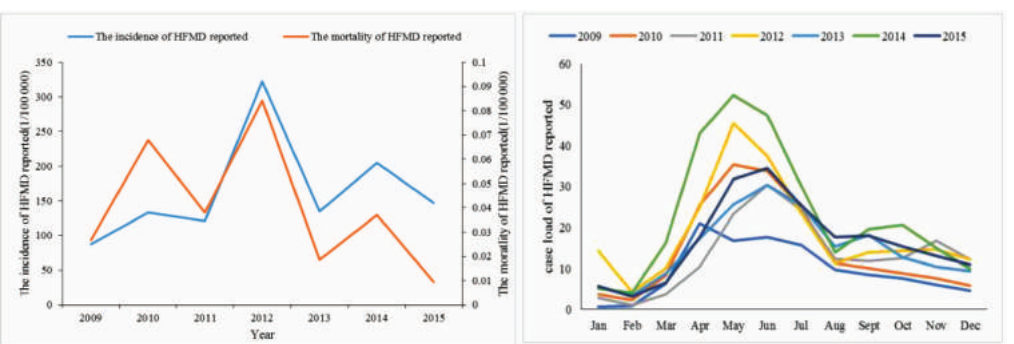


Figure 1

Figure 2 showed the thematic maps of the reported incidence of HFMD in various provinces (excluding Taiwan, Hong Kong and Macau). The reported incidence of HFMD was lowest in Xinjiang, Tibet, Qinghai, Gansu, Sichuan and Heilongjiang. The reported incidence of HFMD in Guangxi, Hainan, Guangdong, Fujian, Zhejiang, Shanghai and Beijing were at a high level. The overall reported incidence in 2009-2015 presented an increasing tendency. The reported incidence in the eastern region was higher than that in the western region, and it in the south was higher than that in the north. Besides, the incidence in the coastal region was higher than that in the inland region, too.

Global spatial trend: Spatial distribution trend of every year from 2009 to 2015 was found that the incidence of HFMD showed a trend of U-shape in the north-south direction. Furthermore, there was a trend of inverted U-shape in the east-west direction (Figure 3).

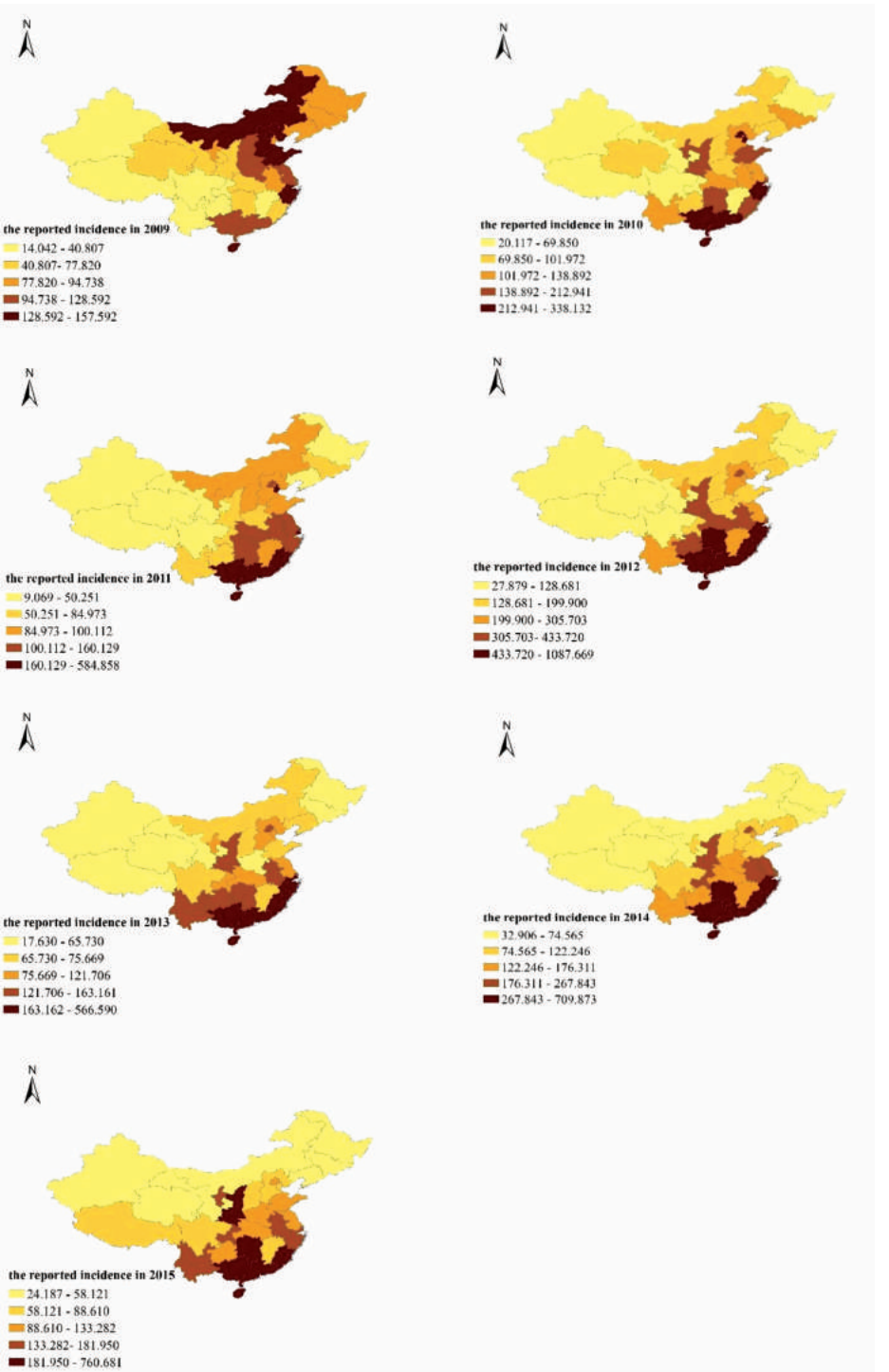


Figure 2

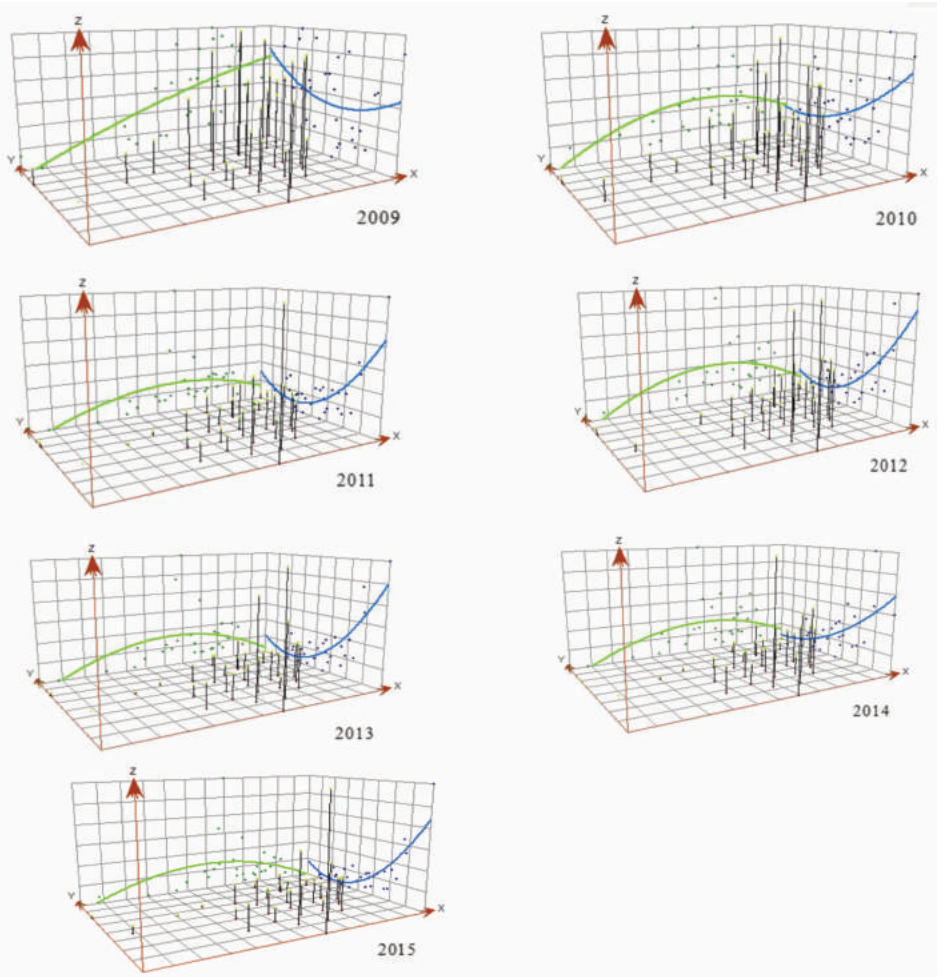


Figure 3

Spatial Autocorrelation Analysis of Global Moran I Index: There was significant global spatial autocorrelation in the mainland of China from 2009 to 2015, and its range was between 0.170 and 0.366 (Table 1). The z-tests were all above the critical value of 1.96, which meant the probability of spatial autocorrelation was higher than 95%. Therefore, further research must be done into the spatio-temporal clustering analysis of HFMD.

Table 1

Year	Moran´s I	Variance	P	Z
2009	0.364895	0.007018	0.000002	4.753568
2010	0.170577	0.006726	0.012904	2.486417
2011	0.238774	0.004520	0.000052	4.047568
2012	0.365107	0.006060	0.000000	5.118443
2013	0.326776	0.005570	0.000001	4.824962
2014	0.358232	0.005757	0.000000	5.160453
2015	0.289265	0.004494	0.000001	4.812061

Moran scatter plot: Table 2 showed the evolution of the corresponding quadrant of the Moran scatter plots for the incidence of HFMD.

Table 2

Regions	Province	2009	2010	2011	2012	2013	2014	2015
Northeast China	Liaoning	1(H–H)	3	3	3	3	3	3(L–L)
	Jilin	1	3(L–L)	3	3(L–L)	3(L–L)	3(L–L)	3(L–L)
	Heilongjiang	1	3	3	3(L–L)	3	3	3(L–L)
East China	Shandong	1(H–H)	4	3	3	3	3	3
	Jiangxi	3	2	2	2(L–H)	2	2(L–H)	2
	Fujian	2	1	1	1	1	1	1
	Anhui	1	4	4	4	4	1	4
	Shanghai	1	1	1	1	1	1	1
	Jiangsu	1(H–H)	2	1	2	2	1	2
	Zhejiang	1	1	1	1	1	1	1
North China	Neimenggu	1	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)
	Hebei	1(H–H)	1	3	3	3	3	3(L–L)
	Shanxi	1(H–H)	3	3	3	3	3	3
	Tianjing	1(H–H)	1	1	3	3	3	3
	Beijin	1	1	1	3	4	4	3
the Central of China	Hunan	3	1	1	1	1	1(H–H)	1
	Hubei	3	3	3	4	3	2	2
	Henan	1	3	3	3	3	3	3
Southern China	Hainan	1	1	1	1	1	1	1
	Guangxi	4	1	1	1	1	1(H–H)	1
	Guangdong	4	1	1	1(H–H)	1(H–H)	1(H–H)	1
Southwest of China	Chongqing	3(L–L)	3	3	2	3	2	4
	Guizhou	3(L–L)	2	2	1	2	2(L–H)	2
	Sichuang	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3	3	3
	Yunnan	3(L–L)	2	3	2	2	2	1
	Tibet	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3	3	3
the Northwest-ern District of China	Xinjiang	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)
	Gansu	3	3(L–L)	3(L–L)	3(L–L)	3	3(L–L)	3
	Qinhai	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)	3(L–L)
	Ningxia	1	4	3	3	3	3	3
	Shaanxi	3	4(H–L)	3(L–L)	3(L–L)	4	4	4

In the past seven years, Zhejiang, Hainan, Beijing, Shanghai, Fujian, Hunan, Guangxi and Guangdong had been in the first quadrant (HH) for a long time or most of the time, and Guangdong was often significant. Some provinces (including autonomous regions and municipalities) were in the third quadrant (LL) at long-term, including Tianjin, Shanxi, Shandong, Ningxia, Inner Mongolia, Liaoning, Jilin, Henan, Heilongjiang, Hebei, Tibet, Xinjiang, Sichuan, Qinghai, Hubei, Gansu and Chongqing. The spatial distribution of incidence in Inner Mongolia, Jilin, Tibet, Xinjiang, Sichuan, Qinghai and Gansu were

statistically significant. Jiangsu, Yunnan, Jiangxi and Guizhou were in the second quadrant (LH) over a long period of time. Besides, the spatial distribution of Jiangxi and Guizhou were significant in 2012 and 2014 respectively. The provinces, including Anhui and Shaanxi were in the fourth quadrant (HL) for quite some time, but only Shaanxi had statistical significance in 2010.

4 Discussion

Spatial statistical analysis is the statistical analysis for spatial data. Recognizing the spatial dependence, spatial correlation or spatial autocorrelation of the data related to geographical location is the core of research. It establishes the statistical relationship between data through spatial location. Compared with traditional statistical analysis, the relationship between spatial data of spatial statistical analysis is not independent. The data has some correlation in space and different degrees of correlation at different spatial resolutions.

It discusses the role of geospatial information in the epidemic of infectious diseases by using the spatial statistics and geospatial information. Spatial statistical analysis will help to further understand the epidemic law further. It can also provide useful information for prevention and control work (Han et al., 2018). Therefore, providing the data on the spatial and temporal aggregation of HFMD to policy makers can help them develop precautionary and control measures for HFMD.

The results suggested that the prevalence of HFMD may have a seasonal trend, which displayed the high-risk month of HFMD in every year from 2009 to 2015. The monthly reported lowest incidence of HFMD was in February each year, and then it began to rise sharply. The peak of the reported incidence of HFMD was concentrated between May to June. After that, it began to decline gradually. The feature was similar to other related literature (Ni et al., 2012; Zhu et al., 2011). Another reported peak appeared between September to November. But the value of peak was still far below the number of previous peaks. According to the reported incidence of each region, the provinces with high reported incidence in the past seven years include Guangxi, Hainan, Guangdong, Fujian, Zhejiang, Shanghai and Beijing, which may be related to large population density and the high turnover of people. Moreover, the climate may be a factor that generated the high rate of reported cases of HFMD in these areas, too. The reported incidence of HFMD in Xinjiang, Tibet, Qinghai, Gansu, Sichuan and Heilongjiang had been at a low level. In addition, the reported incidence continued to rise, and the highest annual incidence had been broken every year. Across the mainland of China, on the one hand, the incidence in the south was higher than that in the north; on the other hand, the incidence in east was higher than that in the west.

Autocorrelation analysis of HFMD outbreaks in 2009-2015 showed that Moran's I statistics were positive, and there was a positive correlation in spatial distribution of HFMD. Among them, the spatial correlation of 2014 was the strongest in all ages. On the side, hotspots were chiefly located in Hunan, Guangxi and Guangdong, the cold spot areas were mainly distributed in Jilin, Inner Mongolia, Xinjiang, Gansu and Qinghai. In terms of national data, it was also important to be in the first quadrant for a long time (Han et al., 2018). The incidence of HFMD in 2009-2015 had evolved over the past seven years, forming a high-high accumulation area dominated by Zhejiang, Hainan, Beijing, Shanghai, Fujian, Hunan, Guangxi and Guangdong. Low-low clusters regions had been formulated out with the Inner Mongolia,

Jilin, Tibet, Xinjiang, Sichuan, Qinghai and Gansu as primary areas, too. It indicated that similar natural environments played a very important role in the transmission of HFMD, such as climate, sunshine, and rainfall (Chang et al., 2012; Han et al., 2018; Zhao et al., 2016).

In this paper, GeoDa was used to map the Moran scatter plots in local autocorrelation analysis (Entrikin et al., 2010; Zhang & Zhang, 2007a; Zhang & Zhang, 2007b). However, this method had certain limitations. If the region had no land connection with other regions, the weight of the region cannot be calculated when calculated of the spatial weight matrix. However, Hainan Province had no land connection with other provinces. But, the reported incidence of HFMD in Hainan Province had been at a high level across the country, so it was still included in the calculation. This paper only considered the spatial aggregation of HFMD. In addition, the research failed to analyze other factors affecting HFMD (such as climate, economic development level, gender, etc.).

Acknowledgments

This work was supported by the National Natural Social Science Found of China (Grant Nos. 17AJY008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that they have no competing interests.

References

Anselin, L. (1995) . Local indicators of spatial association–LISA. *Geogr Anal, 27* (2) , 93–115.

Cardosa, M. J., Krishnan, S., Tio, P. H., Perera, D., & Wong, S. (1999) . Isolation of subgenus B adenovirus during a fatal outbreak of enterovirus 71 –associated hand, foot, and mouth disease in Sibu, Sarawak. *Lancet, 354* (9183) , 987–991.

Chang, H., Chio, C., Su, H., Liao, C., Lin, C., Shau, W., Chi, Y., Cheng, Y., Chou, Y., Li, C., Chen, K., & Chen, K. (2012) . The association between enterovirus 71 infections and meteorological parameters in Taiwan. *PLoS One*, 7 (10) , e46845.

Chang, L., Tsao, K., Hsia, S., Shih, S., Huang, C., Chan, W., Hsu, K., Fang, T., Huang, Y., & Lin, T. (2004) . Transmission and clinical features of enterovirus 71 infections in household contacts in Taiwan. *JAMA Netw Open, 291* (2) , 222.

Chen, C., Lin, H., Li, X., Lang, L., Xiao, X., Ding, P., He, P., Zhang, Y., Wang, M., & Liu, Q. (2013) . Short–term effects of meteorological factors on children hand, foot and mouth disease in Guangzhou, China. *Int J Biometeorol, 58* (7) , 1605–1614.

Chen, S., Qin, L., Du, Z., Jin, Y., Du, J., Chen, Y., Watanabe, C., & Umezaki, M. (2017) . Spatial clustering of severe hand–foot–mouth disease cases on Hainan Island, China. *Jpn J Infect Dis, 70* (6) , 604–608.

Entrikin, J. N., Gale, S., & Olsson, G. (2010) . Philosophy in geography. *N Z Geog, 37* (1) , 40–41.

Getis, A., & Ord, J. K. (2010) . The analysis of spatial association by use of distance statistics. *Geogr Anal, 24* (3) , 189–206.

Han, T., Guo, Y., Xu, W., & Wang, Y. (2018) . Spatial clustering of hand, foot and mouth disease in mainland China from 2008 to 2017. *Bing Du Xue Bao, 34* (5) , 534–542.

Kobayashi, M., Makino, T., Hanaoka, N., Shimizu, H., Enomoto, M., Okabe, N., Kanou, K., Konagaya, M., Oishi, K., & Fujimoto, T. (2013) . Clinical manifestations of coxsackievirus A6 infection associated with a major outbreak of hand, foot, and mouth disease in Japan. *Jpn J Infect Dis, 66* (3) , 260–261.

Li, X., & Chen, K. (2008) . Scan statistic theory and its application in spatial epidemiology. *Zhonghua Liu Xing*

Bing Xue Za Zhi, 29 (8) , 828–831.

Liu, M., Li, Q., Zhang, Y., Ma, Y., Liu, Y., Feng, W., Hou, C., Amsalu, E., Li, X., & Wang, W. (2018) . Spatial and temporal clustering analysis of tuberculosis in the mainland of China at the prefecture level, 2005–2015. *Infect Dis Poverty, 7*, 106.

National Health and Family Planning Commission of PRC. (2018) . Guidelines for the diagnosis and treatment of hand, foot and mouth disease (2018 Edition) . *Chinese Practical Journal of Rural Doctor,* (6) .

Ni, H., Yi, B., Yin, J., Fang, T., He, T., Du, Y., Wang, J., Zhang, H., Xie, L., Ding, Y., Gu, W., Zhang, S., Han, Y., Dong, H., Su, T., Xu, G., & Cao, G. (2012) . Epidemiological and etiological characteristics of hand, foot, and mouth disease in Ningbo, China, 2008–2011. *J Clin Virol, 54* (4) , 342–348.

Ord, J. K., & Getis, A. (2010) . Local spatial autocorrelation statistics: Distributional issues and an application. *Geogr Anal, 27* (4) , 286–306.

Sun, L., Lin, H., Lin, J., He, J., Deng, A., Kang, M., Zeng, H., Ma, W., & Zhang, Y. (2016) . Evaluating the transmission routes of hand, foot, and mouth disease in Guangdong, *China. Am J Infect Control, 44* (2) , 13–14.

Tang, X., Zeng, Q., Zhao, H., Juan, Y., Qin, L., Xiao, D., Xia, Y., Yang, R., & Fang, M. (2014) . Spatial clustering and influential factors of hand–foot–mouth disease (HFMD) in Chongqing, China,2008–2012. *Chinese Journal of Zoonoses, 30* (12) , 1196–1200+1205.

Wu, Y., Yeo, A., Phoon, M. C., Tan, E. L., Poh, C. L., Quak, S. H., & Chow, V. T. K. (2010) . The largest outbreak of hand; foot and mouth disease in Singapore in 2008: The role of enterovirus 71 and coxsackievirus A strains. *Int J Infect Dis, 14* (12) , 1076– 1081.

Xiao, X., Liao, Q., Kenward, M. G., Zheng, Y., Huang, J., Yin, F., Yu, H., & Li, X. (2016) . Comparisons between mild and severe cases of hand, foot and mouth disease in temporal trends: A comparative time series study from mainland China. *BMC Public Health, 16* (1) , 1109.

Zhang, S., & Zhang, K. (2007a) . Contrast study on moran and Getis–Ord indexes of local spatial autocorrelation indices. *Geod Geodyn, 27* (3) , 31–34.

Zhang, S., & Zhang, K. (2007b) . Comparison between general moran´s index and Getis–Ord general G of spatial autocorrelation. Zhongshan Da Xue Xue Bao Zi Ran Ke Xue Ban, 46 (4) , 93–97.

Zhao, D., Wang, L., Cheng, J., Xu, J., Xu, Z., Xie, M., Yang, H., Li, K., Wen, L., Wang, X., Zhang, H., Wang, S., & Su, H. (2016) . Impact of weather factors on hand, foot and mouth disease, and its role in short–term incidence trend forecast in Huainan city, Anhui province. *Int J Biometeorol, 61* (3) , 1–9.

Zhu, Q., Hao, Y., & Yu, S. (2011) . Epidemiological characteristics and space–time analysis of hand–foot–and–mouth disease in Guangdong province from 2008 to 2010. *Xian dai Yu fang Yi xue, 38* (10) , 1824–1826+1831.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Ontology–based Indexing Technologies in Information Retrieval: Building a Topic Map （ISO 13250) for a Mathematics Education Database

Fei Shu

Chinese Academy of Science and Education Evaluation,Hangzhou Dianzi University,Hangzhou,China

ABSTRACT
This paper describes a project that has created a Topic Map search tool for a mathematics educational database containing articles from the journal For the Learning of Mathematics. The resulting website enables users to retrieve research articles based on a variety of topics such as mathematics classification, research methods, educational objectives, in addition to traditional bibliographic information.

KEYWORDS
Ontology; Information Retrieval; Topic Map; Database

1 Problem

Keyword-based search (e.g., Google) is the most popular tool for web searching and information retrieval; however, keyword-based searching also produces many irrelevant results because keywords can have multiple meanings or they often inadequately express users' intent (Michael & Rajesh, 2011). In order to find accurate information, users waste time browsing long lists of often-irrelevant results. In addition, keyword-based search does not provide adequate support to users but strictly returns the documents whose vocabulary matches the keyword terms (Wilson et al., 2009); therefore, it fails to recognize relevant documents that do not match the query (Mann, 2008).

2 Ontology and Topic Maps

Some previous research tries to improve information searching by optimizing information organization. Ontology-based indexing technologies area promising approach (Patkar, 2011). By establishing an ontology on the basis of semantic relationships between concepts, the improved information organization may improve users' searching performance (Jimeno-Yepes et al., 2010; Yi, 2008).

Building ontologies can be done using Topic Maps, which is an international standard (ISO13250) for knowledge representation and exchange. According to Pepper (2000), Topic Maps represent information concepts and their relationships using the following elements:

- Topics: represent any concept, from people, countries, and organizations to software modules, individual files, and events;
- Associations: representing relationships between topics;
- Occurrences: representing information resources relevant to a particular topic.

Compared with the keyword searching, a Topic Map aims to reduce the gap between desired results (i.e., more relevant information) and the large result sets containing much irrelevant information returned by traditional keyword search tools (Kwong & NG, 2003; Yi, 2008). A topic map represents knowledge through topics, associations between topics, and occurrences, which are locators pointing to information resources related to the given topics (Pepper, 2000). A topic map can facilitate information discovery and retrieval because it groups results by topics that users can visually navigate without inputting keywords (Garshol, 2004). Topic Map users can retrieve information associated with a topic and discover information associated with previously unknown but related topics (Melgar, 2011). Assuming a topic map application that is integrated with a keyword search engine, a student can filter the returned articles by topic, choose a topic, navigate the topic map, access other related topics via associations, and retrieve information without conducting additional searches.

This paper describes the design of a Topic Map search tool for a mathematics educational database containing articles from the journal For the Learning of Mathematics. The project aims to enable novice and expert users to retrieve scholarly articles based on a variety of topics such as mathematics classification, research methods, and educational objectives, in addition to traditional bibliographic information.

3 Previous Works

Topic Maps are used as an ontology-based search tool in various domains. Based on measures of recall and search time, an experimental study using 40 participants preforming information retrieval tasks showed that a Topic Maps-based ontology information retrieval system had a significant and positive effect on both recall and search time (Yi, 2008). Topic Maps have been used to generate learning materials for Chinese herb medication (Shin et al., 2007), and they have been applied to language teaching (Urbaniak & Venkatesh, 2012) and aboriginal language preservation (Pelczer et al., 2012) to represent grammatical and task-based structures.

Topic Maps are also used in E-learning applications. For example, they support the location of appropriate learning resources for a specific student in a given context to ensure effective learning (Kolas, 2006), and Dicheva and Dichev (2006) describe the TM4L environment that enables the creation, maintenance and use of ontology-aware online learning repositories based on the Topic Map standard. Widhalm and Mueck (2010) suggest that Topic Maps may be used instead of SCORM/LOM for the characterization of E-learning resources, and Lalingkar and Ramani (2010) developed a Study Facilitation System (SFS) based on Topic Maps.

4 Methodology

This project requires a domain appropriate subject structure (i.e., topics and their relationships), the assignment of these topics to a collection of articles from the domain, and the development and implementation of a live Web site that allows browsing and searching of the collection based on the subject structure. The pilot test was also conducted by interviewing 5 users regarding their experience and satisfaction using the novel Topic Map search tool.

4.1 Creating a Subject Structure

Traditional academic databases (e.g. ERIC) allow users to search and group results by subject heading or descriptor. However, these subject headings and descriptors are pre-desig-

nated based on the articles' content, meaning that users cannot modify them according to their specific needs (e.g. neither the subject "Discrete Mathematics" nor the subject "Combinatorics" exists in ERIC). In this case, users would like to group the searching results not only by the subject discussed in the content but also by metadata not currently available such as the research method or the educational objective applied by the studies. Therefore, a new subject structure was created based on users' requirement. The Topic Map standard was chosen as an appropriate tool for the creation of this subject structure.

Users' needs and requirements were collected by consulting subject-matter experts and conducting a needs assessment. The Mathematics Subject Classification (MSC) 2010 served as the basis for the subject structure topic design, which was modified according to users' requests. For example, new topics were defined (e.g., Academic Level, Research Method, Educational Objective) as well as their associations. Table 1 shows examples of topics, association and occurrence types created for the collection:

Table 1 List of Topics, Associations and Occurrences

Topic	
Article	Academic Level
Author	Pre–School
Topic (Mathematics Classification)	Elementary
Discrete Mathematics	Secondary
Applied Mathematics	College
Mathematics Theory & Philosophy	University
History of Mathematics	General
Pure Mathematics	Article Type
Algebra	Opinions & Discussion
Arithmetic	Research Paper
Calculus	Peer Review
Combinatorics	Project Report
Function	Literature Review
Geometry	Educational Objective
Infinity	Teaching
Mathematics Analysis	Curriculum
Numbers	Pedagogy
Topology	Educator
Trigonometry	Learning
Research Methods	Learning Objective
Quantitative Research	Learning Environment
Qualitative Research	Student
Mixed Methods Design	History of Education
Other Methods	Instruction
Association	**Occurrence**
Write/Written by	Volume No.
Discuss/Discussed by	Issue No.
Applied to	Publish Time
Related to	Page
Classified as	Content
Use/Used by	

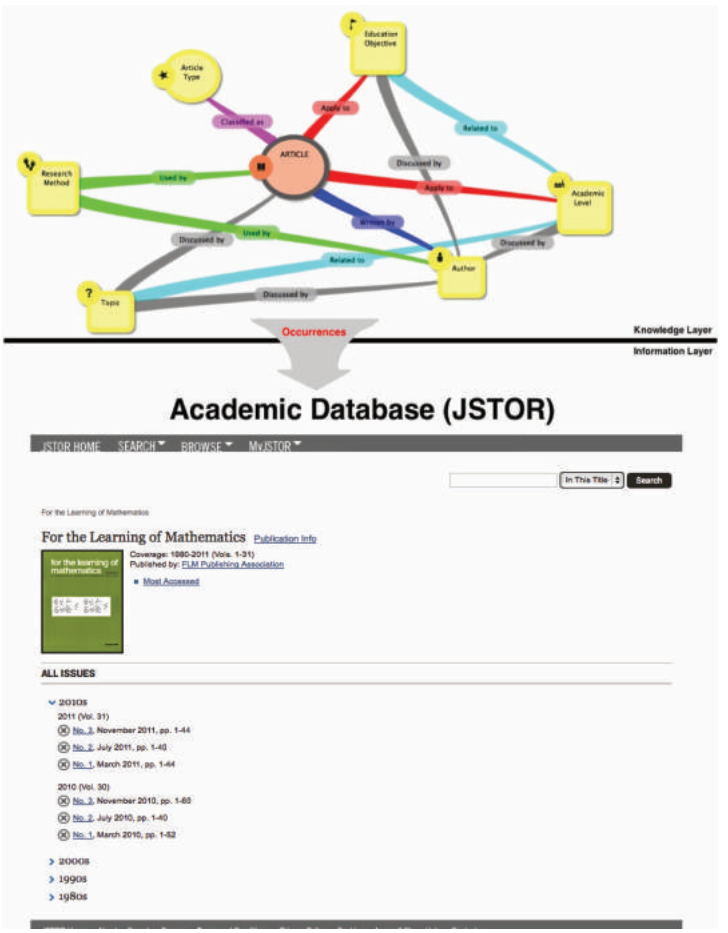


Figure 1 The 2-layer Topic Map model

Figure 1 shows that Topic Maps consist of two layers: the information layer contains a set of information resources (i.e., the database), and the knowledge layer contains a knowledge map consisting of topics and their associations. These two layers are connected by occurrences or individual documents. The 2-layer model separates the topics from the resources that demonstrate these topics, and it establishes the relationships between topics without accessing the database of resources (Venkatesh, 2007). Therefore, search results can be efficiently filtered or grouped by topic and explored using the relationships between topics. Compared to traditional searching applications that strictly retrieve documents matching the user specified keywords, the Topic Map search tool allows users to navigate from topic to topic without entering additional keywords.

4.2 Web Site Development

Ontopia was selected as a Web site development framework. It contains a set of tools whose engine stores and maintains the topic maps. Ontopia also includes Ontopoly and Omnigator, both used to design and display the Topic Maps. The current version provides access to 108 articles; this is a small collection used for the initial prototype and preliminary user testing.

As shown in Figure 2, Omnigator allows users to begin a search by visually exploring topics without entering keywords. When a user chooses a topic, results (e.g., articles) related to this topic are returned by the database. Since the results are grouped by topic, users can find the results related to the same topic quickly and accurately while wasting less time filtering irrelevant results. When users find a relevant article, they can either retrieve the article using an external link, or consult the article's metadata stored with the Topic Map. Users can also inspect related topics to retrieve articles assigned to other topics without restarting the search process. For example, as shown in Figure 2, assuming the user has found an article entitled *"Which Operations? Certainly not Division!"* related topics are listed in the associations (e.g., Applied to: elementary).

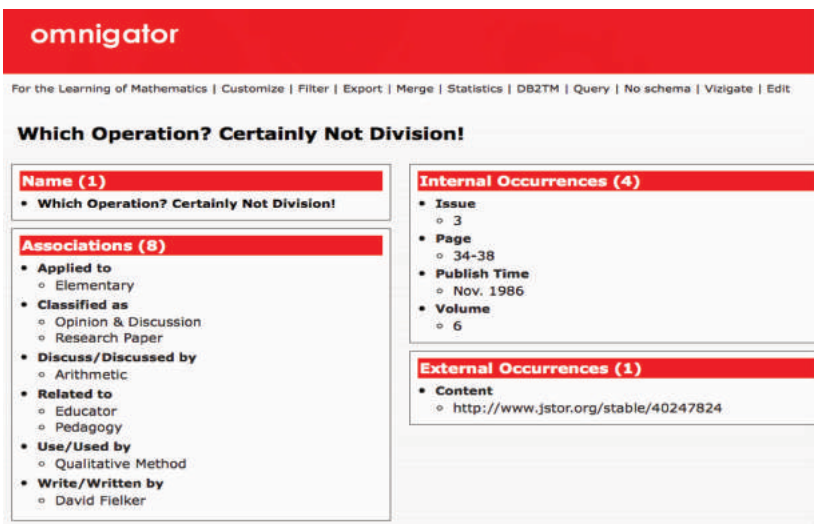


Figure 2 Screenshot of the Omnigator - Article

If the user desires to read more articles concerning Arithmetic, she may click on "Arithmetic" under its "Discuss/Discussed by" association and the system returns all articles assigned to "Arithmetic" as shown in Figure 3.

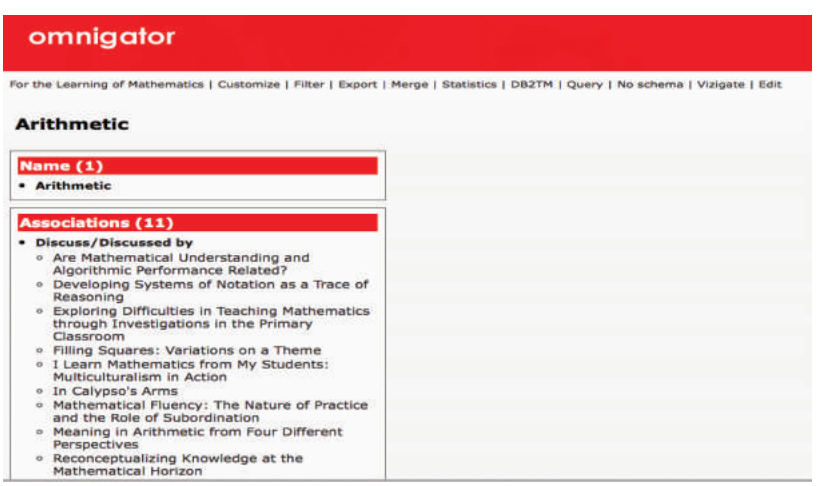


Figure 3 Screenshot of the Omnigator-Arithmetic

4.3 Test

A pilot test was conducted to collect users' experience with the system. This was not a task-based evaluation but a user-centered evaluation placing emphasis on users' satisfaction. Since the system is designed for retrieving scholarly mathematics articles, most users are graduate students or scholars in mathematics or information science. According to this user population, two participants were selected from School of Information Studies at McGill Universityand three participants were selected from Department of Mathematics at Concordia University. The participants were selected by means of a purposive sample using the snowball recruiting method to identify the participants. Participants were recommended by friends and informed the purpose of my research and the method of data collection.
Participants were not required to complete any specific tasks or queries but to share their experience in using this new Topic Maps search tool.They were asked to read a document that described the purpose of the test and brief usage instructions for Omnigator. They were asked to use Omnigator to find mathematics articles of interest. To ensure an adequate familiarization with the system, they were required to spend at least two hours using the system in one week. Afterwards, they would share their stories and express their feelings concerning the system by answering four open-end questions:

- Do you like this novel Topic Maps search tool (Ontopia) in general?
- Does Ontopia provide satisfactory search results in terms of accuracy as compared to searching directly in the academic database?
- Does Ontopia provide a satisfactory performance in terms of search time compared to searching directly in the academic database?
- Are you satisfied with Ontopia in terms of its layout, interface and functionality?

5 Results

All five participants stated they liked the idea of searching for articles by topics using the novel Topic Maps search tool (Ontopia). They believed that Ontopia had positive effects on search accuracy because it filtered the articles that did not concern their chosen topic even when these articles matched the keyword search term(s). All five participants declared that Ontopia saved time because they could navigate from topic to topic, or from article to article, without typing keywords or accessing the database of articles. Four out of five participants were satisfied with Ontopia in terms of its layout, interface and functionality. One participant complained that he could not search through the full-text of the articles since Ontopia offers a full text search restricted to metadata.This could be a weakness of the system since a Topic Map integrated with full-text keyword searching may improve users' searching performance by offering both search techniques in conjunction.

6 Conclusion

This ongoing project illustrates the potential of Topic Map search tools, and ontologies in general, to retrieve articles from an academic database, and improve our understanding of the possible interactions between users and Topic Maps. A Topic Map integrated with a traditional keyword search engine may facilitate searching by grouping the search results by topics, and improve users' searching performance by allowing them to navigate from topic to topic without inputting keywords; thus, allowing users to retrieve more accurate

information in less time.

Reference

Dicheva, D., & Dichev, C. T.(2006). Creating and browsing educational topic maps. *British Journal of Educational Technology, 37* (3), 391–404.

Garshol, L.(2004). Metadata? thesauri? taxonomies? topic maps! making sense of it all. *Journal of Information Science, 30*(4), 378–391.

Jimeno–Yepes, A., Berlanga–Llavori, R., & Rebholz–Schuhmann, D.(2010). Ontology refinement for improved information retrieval. Information *Processing and Management, 46*(4), 426–435.

Kolås, L.(2006, October). Topic maps in e–learning: An ontology ensuring an active student role as producer. In E–Learn: World Conference on E–Learning in Corporate, Government, Healthcare, and Higher Education(pp. 2107–2113). Association for the Advancement of Computing in Education(AACE).

Kwong, L. W., & NG, Y–K.(2003). Performing binary–categorization on multiple–record web documents using information retrieval models and application ontologies. *World Wide Web: Internet and Web Information Systems, 6*(3), 281–303.

Lalingkar, A., & Ramani, S.(2010, June). A topic map–based system for identifying relevant learning objects. In *EdMedia+ Innovate Learning* (pp. 1044–1053). Association for the Advancement of Computing in Education (AACE).

Mann, T.(2008). Will Google´s keyword searching eliminate the need for LC cataloging and Classification?. *Journal of Library Metadata, 8*(2), 159–168.

Melgar, E. L. M.(2011). Topic maps from a knowledge organization perspective. *Knowledge Organization, 38*(1), 43–61.

Michael, R. T. F., & Rajesh, K.(2011). A Comparitive study: FLOWS and PSL model in Selecting the Ontologies for Dynamic Web service Selection in Semantic web Environment. *International Journal on Computer Science and Engineering, 3*(4), 1666–1671.

Patkar, V.(2011). A passage to ontology tool for information organisation in the digital age. *DESIDOC Journal of Library & Information Technology, 31*(2), 90–102.

Pelczer, I., Cook, M. C., Venkatesh, V., Gatbonton, E., & Segalowitz, N.(2012, October). Building an ontology for inuttitut language samples: An application of topic maps in Canadian aboriginal language preservation. *In E–Learn: World Conference on E–Learning in Corporate, Government, Healthcare, and Higher Education*(pp. 1233–1238). Association for the Advancement of Computing in Education(AACE).

Pepper, S.(2000, June). The TAO of topic maps. In *Proceedings of XML Europe*(Vol. 3, p. 77).

Shih, B. J., Shih, J. L., & Chen, R. L.(2007). Organizing learning materials through hierarchical topic maps: An illustration through Chinese herb medication. *Journal of Computer Assisted Learning, 23*(6), 477–490.

Urbaniak, K., & Venkatesh, V.(2012, October). Building an ontology for student models of english as second language essays: A synthesis of literature and results of development of topic map indexes. In *E–Learn: World Conference on E–Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 432–438). Association for the Advancement of Computing in Education(AACE).

Venkatesh, V.(2007, October). Theoretical and practical implications of using topic map technologies in e–learning applications. In *E–Learn: World Conference on E–Learning in Corporate, Government, Healthcare, and Higher Education*(pp. 1311–1319). Association for the Advancement of Computing in Education(AACE).

Widhalm, R., & Mueck, T.(2003). Using topic maps for eLearning. In *E–Learn: World Conference on E–Learning in Corporate, Government, Healthcare, and Higher Education*(pp. 2023–2030). Association for the Advancement of Computing in Education(AACE).

Wilson, M. L., Schraefel, M. C., & White, R. W.(2009). Evaluating advanced search interfaces using established information–seeking models. *Journal of the American Society for Information Science and Technology, 60* (7), 1407–1422.

Yi, M.(2008). Information organization and retrieval using a topic maps–based ontology: Results of a task–based evaluation. *Journal of the American Society for Information Science and Technology, 59* (12), 1898–1911.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Directionality of paper reviewing and publishing of a scientist: A Granger causality inference

Chunli Wei[a], Yi Bu[b], Lele Kang[a], Jiang Li[a*]

a. School of Information Management, Nanjing University, Nanjing, China
b. Department of Information Management, Peking University, Beijing, China

ABSTRACT
It has been evidenced that peer review activities are positively correlated to scientists' bibliometric performance (e.g., Ortega, 2017, 2019). However, how the number of paper 'reviewing' interacts with a scientist's 'publishing' has not been addressed in previous studies. This paper attempts to employ the Granger causality inference to explore the directionality between a scientist's publication performance and his/her review activities. Our dataset comprises scientists' reviewed articles derived from Publons in the Web of Knowledge database, and their publications retrieved from PubMed. We find that scientists who reviewed less or published less tend to have Granger causality between reviewing and publishing activities. In addition, compared with early-career researchers, reviewing advances publishing for senior scientists.

KEYWORDS
Granger Causality Inference; Peer Review; Scientific Publications; Science of Science

1 Introduction

Peer review is a process of subjecting an authors' scholarly work, research, or idea to the scrutiny of others who are experts in the same field (Ware, 2008). This procedure improves the quality of manuscripts and filters the scientific community's scientific outputs. It is the heart of all science through which papers are published, grants are allocated, researchers are promoted, and prizes are awarded (Smith, 2006).

According to the recent peer review system, scholars' activity to review submitted manuscripts is underpaid. There exist at least two reasons why scholars are willing to review activities. On the one hand, some scholars considered peer review as one important part of their academic job; on the other hand, it is believed that the peer review process is an invaluable approach for researchers to stay up-to-date with research trends in their fields. However, the number of submitted manuscripts is increasing rapidly year by year, which has caused the demand for reviewers to outstrip the supply. The overload reviewing work for each scholar may cause their declining review invitations. The primary reason is that the effort put into this procedure has not been adverted into a reward system among the scientific community (Ortega, 2017). The lack of recognition for reviewers can be attributed to the difficulty of identifying or quantifying the quality of review activity because the personal information

* Corresponding author: lijiang@nju.edu.cn

about reviewers and review records in the current peer review system is anonymous. Consequently, There are no metrics to measure reviewing activities of scholars and it is more difficult for researchers to explore relationships between reviewing activities and other academic activities. By this time, scholars will weigh up the pros and cons of reviewing activity to decide whether they accept the request of reviewing manuscripts from journal editors and/or conference chairs.

To this end, some online websites, such as Publons[1], a peer review platform, attempted to identify scholars' contributions for their reviewing activities. In this platform, reviewers' devotion can be acknowledged, and journal editors can find appropriate reviewers for submitted manuscripts (Cuellar, 2018). In the meantime, such a platform has provided an opportunity for scholars to dive into peer review activity and relationship with other research activities profoundly, such as paper publishing.

The relationship between reviewing activity and other research activities, such as publishing activity, has been discussed in recent years (Ortega, 2017, 2019). They primarily implemented correlation-level analyses between reviewing and publication activities. Yet, the directional effect between the two activities has been ignored. In another word, it is unclear whether more review tasks lead to more papers published, more papers published result in more review tasks, or bidirectionality between two activities. Apart from directionality, the time lag between the two time series (i.e., the monthly numbers of publications/reviews) is also neglected. For example, as there is a life cycle of one scientific publication, the number of reviewing articles in this year may affect the number of articles published in the next year. Simultaneously, the lag time should vary for different scholars in various disciplines rather than a fixed value for all scholars.

Therefore, this study aims to explore the directional effects between reviewing and publishing activities based on the publication records derived from PubMed and the reviewing records acquired from Publons from 2012 to 2018. We conducted Granger-causality test to examine directionality between two activities. Although Granger-causality inference cannot indeed illustrate the "real causality" of peer review and publication of scholars, this model's result could still offer us more significant information, such as the directionality between two activities, than correlation analysis. In the meantime, we will conduct Granger-causality inference case by case to identify the fittest lag time for each scholar.

2 Related Studies

In academia, scholars tend to publish academic manuscripts in journals or attend academic meetings to improve the knowledge reserve in professional disciplines and support their scholarly productivity (Newhart et al., 2020). Publications of scholars have become a significant indicator for rewards, funding grants, and promotion (Inoannidis et al., 2014). There exist quantities of subjective and objective factors influencing scholars' publication in their academic careers.

Subjective factors, including gender, family, and time constraint, may affect scholars' publication productivity (Newhart et al., 2020). Although the gender difference in scholars' publications has been decreased over the last 30 years (Caplar et al., 2017), gender inequality still

[1] https://publons.com/.

exists in some disciplines. For instance, in astronomy (Mayer et al., 2017), female authors write 19 ± 7% fewer papers in seven years following their first paper than their male colleagues. In the meantime, publication productivity is related to marriage. For example, for women particularly, the relationship between productivity and marriage varies by type of marriage: first compared with subsequent marriage and spouse's occupation in science (compared with non-scientific employment). Women with preschool children have higher productivity than women without children or school-age children (Fox, 2005). Besides, the publication process is a demanding and take-consuming activity, and time constraints may be the most common barrier to publication of scholars (Chen, 2011).

Objective factors, containing faculty ranks and source of funding, also have a significant influence on publication productivity (Newhart et al., 2020). Historical literature illustrated that researchers with upper levels performed more remarkable than those with lower ranks. For instance, based on all Italian university researchers' performance in the hard sciences for the period 2004-2008, Abramo et al. (2011) found that lower academic ranks typically owned less output than higher grades. The higher levels holding greater seniority and more incredible experience in the professional fields contributed to this apparent phenomenon. Furthermore, funding is positively correlated with increased output, and researchers who received funding from the aerospace engineering program published 2.59 articles more than those not receiving funding support (Goldfarb, 2008).

Peer review is fundamental and essential in the scientific process. It can provide quality control of what science should be published, funded, and who should be promoted (Wagner, 2006). Meanwhile, the peer review is underpaid for reviewers. Therefore, a successful peer review system depends on the reviewer's willingness to review manuscripts. Rapid growth in scientific production puts a burden on the scientific peer review system, and the system is facing a crisis (Kovanis, 2016). Editors have assumed that it is the overload of reviewing activity that makes researchers less willing to perform the anonymous, time-consuming yet underpaid tasks associated with reviewing papers (Breuning, 2015). Fortunately, some online platforms such as Publons attempted to give scholars credit for their reviewing activities and provided a new research perspective matching bibliometrics indicators to assess scholars' output.

Scholars have dived into the relationship between publication activities utilizing bibliometrics indicators and reviewing activity based on the scholars' reviewing records from Publons. For instance, based on publishing records derived from Google Scholar and reviewing data from Publons, Ortega (2017) found that there seems to be a weak correlation between bibliometric indicators, such as the number of publications. Similarly, Ortega (2019) explored the relationship between Publons metrics and altimetric counts, and there is also a week relationship between them. Based on the previous studies, it is obvious that the correlation does not uncover the directionality of two variables or consider the lag time of two-time series. This paper will utilize the Granger-causality inference method to discover the directional relationship between reviewing activity and publication productivity of scholars (Granger, 1969).

The Granger causality inference has been applied into quantities of discipline, such as business economics (Akkemik & Göksa, 2012; Rahimi et al., 2017),mathematics (Inglesi-Lotz et al., 2014), neurosciences neurology (Barnett et al., 2014; Chen et al., 2006), behavior science (Schippers et al., 2011), Psychology (Wang et al., 2007; Zhou et al., 2009), public administra-

tion (Altuzarra & Esteban, 2011; Beyzatlar et al., 2014; Bilen et al., 2017). For example, based on Spanish quarterly data for 1977-1998, Bajo-Rubio (2001) analyzed the relationship between outward foreign direct investment (FDI) and exports. The results indicated that the relationship between two variables is from outward FDI to exports in the short run and bilateral Granger causality in the long run (ajo-Rubio & Montero-Muñoz, 2001). Zhang (2011) utilized some econometric techniques, including the Granger causality test, etc., to explore the influence of financial development on carbon emissions and found that China's economic development acts as an important driver for carbon emissions increase. However, in management science, there seem just a few papers applying this method (Chang et al., 2018). For instance, by tracking 3,390 products on Amazon.com over two months, Ren (2018) found that the volume of negative consumer reviews drives consumers' purchasing decisions, but the magnitude of positive consumer reviews only marginally affects purchasing decisions.

3 Data and Methods

Our dataset consists of time series of scholars' publishing and reviewing records. The two metrics indicators are the number of publishing and reviewing articles per month. The publication data is derived from the PubMed database. PubMed is a free source developed and maintained by the National Center for Biotechnology Information (NCBI), a division of the U. S. National Library of Medicine (NLM), at the National Institutes of Health (NIH). PubMed citations and abstracts include biomedicine and health fields and cover portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering (Canese & Wei, 2013). Simultaneously, the peer-reviewing records are acquired from the platform, called Publons. Publons is the world's leading peer review platform to officially recognize the reviewer's contribution to the Journal of Transcultural Nursing (JTCN) (Cuellar, 2018). In this platform, scholars can create a personal profile to display the information of manuscripts they have reviewed, such as, the journal and numbers of reviewing articles.

The reviewing records covered all the scholars' activity in the Publons database from January 2012 to December 2018. The publication records are also between January 2012 to December 2018 from the PubMed database to ensure the two time comparability series. Firstly, we excluded the review records without ORCID in the Publons database (1,192,255 review records remained). Then, we acquired 4,072,414 articles with ORCID and information about manuscripts' publication time from the PubMed database. We obtained the monthly numbers of publishing and reviewing articles between January 2012 to December 2018 by matching ORCID (1,467,950 records of 49,379 scholars remained). Finally, we excluded the time series pairs whose lengths are less than 15 to ensure that a sufficient length of time series for further analyses (1,219,507 records of 23,126 scholars remained) (Hoffmann et al., 2005). Subsequently, our samples comprised all series whose sizes are at least 15 time points for both variables without missing value. The max length of the series equals 226 months. In other words, the length of time series is between 15 months to 226 months. Finally, we utilized Granger causality inference step by step and case by case to uncover directional effects between reviewing and publishing activities. This model's core concept is to introduce accurate lagged variables for every time series and examine the effect from the lagged form of one variable on the other.

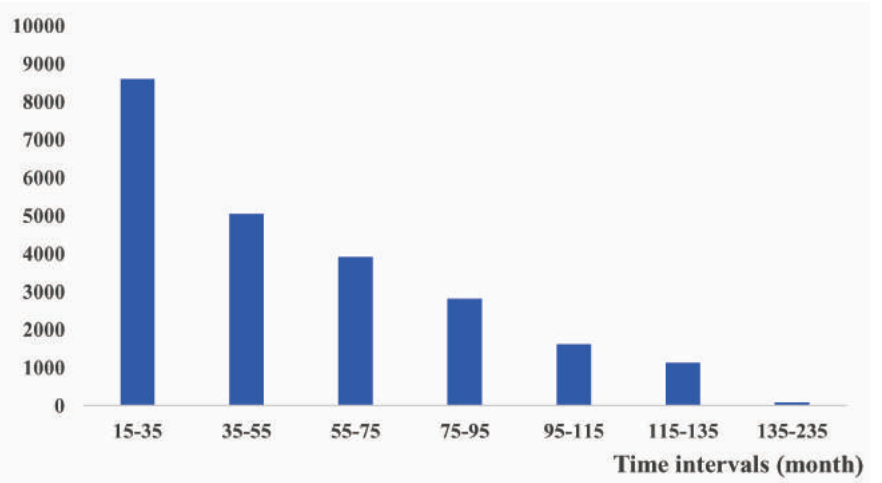


Figure 1 Length of time series in our dataset

Moreover, we examined the different directional effect patterns with different academic ages, different productivity of reviewing and publishing activity utilizing One-way ANOVA.

4 Empirical results

Table 1 has displayed the statistic descriptive of our data sample. Then, we conducted the Granger-causality test to explore the directional relationship between reviewing activities and publication productivity of scholars. This method consists of four steps (Hu et al., 2021), including the stationarity test, confirming each time series pair's accurate lag time, cointegration tests, and Granger-causality tests.

Table 1 Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
year	1,219,507	2015.147	2.535	2012	2018
month	1,219,507	6.711	3.406	1	12
# pub	1,219,507	.351	.477	0	1
# review	1,219,507	.223	.478	0	15

4.1 Stationarity tests

Time series analysis has an assumption that the utilized time series should be stationary. In fact, in many cases, the time series are non-stationary. If we run a non-stationary time series, a spurious result will be acquired, and then we will fail to speculate the true trend of the time series. The core idea of stationarity tests is to check whether both time series for each scholar has a unit root. If there is a unit root, the time series seems non-stationary.

This paper will utilize the augmented Dickey-Fuller tests (ADF) to test each time series pair case by case. If a time series passes the ADF test as p values are significantly less than a threshold (the value of the threshold is set as 0.05 in this paper), we can identify that the time series is stationary.

We checked the two time series, which describe monthly publication records and reviewing records for scholars case by case. Finally, based on the results of the test for all scholars, we divided scholars into three types :

- Type 1: The publication records and reviewing records both passed the ADF test, which suggested that both time series are stationary;
- Type 2: Neither the publication records nor reviewing records passed the ADF test, illustrating that neither time series is stationary;
- Type 3: The publication records and reviewing records don't pass the ADF test, which indicated that publication and reviewing time series are both non-stationary.

Because time serials pair of type 1 are stationary, the next step for this kind of pair is to confirm the accurate lag time. Concerning types 2 and 3, we should conduct first-order differences for both time series, and then we retested differential time series. Finally, we excluded the scholars whose time serials after first-order differences still don't pass the ADF test.

Table 2 Stationarity test results.

Stationarity test		Stationary	Non–stationary
Before difference	# scholars	22238	888
	% scholars	96.16%	3.84%
After difference	# scholars	23103	23
	% scholars	99.01%	0.99%

Table 2 shows that 99.01% of scholars have stationary time series pairs after first-order differences. Finally, the 23,103 scholars' records will be utilized in the following steps.

4.2 Confirming the fittest lag time for each time series pair

Many researchers have analyzed two time series and compared their relations in different fields. However, the time lag is often a fixed or a simple value (Moed, 2016). This paper will confirm a more accurate time lag for each scholar's time series pair.

An approach, called vector autoregression, has been introduced to confirm the lag time for time series pairs in Granger-causality tests. There exist amounts of indicators as the criterion for us to identify the fittest lag time for time series, such as AIC (Akaike Information Criterion) (Aho et al., 2010), BIC (Bayesian Information Criterion) (Bhat & Kumar, 2010) and LR (likelihood ratio) test (McGee, 2002). These indicators estimate prediction error and speculate the relative quality of statistical models for a given set of data. In this paper, we utilized AIC and BIC as the criteria to select the fitness lag time for each time series pair. If the values of AIC and BIC in one model are the minimum compared to other models for one time series pair, the lag time in this model is the fittest one for this scholar.

We conducted a VAR model for each series time pair that has remained after being filtered before and set the maximum lag to eight. Ultimately, we acquired an accurate lag time for each scholar's time-series pair. Figure 1 showed the time lag distribution for all scholars in our dataset.

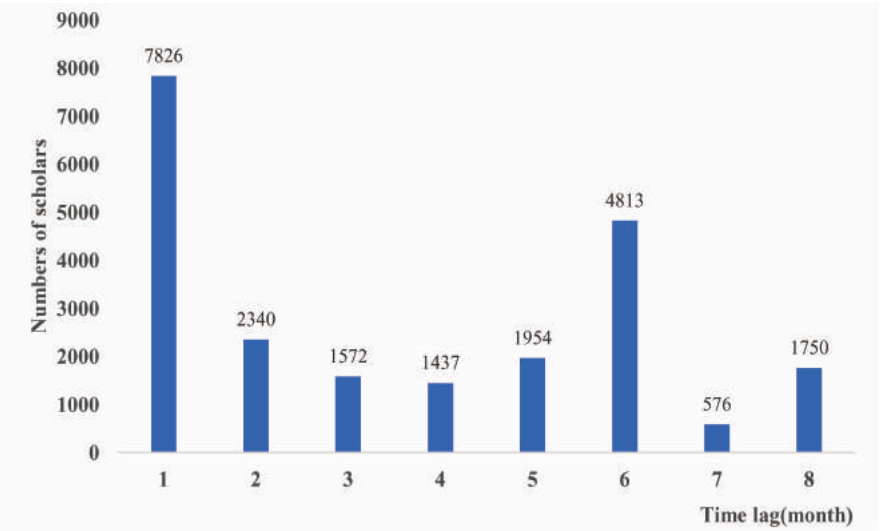


Figure 2 Distribution of time lag in our dataset

4.3 Cointegration tests

We have acquired the 23,103 scholars' stationary time-series records. These records can be divided into two categories: One is the raw time series, which are both stationary, and another one is the time series whose raw time series are non-stationary but stationary after first-order difference. As for the former series, we don't need any extra operation in this step. But it is possible that although time-series pairs are non-stationary, there may remain a statistically significant connection between the two variables. Therefore, we need to conduct cointegration tests case by case for latter records to check for a cointegrated combination of the two series.

There are three main methods for cointegration tests: Engle-Granger two-step method (Engle et al., 1987), Johansen test (Johansen, 1995), and Phillips-Ouliaris cointegration test (Phillips & Ouliaris, 1990). The Johansen test is a test for cointegration that allows for more than one cointegrating relationship for a large sample (Pesaran et al., 2001). This paper will apply the Johansen test to check the cointegrated combination of two series.

There are 1000 scholars' monthly time series pairs whose raw records are non-stationary, yet the differential records are stationary that should be conducted with cointegration tests. Finally, 146 scholars passed the test, and the rest are excluded.

4.4 Granger–causality test

The Granger causality test is a statistical hypothesis test for determining whether one-time series is useful in forecasting another, first proposed in 1969 (Granger, 1969). According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone (Granger, 1969). The Granger-causality test's null hypothesis is that X1 doesn't Granger cause X2, or X2 doesn't Granger cause. If the time series pair passes the grange-causality as the *p*-value is significantly less than 0.05, we will reject the hypothesis. For example, If the number of publications "Granger-causes" (or "G-causes") the number of reviewing articles, the number of publishing articles of a scholar

has a significant effect on the reviewing articles.

Based on the stational time series pairs, and the fittest time lag of each time series pair, we can conduct the Granger-causality test for all stationary series pairs in our dataset. Table 3 shows the Granger-causality test result based on monthly publication and reviewing records for scholars. In Table 3, "Publication→Reviewing" means that the publication productivity of scholars "Granger-causes"the reviewing activity, while"Reviewing→Publication" indicates that reviewing activity may influence the bibliometric performance of scholars." Publication⟷Reviewing" suggests that the number of reviewing articles and publishing manuscripts affect each other bidirectionally. As shown in Table 3, 42.3% of scholars have no significant effect between two-time series. 32.5% of scholars show a one-way product between two-time series, including 3720 scholars whose reviewing activity Granger cause publication productivity and 3518 scholars whose publication activity influences reviewing activity. In the meantime, 25.2% of scholars show a bidirectional effect between two-time series pairs.

Table 3 The result of Granger-causality test

	Reviewing→ Publication	Publication→ Reviewing	Publication⟷ Reviewing	No sig.
# scholars	3720	3518	5601	9422
% scholars	16.7%	15.8%	25.2%	42.3%
Academic age	5.50	5.43	4.39	5.63
# publishing articles per year	1.70	1.73	1.72	2.18
# reviewing articles per year	2.71	2.63	2.87	3.47

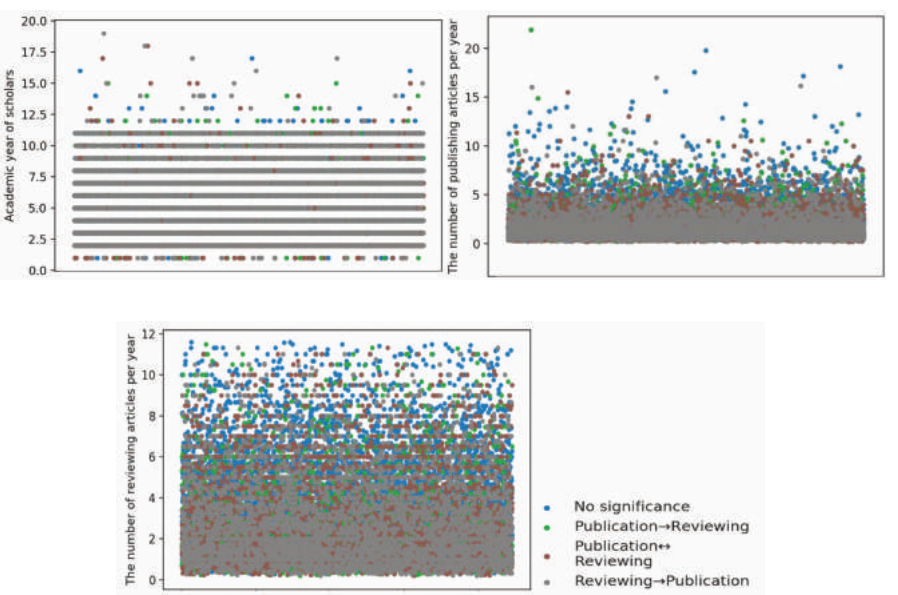


Figure 3 The left upper corner of the figure displays the distribution of academic year of scholars in four groups, and the right upper one displays the distribution of publishing productivity of scientists in four group. The distribution of academic year of scholars has been displayed on the lower.

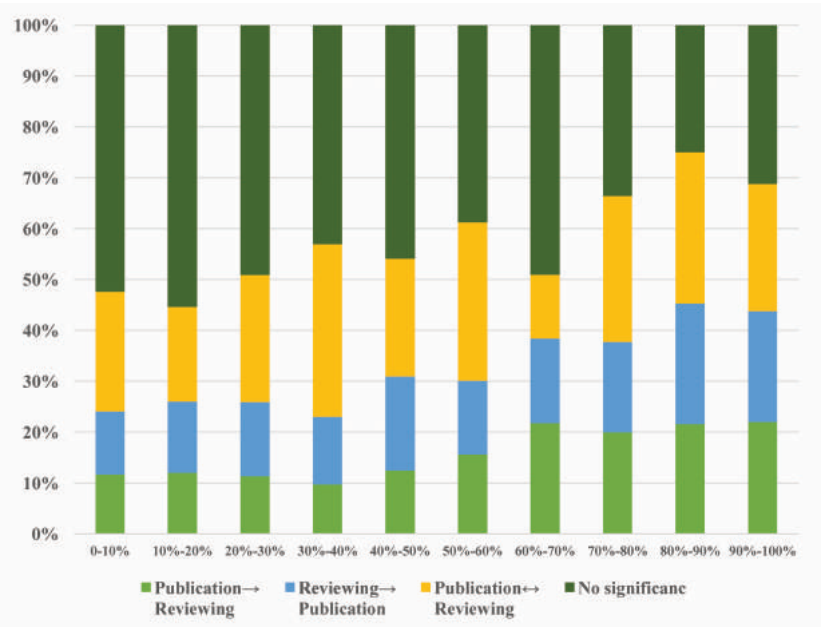
Moreover, we conducted One-way ANOVA to demonstrate the difference in publishing articles, reviewing articles, and academic age of scholars among four groups, including "Publication → Reviewing," "Reviewing → Publication," "Publication↔Reviewing," and "No significance." As shown in Table 4, we compared the number of publishing articles of scholars among four groups. We identified that the scientists in the "no significance" group own more publishing articles per year than in other groups with $p<0.05$. We also compared the numbers of reviewing items per year among four groups and found that scholars showing no significance between publishing and reviewing activities owned more reviewing articles than others. Meanwhile, scientists who displayed a bidirectional effect between two activities had more publishing articles than those who owned a one-way influence. Surprisingly, the scholars in "Reviewing→ Publication" group is significantly older than other groups.

Table 4 Result of One-Way ANOVA

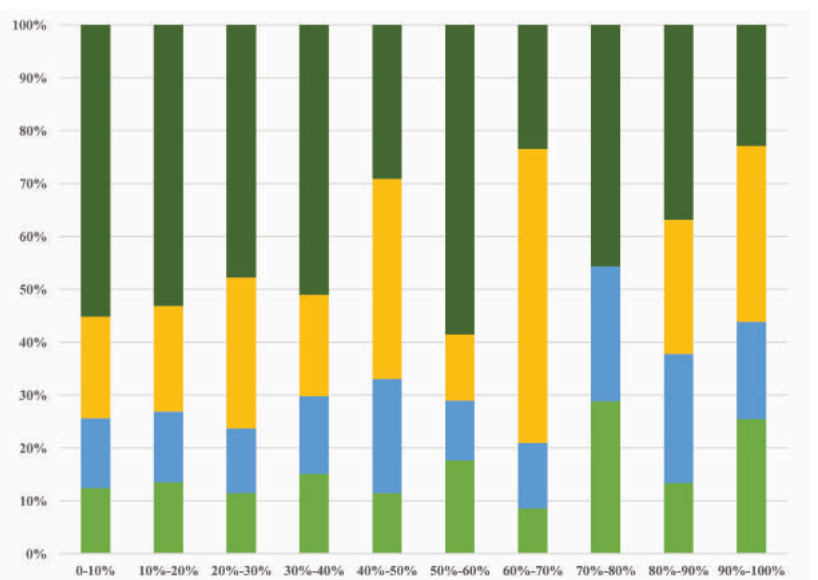
	Mean	Std. Err.	t	P>|t|
The number of publishing articles per year				
Publication→ Reviewing vs No significance	−.450	.029	−15.38	−0.000*
Reviewing→ Publication vs Publication↔Reviewing	−.011	.031	−0.36	0.984
Publication→ Reviewing vs Publication↔Reviewing	.019	.032	0.60	0.934
Publication→ Reviewing vs Reviewing→ Publication	.030	.034	0.87	0.822
No significance vs Publication↔Reviewing	.470	.025	18.77	−0.000*
No significance vs Reviewing→ Publication	.480	.029	16.74	−0.000*
The number of reviewing articles per year				
Publication→ Reviewing vs No significance	−.836	.041	−20.29	−0.000*
Publication→ Reviewing vs Publication↔Reviewing	−.243	.045	−5.41	0.000*
Reviewing→ Publication vs Publication↔Reviewing	−.167	.044	−3.77	0.001*
Publication→ Reviewing vs Reviewing→ Publication	−.076	.049	−1.55	0.407
No significance vs Publication↔Reviewing	.594	.035	16.87	−0.000*
No significance vs Reviewing→ Publication	.760	.040	18.82	−0.000*
Academic age				
Publication→ Reviewing vs No significance	−.212	.052	−4.04	0.000*
Publication→ Reviewing vs Reviewing→ Publication	−.075	.062	−1.20	0.627
No significance vs Reviewing→ Publication	−.137	.051	2.67	0.038*
Publication→ Reviewing vs Publication↔Reviewing	1.033	.057	18.09	−0.000*
Reviewing→ Publication vs Publication↔Reviewing	1.108	.056	18.74	−0.000*
No significance vs Publication↔Reviewing	1.245	.045	27.80	−0.000*

In the meantime, we also examined the directional patterns with various reviewing and bibliometric performance, different academic age in Figure 3. We divided all scholars in the dataset into 10 groups based on scholars' academic age, the number of publishing articles per year and the number of reviewing manuscripts per year, respectively: 0-10%, 10%-20%, 20%-30%, 30-40%, 40%-50%, 50%-60%, 60%-70%, 70%-80%, 80%-90% and 90%-100%. For instance, 0-10% denoted that the scholars whose academic age, number of reviewing or publishing productivity ranks among the top 10%. From Figure 4 (a), It is obvious that

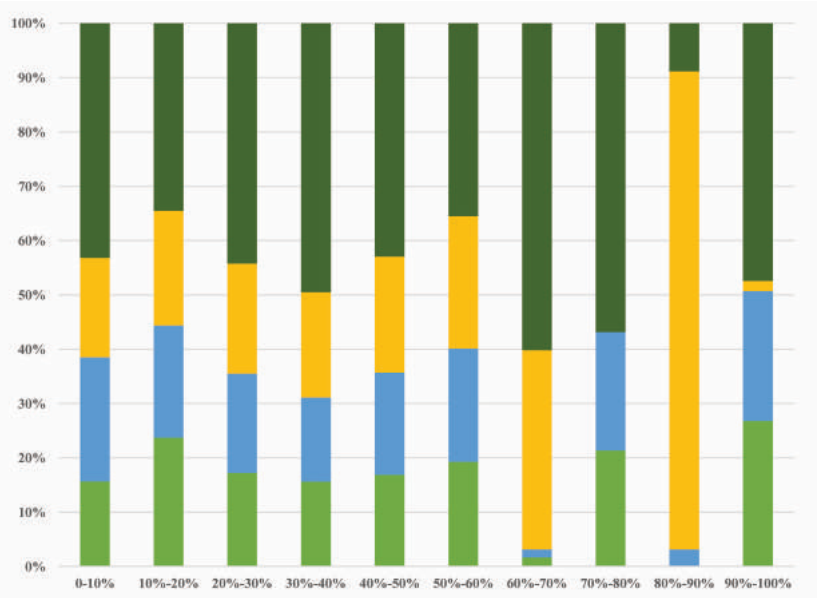
scholars with lower reviewing articles showed more Granger causality relationship between the two activities. It is consistent with the aforementioned findings in One-Way ANOVA. For example, for the scholars in group 90% -100%, 68.82% owned the Granger causality relationship between publishing and reviewing articles. However, for scholars whose numbers of reviewing rank among 0-10%, 47.62% displayed Granger causality inference. However, from Figure 4(b) and 4(c), There exists no apparent law in directional patterns of scholars with different publishing performance and various academic age.



(a)



(b)



(c)

Figure 4 Directional patterns with different groups (Figure (a) and Figure (b) illustrated the Granger causality result for the scholars with different numbers of reviewing and publishing articles, respectively. Figure (c) displayed the different directional patterns of scholars with various academic age)

5 Conclusions

This paper applied the Granger-causality test to uncover the directional relationship between scholars' publishing and reviewing articles. We focused on settling two issues. One is to confirm an accurate time lag for each time series, rather than a fixed value and the other one is to uncover directionality effect between two activities. We collected reviewing records and publishing records from Publons and PubMed, respectively, and utilized ORCID connecting publication records with reviewing records to acquire the time series pairs between January 2012 and December 2018 for each scholar. By conducting the Granger causality test step by step for each scholar, we found that 57.7% of scholars show a significant directional effect between publication and reviewing articles. The scientists who own fewer reviewing articles may have a Granger causality inference between reviewing and publishing activities compared to higher ones. Furthermore, the scholars who publish lesser articles tend to have more significant causality between two activities. Surprisingly, scientists with elder academic age tend to be in the "Reviewing→ Publication" group.

The peer-review system played an essential role in academic development. Unfortunately, the system breaks down with the rapidly increasing submitted manuscripts and the lack of acknowledgment for reviewers. Therefore, the attempt to explore the directional relationship between peer review activity of scholars and publishing activity, can provide more valuable suggestions for editors to select appropriate reviewers and scholars to decide whether they should accept requests from journal editors.

Although the Granger-causality test can't uncover true causality between two variables, this approach can introduce accurate lag time into the model and uncover the directionality

of effect between two-time series, which the correlation coefficient can't display. In the future, we can certainly focus on the causality of two activities to explore the intrinsic mechanism of unidirectional or bidirectional effect between two activities. For instance, if we can find the characteristic of scholars who shows the unidirectional effect from reviewing activity to publication productivity of scholars, It may provide scientific evidence for journal editors to allocate more manuscripts logically to improve the whole peer review system.

There also exits several limitations of this study. Publons is biased in disciplines and publishers (Ortega, 2019). For indiscipline, Health Sciences and Life Sciences, and Physical Sciences and Engineering are underrepresented in this platform. In publishers, Publons includes more articles from open access platforms. These biases could be one of the limitations of our study.

Reference

Abramo, G., D´Angelo, C. A., & Di Costa, F.(2011). Research productivity: Are higher academic ranks more productive than lower ones?. *Scientometrics, 88* (3), 915–928.

Aho, K., Derryberry, D., & Peterson, T.(2014). Model selection for ecologists: the worldviews of AIC and BIC. *Ecology, 95* (3), 631–636.

Altuzarra, A., & Esteban, M.(2011). Land prices and housing prices: the case of Spain. *Journal of Housing and the Built Environment, 26* (4), 397–409.

Akkemik, K. A., &Göksal, K.(2012). Energy consumption–GDP nexus: Heterogeneous panel causality analysis. *Energy Economics, 34* (4), 865–873.

Breuning, M., Backstrom, J., Brannon, J., Gross, B. I., &Widmeier, M.(2015). Reviewer fatigue? Why scholars decline to review their peers´ work. *Political Science & Politics, 48* (4), 595–600.

Barnett, L., Barrett, A. B., & Seth, A. K.(2018). Misunderstandings regarding the application of Granger causality in neuroscience. *Proceedings of the National Academy of Sciences*, 201714497.

Bhat, H. S., & Kumar, N.(2010). On the derivation of the Bayesian Information Criterion. *School of Natural Sciences, University of California*, 99.

Beyzatlar, M. A., Karacal, M., &Yetkiner, H.(2014). Granger–causality between transportation and GDP: A panel data approach. *Transportation Research Part A: Policy and Practice, 63*, 43–55.

Bajo–Rubio, O., & Montero–Muñoz, M.(2001). Foreign direct investment and trade: a causality analysis. O*pen Economies Review, 12* (3), 305–323.

Bilen, M., Yilanci, V., &Eryüzlü, H.(2017). Tourism development and economic growth: a panel Granger causality analysis in the frequency domain. *Current Issues in Tourism, 20* (1), 27–32.

Cuellar, N. G.(2018). Recognition for reviewers: PUBLONS! . *Journal of Transcultural Nursing, 29* (3), 221–221.

Chen, Y., Bressler, S. L., & Ding, M.(2006). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of Neuroscience Methods, 150* (2), 228–237.

Chen, X. P.(2011). Author ethical dilemmas in the research publication process. *Management and Organization Review, 7* (3), 423–432.

Chang, V., Chen, Y., & Xiong, C.(2018). Dynamic interaction between higher education and economic progress: A comparative analysis of BRICS countries. *Information Discovery and Delivery, 46* (4), 225–238.

Caplar, N., Tacchella, S., & Birrer, S.(2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy, 1* (6), 1–5.

Canese, K., & Weis, S.(2013). *PubMed: The bibliographic database*. In the NCBI Handbook [Internet]. 2nd edition. National Center for Biotechnology Information(US).

Engle, R. F., Granger, C. W. J., & Grangeri, C. W. J.(1987). COINTEGRATION AND ERROR CORRECTION: REPRESENTATION, ESTIMATION, AND TESTING. *Econometrica, 55* (2), 251–276.

Fox, M. F.(2005). Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science, 35* (1), 131–150.

Goldfarb, B.(2008). The effect of government contracting on academic research: Does the source of funding af–

fect scientific output?. *Research Policy, 37* (1), 41–58.

Granger, C. W. J.(1969). Investigating Causal Relations by Econometric Models and Cross–spectral Methods. *Econometrica, 37* (3), 424–438.

Hu, B., Ding, Y., Dong, X., Bu, Y.,, Ding, Y.(2021). On the relationship between download and citation counts: An introduction of Granger–causality inference. *Journal of Informetrics, 15* (2), 101125.

Hoffmann, R., Lee, C. G., Ramasamy, B., & Yeung, M.(2005). FDI and pollution: a Granger causality test using panel data. Journal of International Development: *The Journal of the Development Studies Association, 17* (3), 311–317.

Inglesi–Lotz, R., Balcilar, M., & Gupta, R.(2014). Time–varying causality between research output and economic growth in US. *Scientometrics, 100 (*1), 203–216.

Ioannidis, J. P., Boyack, K. W., &Klavans, R.(2014). Estimates of the continuously publishing core in the scien–tific workforce. *PloS One, 9* (7), e101698.

Johansen, S.(1995). Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. *Journal of Econometrics, 69* (1), 111–132. https://doi.org/10.1016/0304–4076(94)01664–L.

Kovanis, M., Porcher, R., Ravaud, P., &Trinquart, L.(2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PloS One, 11* (11), e0166387.

McGee, S.(2002). Simplifying likelihood ratios. *Journal of General Internal Medicine, 17* (8), 647–650.

Moed, H. F., & Halevi, G.(2016). On full text download and citation distributions in scientific–scholarly journals. *Journal of the Association for Information Science and Technology, 67* (2), 412–431.

Mayer, E. N., Lenherr, S. M., Hanson, H. A., Jessop, T. C., & Lowrance, W. T.(2017). Gender differences in publication productivity among academic urologists in the United States. *Urology, 103,* 39–46.

Newhart, S., Mullen, P. R., Blount, A. J., & Hagedorn, W. B.(2020). Factors influencing publication rates among counselor educators. *Teaching and Supervision in Counseling, 2* (1), 5.

Ortega, J. L.(2017). Are peer–review activities related to reviewer bibliometric performance? A scientometric analysis of Publons. *Scientometrics, 112* (2), 947–962.

Ortega, J. L.(2019). Exploratory analysis of Publons metrics and their relationship with bibliometric and altmetric impact. *Aslib Journal of Information Management, 71* (1), 124–136.

Phillips, P. C. B., &Ouliaris, S.(1990). Asymptotic Properties of Residual Based Tests for Cointegration. *Econo–metrica, 58* (1), 165. https://doi.org/10.2307/2938339

Pesaran, M. H., Shin, Y., & Smith, R. J.(2001). Bounds testing approaches to the analysis of level relationships. *Journal of Applied Econometrics, 16* (3), 289–326.

Rahimi, A., Chu, B. M., & Lavoie, M.(2017). Linear and non- linear Granger causality between short- term and long- term interest rates: A rolling window strategy. *Metroeconomica, 68*(4), 882–902.

Ren, J., Yeoh, W., Shan Ee, M., &Popovič, A.(2018). Online consumer reviews and sales: Examining the chick–en–egg relationships. *Journal of the Association for Information Science and Technology, 69* (3), 449–460.

Smith, R.(2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Soci–ety of Medicine, 99* (4), 178–182.

Schippers, M. B., Renken, R., &Keysers, C.(2011). The effect of intra–and inter–subject variability of hemody–namic responses on group level Granger causality analyses. *Neuroimage, 57* (1), 22–36.

Wagner, E.(2006). Analysing the purpose of peer review. *Nature.* doi:10.1371/journal.pone.0001636

Ware, M.(2008). Peer review: benefits, perceptions and alternatives(p. 2008). London: Publishing Research Con–sortium.

Wang, X., Chen, Y., Bressler, S. L., & Ding, M.(2007). Granger causality between multiple interdependent neuro–biological time series: blockwise versus pairwise methods. *International Journal of Neural Systems, 17* (02), 71–78.

Zhang, Y. J.(2011). The impact of financial development on carbon emissions: An empirical analysis in China. *Energy Policy, 39* (4), 2197–2203.

Zhou, Z., Chen, Y., Ding, M., Wright, P., Lu, Z., & Liu, Y.(2009). Analyzing brain networks with PCA and condi–tional Granger causality. *Human Brain Mapping, 30* (7), 2197–2206.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Mapping of Research Output in the Indian Veterinary Journal through Google Scholar

Kutty Kumar

College of Veterinary Science, Sri Venkateswara Veterinary Unviersity, Andhra Pradesh, India

ABSTRACT

This paper presents a scientometric study of Indian Veterinary Journal (IVJ) using Publish or Perish (PoP) during the period 1977-2016 (39 years). The study used Google Scholar to obtain raw citations and analyze various citation metrics to find the impact of Indian Veterinary Journal on emerging research. The growth of contributions by year, authorship trends, author productivity by Lotka's law, single and multi-authored papers by year, and the most prolific contributors was examined in a total of thousands of research papers. Also, the relative IVJ growth rate and doubling time are evaluated for the period of the sample. The data analysis revealed that the highest number of submissions, i.e. 223 (22.30 percent), was published in the years 1992-1996. The total number of authors per paper is 2.97, the highest number of the output of authors, i.e. 15 research papers; the collaboration degree is 0.91% . For a more comprehensive evaluation of the effects of research and scholarly production, the paper suggests journal and author productivity collaborative practice using sensibly selected metrics.

KEYWORDS

Scientometric Study; Indian Veterinary Journal; Lotka's Law; Publications; Author Productivity.

1 Introduction

The reputation of a scientific journal dominates investigators' choice of publication and is strongly influenced by the impact factor-a high impact factor demonstrates that recent journal publications are consistently cited in other journals (Garfield, 2006). The Impact Factor is determined by applying the citations for the previous two years to papers in the journal, separated by the number of items cited for those two years in the journal (Dong et al., 2005). Nonetheless, different fields show variable citation patterns (Kear & Colbert-Lewis, 2011). Usually, publication citation metrics provide a broad range of accomplishments focused on scientific and scholarly practices, and others serve as a valuable way to illustrate the success of the researchers and the importance of their current literature (Narin, 1976). The impact of a published work (and its host journal) in a subsequent publication through acknowledgement is monitored in the form of a citation.

1.1 About Indian Veterinary Journal

The Indian Veterinary Journal (https://ivj.org.in/en/webhome.aspx) is an official organ of the Indian Veterinary Association. It is the only publication in India representing the workers of academic veterinary science, growth and extension. The journal has been published since 1924, initially as a bimonthly publication and later as a regular monthly publication of the In-

dian Veterinary Association. The Indian Veterinary Journal's office has kept volumes of the Indian Veterinary Journal (IVJ) right from its inception in 1924. At this moment, a mapping of IVJ's contributions to academic science over three decades seems opportune. Bibliometric metrics are widely used to calculate study efficiency since they include views of a field that might not be apparent otherwise. The current study aimed to explore data on the output of publications to establish a picture of IVJ's research efficiency that could be useful to veterinary professionals and researchers. Figure 1 specifies the aims of the Indian Veterinary Journal.

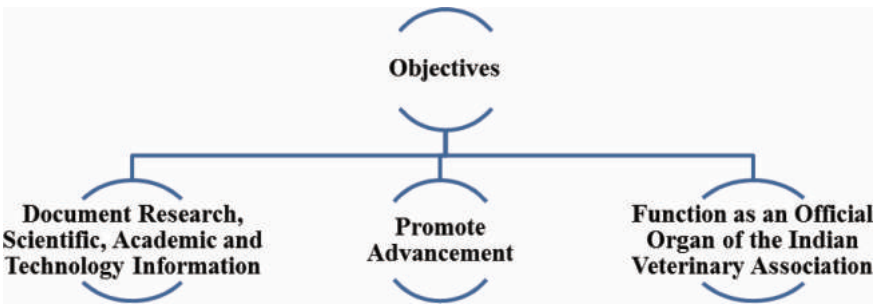


Figure 1 Objectives of the Indian Veterinary Journal

The Indian Veterinary Journal is dedicated to the cause of veterinary science and the advancement of the veterinary profession, with international status. The journal publishes original work as an official organ of the Indian Veterinary Association in the fields of veterinary medicine, surgery, reproduction, husbandry, fisheries and other related subjects, useful to professionals in the veterinary, dairy, livestock and poultry sectors. Of special interest to the journal are both large and small animal general veterinarians, researchers, field inspectors, cattle, poultry and dairy production officers, marketing officers and all animal health professionals. As a pioneer in veterinary journalism in the East, Balaraman (2018) has a much-unrivalled reputation as an authentic source of knowledge on all tropical diseases. This is a globally recognized arbitration publication. (http://www.connectjournals.com/ivj). This is real. This journal ranked by the National Academy of Agricultural Sciences (NAAS) with a mark of 6.0 on a scale of 1 to 10.

2 Overview of bibliometric research

Unsurprisingly, for a research method rooted in information science, many bibliometric studies have examined aspects of information science research and authors in information science. Voos (1974), for example, researched authors' effectiveness in the field of information science. Two recent books guide librarians on bibliometrics and altmetrics, and the contested area of research evaluation using metrics linked to publications tested a version of Lotka's law for writers writing in the field of knowledge science between 1996 and 2007 (Sobrino et al., 2009). Roemer and Borchardt (2015) present a guide for librarians on bibliometrics, altmetrics and research impact, Cronin and Sugimoto (2015) discussed multidimensional indicators of scholarly impact. Agarwal et al. (2016) offered a broad overview of the broad range of metrics commonly used in science and academia, and Dhiman (2015) discussed some of the newer metrics such as h-index, g-index, and I-index. Michael et al. (2010) discussed the benefits and drawbacks of the h-index.

There has been considerable interest in the applicability of Lotka's law to assess the author's efficiency in a sector. Gupta (1987) researched and analyzed writers' productivity models and checked the applicability of Lotka's law to four separate groups of data. Vlachy (1978) provided a bibliography of Lotka's and related work. Ahmed and Rahman (2009) checked the validity of the Lotka's law on the distribution of authorship in the field of nutrition study in Bangladesh. A list of periodic articles were published during 1972-2006 on various aspects of Bangladesh's nutrition research compiled for review. Using 'absolute productivity' of authorship, 998 personal author names were defined. Using both generalized and modified fonns-test and Kolmogorov-Smirnov goodness-of-fit tests, Lotka's law was tested. The findings suggested that in the generalized inverse square Lotka's law, the distribution of author productivity predicted did not apply to nutrition research in Bangladesh. Lotka's law, excluding highly efficient authors and maximum likelihood methods, was found to apply to Bangladesh's nutrition study using least-squares. Narendra (2016) discussed the applicability of Lotka's law in the Science and Industrial Council as a general inverse force for the distribution of research productivity. Wildgaard et al. (2014) explored the characteristics, including effect indicators over time, of 108 bibliometric indicators at the level of the author.

The subject of research into bibliometrics is very varied. For example, Majhi et al. (2016) aimed to analyze the content of wiki articles published in the journals of the Science Direct database. The identified research methods used, the type of data analysis techniques used for wiki articles, the most common country contributes to the largest number of articles, the largest contributing author, the annual publication and the history of the authors. Much research examines changes in the patterns of research and publication. Kumar (2016) analyzed 380 peer-reviewed articles published in IETE Technical Review-journal during 2007 to 2014, examining the growth pattern of research output, authorship patterns/co-authorship index/, collaboration coefficient, the geographical distribution of output and the average length of articles. Jesubright et al. (2014) studied the growth of forensic science literature from 1975 to 2011, the productivity of authors, the top-ranking source journal, and the productivity of the country.

Wan et al. (2009) reviewed bibliometric studies on single journals, noting that 28% of studies examined Indian journals. The study of Indian Economic Analysis (Nandi & Bandyopadhyay, 2008) generally examined the pattern of authorship, the degree of collaboration between authors, and the distribution of authors geographically. Swain (2014) completed a 10-year bibliometric overview of the International Information and Library Review. In a bibliometric analysis of the 104 African medical and health journals hosted in the African Journal Online database, Ezema and Onyancha (2016) used Harzing's Publish or Perish app.

3 Objectives

The primary objective of this review was to understand the development of the Indian Veterinary Journal during the period from 1977 to 2016 and the research output of contributors worldwide. The concrete goals were:

- Analyze the impact of IVJ on publication productivity through citation metrics.
- Study the distribution of articles and authorship patterns by year.
- Identify author collaboration, single and multi-authored papers by year.
- Find the Relative Growth Rate (RGR) and double the duration of the papers for study.
- Determine the application of the research productivity of Lotka's law of Author in IVJ.

4 Research Methodology

The research data was gathered from an online edition of the Indian Veterinary Journal accessible from 1977 to 2016 using Publish or Perish (PoP) (www.harzing.com). PoP is a Microsoft Windows program that, with the support of an appropriate emulator such as cross-over Mac or Wine, can also be installed and compatible on OSX and GNU/Linux computers; PoP retrieves and analyses scholarly citations. To evaluate different metrics, this analysis used Google Scholar to obtain raw citations. An important and realistic explanation for this is that Google Scholar is widely accessible and well known for its speed to anyone with an Internet connection (Notess, 2005). In contrast to other databases, Google Scholar offers a full image of academic effect (Pauly & Stergiou, 2005). A broad variety of publications are covered by the Indian Veterinary Journal, including academic articles, brief correspondence, reviews, and case studies. Based on citation metrics, necessary data were collected to evaluate the impact of the Indian Veterinary Journal and analyze bibliometric components such as article contributions by year, number of writers, authorship pattern, and authors productivity through Lotka's law to meet the objectives of the present study. As a final point, the data was organized, weighed, tabulated, assessed and presented as tables and graphs for interpretation and discussion.

The current standards for evaluating journal quality need to be understood by every reader. The evaluation of a particular journal's academic value helps to determine its merits and relevance to academic research and distinguishes it among other journals. The higher the impact metrics, the more highly ranked the journal is, but opinions differ as to what constitutes a "good" impact factor (Majhi et al., 2016). However, opinions differ. Although there is no 'correct' answer to this issue, in terms of various simple statistics (number of articles, number of citations and number of authors) and various other citation metrics of the Indian Veterinary Journal, there is a certain background in Table 1.

Table 1 Impact of Indian Veterinary Journal through citation metrics

S.No	Citation metrics	Value
1	Papers	1000
2	Citations	4863
3	Years	39
4	Cites/Year	124.69
5	Cites/Paper	4.86
6	Cites/Author	1885.11
7	Papers/Author	388.63
8	Authors/Paper	2.97
9	h index	17
10	g index	21
11	hc index	6
12	hI index	5.25
13	hI norm	9
14	AWCR	296.31
15	AW index	17.21
16	AWCRpA	111.89
17	e index	10.68
18	hm index	13.77
19	Cites Author Year	48.33
20	hIannual	0.23
21	h coverage	8
22	g coverage	10

The benefit of using Publish or Perish with Google Scholar is that it offers a much more accurate image of the impact of a journal than what would be possible with ISI impact factors / Thomson Journal Citation Reports. While the total number of publications (1000) provides useful information on productivity, which is strongly influenced by the number of years in which the journal has been producing research (39), the impact of their work, which is a limitation of the study to date, is not described. To assess the impact of a journal, different citation metrics are considered. Hirsch's h-index attempts to provide a rigorous single-number measure of an academic's impact, balancing quality with quantity (Hirsch, 2005). To calculate the performance of papers, Braun et al. (2006) suggest using the h index as an alternative to the impact factor given by Thomson Reuters (2019). The h-Index from IVJ is 17. The example worked on accounting journals in the Publish or Perish book (Harzing, 2013) has an h-index of 17 for the Accounting Horizons journal, with citations/paper of 8.45 (compared to IVJ's 4.86). The h-index may not, however, be a reliable indicator of recent results (Bornmann & Daniel, 2007). Egghe's g-index aims at boosting the h-index by giving more weight to frequently cited papers (Egghe, 2006; Sidiropoulos el at., 2007). The G-Index of the Indian Veterinary Journal is 21. In contrast, in accounting journals, the very high-impact (and international) ranges are 15-20 citations per article, h-indexes (28-43) and g-indexes (45-74). For every article, per author count (388.63) is determined to give the normalized author count for the paper. The sum of the author counts, separated by the total number of articles, across all papers was 2.97. The AWCR (296.31) calculates the total number of citations for the whole body of work, modified for the process of each paper (Jin, 2006). The individual h-index (5.25) and hI norm (9) as adjusted by Publish or Perish normalize the number of citations for each paper by dividing the number of citations by the number of authors for that paper and measuring the h-index of the uniform quotation count. Instead of minimizing citation counts, the multi-authored h-index (hm) uses fractional paper counts to account for the shared authorship of papers and then calculates the multi-authored hm index (13.77) based on the corresponding active rank of papers using undiluted citation counts (Schreiber, 2008). These seem to be more valid metrics where journal impact variables are immediately available and provide a simple way to test individual scientists or research groups. To obtain an objective and quantitative measure of the scientific achievement of the author, the journal impact factors of an author's publications can simply be applied, assuming that the journal is representative of its papers. Nevertheless, journal citation metrics are not statistically representative of individual journal publications and are poorly related to actual individual citations of articles (Seglen, 1997). Looking at the statistics provided in Table 1 and also as stated by Starbuck (2005), it is possible to conclude the impact of a journal on the productivity of publishing, but the confidence limits for such estimates are broad, particularly as the Journal Impact Factor changes every year.

5 Data Analysis and Interpretation

In this research, data were obtained from the Google Scholar online search engine on the bibliometrics records of the Indian Veterinary Journal for the period 1977-2016. A total of 1000 papers were gathered that produced the source data for the report. One of the most important metrics for determining the annual grade of publication growth and identifying the most efficient year of publication is year-by-year improvements in many published papers. Through Figure 2 it could be understood that the maximum number of articles were published during the years 1992-1996 (22.30%) and 19.20% articles during the years 2002-2006 and research publication was smaller during 1977-1986.

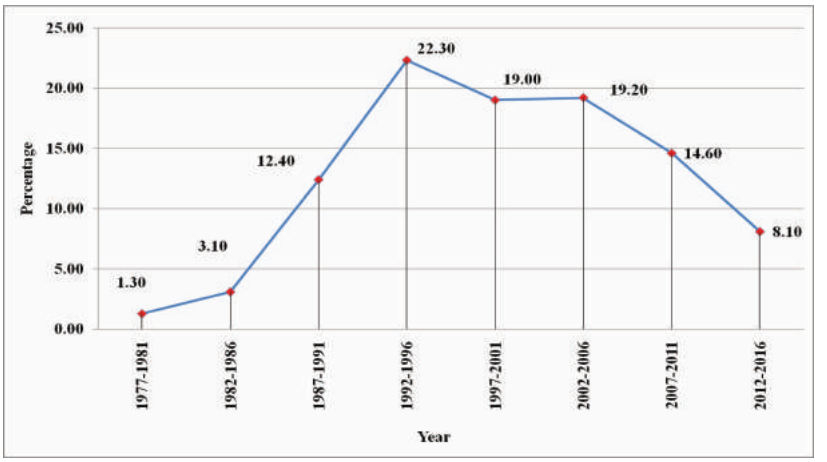


Figure 2 The growth rate of articles from 1977 to 2016

5.1 Relative Growth Rate (RGR)

Employed to detect a rise in the number of articles/pages per time unit. From the following equation, the mean Relative Growth Rate over a given period of the interval can be estimated (Hunt & Cornelissen, 1997).

$$\text{Relative Growth Rate (RGR)} = \frac{\log x_e w_2 - \log x_e w_1}{T_2 - T_1}$$

RGR = Average relative growth rate over the stated duration.

$\text{Log}x_e\ W_1$ =The initial number of papers is logged.

$\text{Log}x_e\ W_2$ = Log of the final number of articles a specific period of interval.

$T_2 - T_1$= The difference in units between the original time and the final time.

The year is taken as a unit of time for the RGR calculation in this study.

5.2 Doubling Time (DT)

The Doubling Time (DT) parameter is specifically correlated to RGR and signifies the time required to double the current volume for publications. Double Time is the exponential growth equation unit. The Doubling Time is computed as follows: Doubling Time = {(t2 - t1)* ln 2} / (ln c2 - ln c1). Again, in the per year growth case, the expression for Doubling Time can be written as: Doubling Time = ln 2/RGR

Table 2 Relative Growth Rate and Doubling Time of Indian Veterinary Journal

Year	Year Output	Cumulative total output	LogXeW1	LogXeW2	RGR	Block Period	Doubling time (Dt)	Block Period
1981	13	13	2.5649	2.5649	0		0	
1986	31	44	3.4340	3.7842	0.07004		9.894277449	
1991	124	168	4.8203	5.1240	0.060736	0.0608	11.40994619	−130.2326
1996	223	391	5.4072	5.9687	0.112307		6.170577318	
2001	190	581	5.2470	6.3648	0.223545		3.100042298	−137.1013
2006	192	773	5.2575	6.6503	0.278557		2.487823528	
2011	146	919	4.9836	6.8233	0.367936	0.2172	1.883480247	
2016	81	1000	4.3944	6.9078	−0.00125		555.8765749	

The data relating to the growing output of IVJ are presented in Table 2. To calculate the mean RGR and mean DT, the study period (1981-2016) is divided into two block periods, i.e. 1981-1996 and 2001-2016. The quantum output of IVJ has increased from 13 in the year 1981 to 81 in the year 2016, however, research publication is found to be maximum in the year 1996. It is therefore noted that the average RGR has increased in the second block from 0.0608 in the first block to 0.217. Mean DT, on the other hand, has decreased to -137.1013 in the second block from -130.2326 in the first block. Also, RGR has decreased from 0.070 in the year 1986 to -0.001 in the year 2016; correspondingly DT has gradually decreased from 9.8 to -555.8 in the same period.

Table 3 Authorship Pattern

S.No	Year	Single	Double	Three	Four	Five	Total	%
1	1977–1981	2	10	1	0	0	13	1.3
2	1982–1986	3	24	4	0	0	31	3.1
3	1987–1991	29	112	28	1	0	170	17
4	1992–1996	20	130	52	17	1	220	22
5	1997–2001	11	68	68	28	1	176	17.6
6	2002–2006	9	49	83	35	4	180	18
7	2007–2011	6	34	65	38	3	146	14.6
8	2012–2016	6	22	21	15	0	64	6.4
	Total	86	449	322	134	9	1000	100

Authorship patterns in Indian Veterinary Journal publications are shown in Table 3. It is known that 2531 authors, either single or multi-authored, published 1000 articles. It is evident from the table that 44.9% of publications contained double-authored articles, with 130 double authored articles published during 1992-1996. Multi-authored papers (5 authors) showed a declining trend (0.9%) during the study period. Further, single-authored articles accounted for 8.6% of the total. Far fewer papers were published during more recent periods (2007-2011) (2012-2016) than in the 1990s.

5.3 Authors´ Collaboration

Table 4 Degree of Collaboration during the study period

S.No	Year	Single Authors (Ns)	Multiple Authors (Nm)	Total	Degree of Collaboration
1	1977–1981	2	11	13	0.85
2	1982–1986	3	28	31	0.90
3	1987–1991	29	141	170	0.83
4	1992–1996	20	200	220	0.91
5	1997–2001	11	165	176	0.94
6	2002–2006	9	171	180	0.95
7	2007–2011	6	140	146	0.96
8	2012–2016	6	58	64	0.91
	Total	86	914	1000	0.91

In this study, where Degree of Collaboration $C = \frac{Nm}{Nm+Ns}$

C= 914/914+86 = 0.91%

The degree of cooperation C is therefore 0.91 percent. Statistics on the degree of collaboration between single-authored research and multi-authored research are provided in Table 4 during the study period. On a total of 1000 research articles, 86 were contributed by the single authors whereas 914 contributed by multi authors. This is high although a study on chemical sciences (Goyal et al., 2013) found a degree of collaboration of 0.97. "Is there a significantly higher probability for highly productive researchers to produce top-cited papers? Or, a sea of irrelevant papers is mainly produced by highly productive researchers. The response to these questions is important because it can help answer the question of whether or not there are perverse effects of increased competition and increased use of research evaluation and accountability focus metrics (Sandström & Besselaar, 2016). Highly active and cited researchers seem to have fresh prospects. Perceptibly, such researchers should be considered for different reasons, including policymaking and scholarly awareness of the related discipline (Klavans & Boyack, 2016). The decisive difference in this perspective, instead of counting publications and citations, is whether or not a researcher contributes to the limited number of very high-cited papers (Glänzel & Schubert, 1998). Table 5 shows statistics on the number of citations per article. It could be noticed from the table, around 1215 citations were received for publications during 1992-1996 and 22.31% citations during the year 1997-2001. However, as noted by Ioannidis et al. (2014), less than 1 percent of all researchers who published anything (indexed in Scopus) between 1996 and 2011 published in each of these 16 years, and that this limited set of core scientists is far more cited than others. What is noticeable is the decline in some citations for more recent periods - this may be due to a time lag between publication and citation but there are fewer papers published in more recent years, and that is likely to have an impact on the number of citations.

Table 5 Citation per Article

S.No	Year	No of Articles	Cited Articles	Percentages
1	1977–1981	13	78	1.60
2	1982–1986	31	134	2.76
3	1987–1991	170	1009	20.75
4	1992–1996	220	1215	24.98
5	1997–2001	176	1085	22.31
6	2002–2006	180	836	17.19
7	2007–2011	146	489	10.06
8	2012–2016	64	17	0.35
Total		1000	4863	100.00

In 1926, his pioneering article *The Frequency Distribution of Scientific Productivity* was published by Alfred J. Lotka (1926), in which he described a predictable pattern for the relative contributions of a body of authors to a body of literature. Out of the 2531 unique authors, Table 6 provides a list of the most productive authors.

Table 6 Core Authors Frequency

S.No	Name of the Author	No. of Articles	Cited Articles	Rank
1	A Kumar	15	122	1
2	S Kumar	11	42	2
3	M Singh	9	30	3
4	A Singh	8	53	4
5	KK Baruah	7	32	5

A. Kumar is the most productive of all, with 15 articles to his name, followed by S. Kumar holding the second rank with 11 articles. It is interesting to note an article entitled "Efficacy of some indigenous drugs in tissue-repair in buffalos (1993) authored by Kumar (1993) and two others had received 54 cites, and he had published research articles with 45 co-authors on various topics. Subsequently, S. Kumar and Khan et al. (2008) had contributed a lot of research work in this journal, coauthoring with 34 scholars and with 42 citations. Their research on "Prevalence of Phthirapteran ectoparasite on poultry (2008)" had received 10 citations. As reported by Ian Rowlands (2005), repeat-publishing authors are of explicit interest to publishers because they can be expected to submit manuscripts in the future as current and perceptibly please patrons, thereby guaranteeing an editor's access to a robust flow of research findings. Besides, these researchers transmit unintended advantages to the publisher of the journal, such as a position of advocacy within the academic environment, motivating their research students and colleagues to consider publishing with that journal and to subscribe and send to IVJ. A peer-reviewed study paper acts as a forum for disseminating the findings of a scientific inquiry, providing an opportunity to publicly uncover the work and support the available information for other researchers (Pendlebury, 2009). Other researchers may further validate, refute, or change the hypotheses in improving their research or clinical practice by consuming the findings of the analysis (Steele et al., 2006). As Christopher et al. (2014), noted, however, that publication data is merely a single chapter in an author's academic and research history. Publication data alone does not provide a full narrative of an author's effect or effects, nor is it necessarily reflective of meaningful empirical results that may have resulted from an author's investigation.

Lotka´s law = Productivity of Scientific Research

No.of Pairs	No.of Articles	No.of Authors Observed (Y)	Log Value Articles (X)	Log Value of Authors (Y)	X Y	X^2

$$\text{Lotka's Law} = n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)2}$$

N = Number of Pairs of Authors

X = Logarithm of article X

Y = Logarithm of Authors Y

More simply, the law can be expressed as $Y=C/X^n$, where X is the number of publications, Y is the relative frequency of authors with X publications, and n and C are constants, depending on the field, with n usually around 2. So for C=100, X=2, $Y=100/2^2$ (=25), and for X=3, $Y=100/3^2$ (11.1) (see Table 7).

Table 7 Author productivity of Indian Veterinary Journal established on Lotka's law

Number contribution	The observed number of Authors	Observed % Authors	Expected % Authors	Expected number of authors predicted by Lotkas Law (P)	(F–P) ²/P
1	1392	100.00	100.00	1392.00	0.00
2	520	37.36	25.00	348.00	85.01
3	276	19.83	11.11	154.67	95.18
4	124	8.91	6.25	87.00	15.74
5	90	6.47	4.00	55.68	21.15
6	54	3.88	2.78	38.67	6.08
7	21	1.51	2.04	28.41	1.93
8	8	0.57	1.56	21.75	8.69
9	9	0.65	1.23	17.19	3.90
11	22	1.58	0.83	11.50	9.58
15	15	1.08	0.44	6.19	12.56
Total	2531				

5.4 Applicability in the Indian Veterinary Journal of the data collection of Lotka´s law of author productivity

Table 7 describes the productivity of the authors produced during the study period by the PoP software application. Lotka's law is often referred to as the inverse square law, meaning that there is an inverse relationship between the number of publications and the number of writers writing these publications (Araú jo, 2014). The proportion of writers at different productivity levels is estimated by Lotka's law. Newby et al. (2003) presented empirical findings suggesting that the Lotka's law was not intended to predict a particular author's performance. Instead, its prediction lies in the cumulative and collective actions of a great number of authors. The versatility of Lotka's law has been essential in bibliometric studies since its introduction by Lotka (1926), and expanded over the years (Leydesdorff et al., 2013); Lotka estimated that the number of authors making x contributions is about 1/x of those making one and that the proportion of all those making a single contribution is 60%. This means that 60% of all writers in a given field will each have only one publication, 15% will each have two publications (1/22 times 60), 7% of authors will each have three publications (1/32 times 60) and only about 6% of authors will each produce up to 10 contributions in any field's literature. According to this data collection, out of 1000 papers contributed, a total of 2531 authors were interested. Where 1392 contributors have one article each (54.99 percent), 520 authors have contributed two articles (20.54 percent), 276 authors (10.90 percent) have three articles sponsored and 124 authors have four articles (4.89 percent) each and to credit, it demonstrates that all these values disprove Lotka's law at every stage. Furthermore, it is found that, as per the Lotka's law, the values observed do not correspond to the predicted values. Therefore, the findings of the analysis do not follow the Indian Veterinary Journal's Lotka's law of Author Productivity. In Figure 3, a graphical plot of Lotka's law on author productivity is presented.

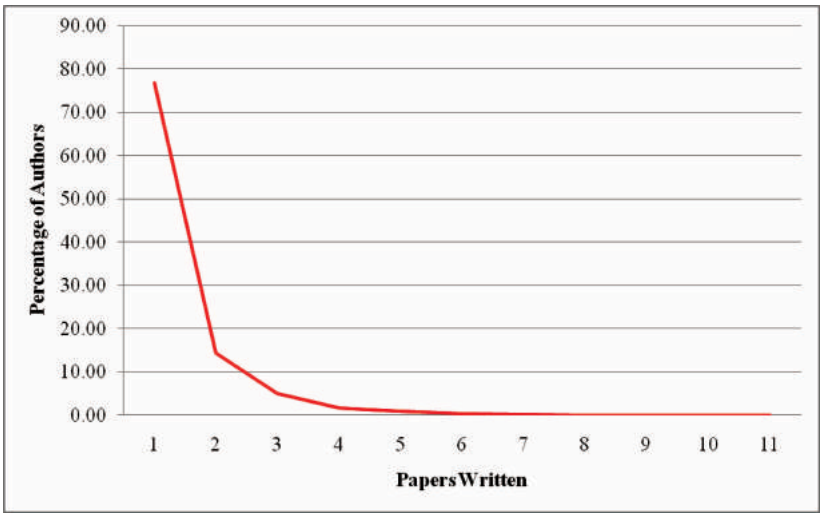


Figure 3 Graphical plot of the Lotka's law

6 Discussion

As indicated by Wan et al. (2009) bibliometric studies on the single journal:

- depicts the picture of the journal,
- Provides an interpretation that is above the trivial,
- Indicate the consistency, maturity and productivity of the journal irrespective of field/country/area,
- Offer details on studies that it supports

The IVJ is almost always considered significant in the veterinary community, valuable enough to be studied, to make decisions that the journal talks to authors who write in this field and represents the research activity in the field. During 2006, bibliometric studies showed that the IVJ was particularly concerned with animal health disciplinary articles and major junk of articles were from South India educational institutions. The report indicated that better communication between the researchers and the editorial offices of the journals in Chennai may be the explanation for this, and added 70% publications were from India and 28% were from other countries including Turkey, Iran, Malaysia, Korea and Poland (Vishwakarma, 2013). However, the present study showed a decline in publication output from 2007 (14.6) to 2016 (8.1). As noted by Rathinasabapathy et al. (2014), the NAAS (National Academy of Agricultural Sciences) rating of IVJ has come down from 6.00 (2011) to 4.33 (2015) and maintains the same position during the study period which is considered to be steep fall as it is one of the reputed journals. Nebelong-Bonnevie and Frandsen (2006) clarified that a big picture of that journal was given by single journal studies. To show the features, quality and status of the journal, the evaluation method used is mostly bibliometric indicators. Mahendra Kumar (2014) conducted a similar bibliometric analysis for the period 2011 to 2014 in the journal entitled "Library Herald." The research included several articles, the pattern of authorship, the distribution by topic of articles, the average number of references per article, the types of cited papers, the year-round distribution of cited journals, etc. The author pointed out that single journal studies represent the journal's merits and

limitations that would be useful for its further growth. In this analysis the decline in the number of papers published per year could be affecting the number of citations, and hence the wider impact of the journal. Authors have to pay (indirectly) to publish in the journal. Correspondingly Abdi et al. (2018) made a bibliometric analysis of the journal "Information Processing & Management (IP & M)" for the period from 1980 to 2015. Seglen (1997) notes that journal impact factors are contingent on their search field and high impact factors may be associated with journals covering a wide range of basic research with intensifying but unstable literature that uses many references per article and cites recent literature. Besides, the journal impact factor is regulated by article citation rates, not vice versa. Journal impact factors do not reflect individual journal articles statistically and are poorly associated with actual citations of individual articles.

The followings are several useful facts discovered from the analysis of the journal IVJ.

- The analysis displays a trend of growth in contributions published from 1992 to 1996 and of an average number of contributions per year is 125.
- The Number of documents cited per year is 124.69 with 4.86 citations per paper. It shows that during the study period from 1977 to 2016 dual authorship is the most frequent authorship arrangement.
- The mean number of authors per article was 2.97.
- The authorship pattern study aimed to identify the percentage of single, and multiple authorships. The results showed that the number of multi-authored articles increased rapidly and the degree of collaboration was found to be 0.91
- The findings of year-wise distribution of citations showed that a good number of citations was in 1992 to 1996 (1215 citations) followed by 1997 to 2001 with 1085 citations, and 1987 to 1991 with 1009 citations respectively.
- A.Kumar was found to be the most productive author with 15 publications and 122 citations followed by Kumar (11 publications and 42 citations).

Generally, Lotka's law determines the frequency of publications by writers in a given topic/discipline. In this paper, an attempt was made to analyze the applicability of the Lotka's law to journal publications instead of a subject or discipline. The findings acquired in this research do not comply with Lotka's law of author productivity as such. It may be due to long periods of research involvement, and maybe a changed Lotka's law may be a better match, as found in the nutrition report in Bangladesh (Ahmed & Rahman 2009).

7 Conclusion

This study may trigger more such research to evaluate an academic research journal and author productivity of those who published their work in this or another journal. Future research could be directed toward examining the patterns of collaborative authorship. For the journal itself, an understanding of the minimum number of high-quality papers required each year to ensure a reasonable impact factor would be desirable, and ways of encouraging authors to publish in the journal should be explored.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

Reference

Abdi, A., Idris, N., Alguliyev, R.M., & Aliguliyev, R.M.(2018). Bibliometric analysis of IP&M Journal (1980–2015). *Journal of Scientometric Research, 7* (1):54–62. http://doi.org/10.5530/jscires.7.1.8

Agarwal, A.,Durairajanayagam, D.,Tatagari, S., Esteves, S.C.,Harlev,A.,Henkel, R.,Roychoudhury,S.,Homa, S., Puchalt, N.G., Ramasamy, R., Majzoub, A., Ly, K.D., Tvrda, E., Assidi, M., Kesari, K., Sharma, R., Banihani, S., Ko, E., Abu–Elmagd, M., Gosalvez, J., & Bashiri, A.(2016). Bibliometrics: Tracking research impact by selecting the appropriate metrics. *Asian Journal of Andrology,18* (2), 296–309. http://doi.org/10.4103/1008–682X.171582

Ahmed S.M.Z., & Rahman M.A. (2009). Lotka´s law and authorship distribution in nutrition research in Bangladesh. *Annals of Library and Information Studies, 56*, 95–102.

Araújo, E. B., Moreira, A. A., Furtado, V., Pequeno, T.H.C., & Jr J.S.A.(2014). Collaboration networks from a large CV database: Dynamics, topology and bonus impact. *Plos One, 9* (3), e90537. https://doi.org/10.1371/journal.pone.0090537

Balaraman, N.(2018). Indian Veterinary Journal Details Vol 95. Available at: http://www.connectjournals.com/ivj

Borchardt, R., & Roemer, R.C.(2015). Meaningful metrics: *A 21st century librarian´s guide to bibliometrics, altmetrics and research impact.* Chicago, Illinois: Association of College and Research Libraries.

Bornmann, L., & Daniel, H.D.(2007). What do we know about the h index?. Journal of the American Society for Information Science, 58, 1381–1385. http://doi.org/10.1002/asi.20609

Braun, T., Glänzel, W., & Schubert, A.(2006). A Hirsch–type index for journals. *Scientometrics, 69*, 169–173. https://doi.org/10.1007/s11192–006–0147–4

Christopher, R., Carpenter, M.D., MSc, David, C., Cone, M.D., & Cathy, C.(2014). Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine, 21* (10), 1160–1172. https://doi.org/10.1111/acem.12482

Cronin, B., & Sugimoto, C. R.(Eds.).(2014). *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact.* MIT Press.

Dhiman, A.(2015). Bibliometrics to altmetrics: Changing trends in assessing research impact. *DESIDOC Journal of Library and Information Technology, 35*, 310–315. http://10.14429/djlit.35.4.8505

Dong, P., Loh, M., & Mondry, A.(2005). The "impact factor" revisited. *Biomedical Digital Libraries, 2* (1), 7. https://doi.org/10.1186/1742–5581–2–7

Egghe, L.(2006). Theory and practice of the g–index. *Scientometrics, 69* (1), 131–152. http://doi.org/10.1007/s11192–006–0144–7

Ezema, I.J., & Onyancha, O.B.(2016). A bibliometric analysis of Health and Medical Journals: Issues in Medical Scholarly Communication in Africa. *Serials Review, 42* (2), 116–128. http://doi.org/10.1080/00987913.2016.1182881

Garfield, E.(2006). The history and meaning of the Journal Impact Factor. *JAMA. 295* (1), 90–93. http://doi.org/10.1001/jama.295.1.90

Glänzel, W., & Schubert, A.(1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science, 14* (2), 123–127. https://doi.org/10.1177/016555158801400208

Goyal, V., Gupta, G.K., & Kumar, A.(2013). Authorship patterns and collaborative research trends in the field of chemical sciences. *International Journal of Information Dissemination and Technology, 3* (3), 184–186.

Gupta, D.K.(1987). Lotka´s law and productivity patterns of entomological research in Nigeria for the period, 1900–1973. *Scientometrics,12*, 33–46. https://doi.org/10.1007/BF02016688

Harzing, A.W.(2013). *Publish or Perish*. Retrieved from http://www.harzing.com/pop.htm

Hirsch, J.E.(2005). An index to quantify an individual´s scientific research output. *Proceedings of the National Academy of Sciences, 102* (46), 16569–16572. http://doi.org//10.1073/pnas.0507655102

Hunt, R., & Cornelissen, J.H.C.(1997). Components of relative growth rate and their interrelations in 59 temperate plant species. *New Phytologist, 135,* 395–417. http://doi.org/10.1046/j.1469–8137.1997.00671.x

Ioannidis, J.P.A., Boyack, K.W., & Klavans, R.(2014). Estimates of the continuously publishing core in the scientific workforce. *Plos One, 9* (7), e101698. https://doi.org/10.1371/journal.pone.0101698

Jesubright, J.J., & Saravanan, P.(2014). *A scientometric analysis of Global Forensic Science Research Publications*. Library Philosophy and Practice.

Jin, B.(2006). H–index: An evaluation indicator proposed by scientist. *Science Focus, 1* (1), 8–9.

Kear, R., & Colbert–Lewis, D.(2011). Citation searching and bibliometric measures: Resources for ranking and tracking. *College & Research Libraries News, 72* (8), 470–474. https://doi.org/10.5860/crln.72.8.8620

Khan, V., Kumar, S., Gupta, N., Ahmad, A., & Saxena, A.K.(2008). Prevalence of phthirapteran ectoparasite on poultry. *Indian Veterinary Journal, 85* (4), 447–448.

Klavans, R., & Boyack, K.W.(2016). Scientific superstars and their effect on the evolution of science. *ENID–Science and Technology Indicators Conference*.

Kumar, A. S., & Singh, H. P.(1993). Efficacy of some indigenous drugs in tissue repair in buffaloes. *Indian Veterinary Journal, 70*, 42–44.

Kumar, M.(2014). Library Herald Journal: A bibliometric study. *Journal of Education & Social Policy, 1*(2),123. http://jespnet.com/journals/Vol_1_No_2_December_2014/17.pdf

Kumar, N.(2010). Applicability to Lotka´s law to research productivity of Council of Scientific and Industrial Research(CSIR), India. *Annals of Library and Information Studies, 57* (1), 7–11.

Kumar, R.(2016). A scientometric analysis of IETE Technical Review Journal (2007–2014). *VSRD International Journal of Library & Information Science, II*. 27–32.

Leydesdorff, L., Rafols, I., & Chen, C.(2013). Interactive overlays of journals and the measurement of interdisciplinary on the basis of aggregated journal – journal citations. *Journal of the American Society for Information Science and Technology, 64* (12), 2573–2586. http://doi.org/10.1002/asi.22946

Lotka, A.J.(1926). The frequency distribution of scientific productivity. *Journal of Washington Academy of Sciences, 16*, 317–323.

Majhi, S., Chanda, J., & Maharana, B.(2016). *Content analysis of journal articles on Wiki in Science Direct Database*. Library Philosophy and Practice.

Michael, Norris, Charles, & Oppenheim.(2010). The h–index: A broad review of a new bibliometric indicator. *Journal of Documentation, 66*, 681–705. http://doi.org/10.1108/00220411011066790

Nandi, A., & Bandyopadhyay, A.K.(2008). Indian economic review (1998–2002): A bibliometric study. *SRELS Journal of Information Management, 45* (1), 95–100.

Narin, F.(1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ: Computer Horizons.

Nebelong–Bonnevie, E., & Faber Frandsen, T.(2006). Journal citation identity and journal citation image: A portrait of the Journal of Documentation. *Journal of Documentation, 62* (1), 30–57. https://doi.org/10.1108/00220410610642039

Newby, G.B., Greenberg, J., & Jones, P.(2003). Open source software development and Lotka´s law: Bibliometric patterns in programming. *Journal of the American Society for Information Science and Technology, 54* (2), 169–178. http://doi.org/10.1002/asi.10177

Notess, G.(2005). Scholarly Web searching: Google Scholar and Scirus. *Online, 29,* 39–41.

Pauly, D., & Stergiou, K.I.(2005). Equivalence of results from two citation Thomson ISI´s Citation Index and Google´s Scholar service. *Ethics in Science and Environmental Politics, 5*,33–35.

Pendlebury, D.A.(2009). The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis, 57* (1), 1–11. http://doi.org/10.1007/s00005–009–0008–y

Rathinasabapathy, G., Rajendran, L., & Kopperundevi, S.(2014). NAAS rating of Indian journals in the field of Veterinary and Animal Sciences: A study. *Asian Journal of Library and Information, 6*, 3–4.

Reuters, T.(2019). Thomson Reuters ESG Scores. *Thomson Reuters, February*.

Rowlands, I.(2005). Emerald authorship data, Lotka´s law and research productivity. *Aslib Proceedings, 57* (1), 5–10. https://doi.org/10.1108/00012530510579039

Sandström, U., & Besselaar, P.V.D.(2016). Quantity and/or quality? The importance of publishing many papers.

Plos One, 11 (11), e0166149. https://doi.org/10.1371/journal.pone.0166149

Schreiber, M.(2008). To share the fame in a fair way hm modifies h for multi–authored manuscripts. *New Journal of Physics, 10* (4), 1–8. http://doi.org/10.1088/1367–2630/10/4/040201

Seglen, P.O.(1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal(Clinical research ed.), 314* (7079), 498–502. https://doi.org/10.1136/bmj.314.7079.497

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y.(2007). Generalized Hirsch h–index for disclosing latent facts in citation networks. *Scientometrics, 72*, 253–280. https://doi.org/10.1007/s11192–007–1722–z

Sobrino, M.M.I., Caldes, A.I.P., & Guerrero, A.P.(2009). Lotka´s law applied to scientific production in the Information Science area. *Brazilian Journal of Information Science, 2* (1), 16–30.

Starbuck, W.H.(2005). How much better are the Most–Prestigious Journals? *The Statistics of Academic Publication Organization Science, 16* (2),180–200. http://doi.org/10.1287/orsc.1040.0107

Steele, C., Butler, L., & Kingsley, D.(2006). The publishing imperative: The pervasive influence of publication metrics. *Learned Publishing, 19*, 277–290. http://doi.org/10.1087/095315106778690751

Swain, D.(2014). International information and library review: A ten year bibliometric study. *International Information & Library Review, 46*, 113–124. http://doi.org/10.1080/10572317.2014.951589

Vishwakarma, M.L., Maurya, S.L., & Parashar, V.(2013). Indian Veterinary Journal and Indian journal of Animal Sciences: A comparative bibliometric investigation. *International Journal of Emerging Technology and Advanced Engineering, 3* (5).

Vlachy, J.(1978). Frequency distribution of scientific performance: A bibliography of Lotka´s law and related phenomena. *Scientometrics, 1* (1), 109–130.

Voos, H.(1974). Lotka and information science. *Journal of the American Society for Information Science, 25* (4), 270–272. http://doi.org/10.1002/asi.4630250410

Wan, K., Anyi, U., Zainab, A.N., & Anuar, N.B.(2009). Bibliometric studies on single journals: A review. *Malaysian Journal of Library & Information Science, 14* (1),17–55.

Wildgaard, L., Schneider, J. W., & Larsen, B.(2014). A review of the characteristics of 108 author–level bibliometric indicators. *Scientometrics, 101* (1), 125–158.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

Corpus construction and mining for Citation Context Analysis

Danqun Zhao, Qianying Guo*, Hongpu Chen, Zhujuan Cai and Xiangyu Wang

Department of Information Management, Peking University, Beijing, China

ABSTRACT

Citation Context Analysis (CCA) is a typical data-driven research field based on full-text information, which breaks the limitations of traditional citation analysis using only bibliographic data, and benefits further studies on various citation behaviors and other core issues behind them, such as citation motivation, citation function and citation sentiment. Corpus for CCA is the most important guarantee and support for these issues. This paper attempts to discuss the corpus construction and mining for CCA in order to comprehensively review the research significance, research status and existing deficiencies in this area. Two main sections in our paper are:1) corpus construction for CCA, its three building tasks, such as citation sentence extraction, citation-reference mapping and citation context extraction, are discussed; 2)corpus mining and utilization for CCA, following related topics or situations are explored, including classification of citation motivation (or behavior) and citation sentiment, indexing and retrieval based on citation, citation recommendation and evaluation, citation-based abstracting and review generation automatically, and domains knowledge metrics. Finally, some suggestions and future research directions are briefly listed.

KEYWORDS

Citation Context Analysis; Citation Content Analysis; Citation Corpus; Citation Analysis

1 Introduction

As one of the core research fields of Bibliometrics, Scientometrics and Informetrics (call these different research areas "information metrics" and abbreviate it as "iMetrics" (Milojecvic & Leydesdorff, 2013)), citation analysis has been studied for more than half a century since it was founded by E. Garfield in 1960s. Looking back, the research development of citation analysis has gone through the following four stages: ①Citation Analysis 1.0, the stage of citation counting by using papers or their bibliographic elements as the quantitative analysis units or objects (before 1970s); ②Citation Analysis 2.0, the stage of clustering analysis of bibliographical relationships, such as bibliographical coupling and co-citation analysis (from 1970s to 1990s); ③Citation Analysis 3.0, the stage of citation network analysis by using complex network theory and SNA tools(since 2000s); ④Citation Analysis 4.0, the stage of citation context (or content) analysis based on full-text information (since 2010s).

Citation Context Analysis (CCA) is one of new research frontiers of citation analysis. By using the full text of a citing paper, CCA tries to obtain and use all citation information about

*Corresponding author: guoqianying@126.com

every reference listed in the end of the citing paper, such as citation position or section, citation frequency (or strength) and citation context, and make the quantitative analysis of citation content on a finer granularity. The synonymous terms of CCA include Citation Content Analysis (Zhang et al., 2013), Full-text Citation Analysis (Liu et al., 2013), etc. Although these terms are different in expression, there is no obvious difference in what they are referring to.

Booming and fast-growing of CCA are mainly due to the joint influence and promotion by the following factors: ①the popularity of full-text database makes it no longer difficult to obtain full-text corpus, which lays the data foundation for CCA; ②the progress of NLP technology, such as text mining, sentiment analysis, Named Entity Recognition (NER), Knowledge Graph(KG) and various advanced machine learning algorithms, has provided strong technical support for CCA; ③the promotion of Open Citation and Initiative for Open Citation ("I4OC"). "I4OC" advocates semantic publishing and citation opening, and is committed to using semantic Web technology to publish and open citation information in RDF format, so that it can be easily tracked and accessed like Web links information and can be understood and used by machines."I4OC" ensures that citation data is exposed and accessed unrestricted in more disciplines (or fields), and the OpenCitations Corpus (OCC) created based on SPAR Ontologies (OpenCitations, 2020) can gradually alleviate the long-standing problems of citation data, such as being difficult to parse, inconvenient to track continuously and unable to be understood by machines, until it is completely solved.

Due to the limitation of bibliographic data, traditional citation analysis is always difficult to accurately identify various citation behaviors and their hidden (implied) motivations, purposes and sentiments behind them, and also difficult to effectively judge ecological environment and quality of citations in academic literature collection. The advent of CCA has greatly broken the constraint of traditional research and become a leading direction in the field of NLP-based Bibliometrics (Amjad et al., 2013).

As a kind of typical data-driven research, CCA's corpus construction is the most important guarantee and support for its research. Through the in-depth cognition and complete mining of the implied value of CCA corpus, it is not only to maximize the value of the corpus, but also to lead the innovation and development of citation analysis both in theory and methodology. This paper attempts to comprehensively discuss the construction, mining and utilization of CCA corpus, which is mainly divided into two sections: one is the construction or building of CCA corpus, and its three building tasks, such as citation sentence extraction, citation-reference mapping and citation context extraction, are analyzed respectively; the other is the mining and utilization of CCA corpus, five related research topics or situations are discussed, including classification of citation motivation (or behavior) and citation sentiment, indexing and retrieval based on citation, citation recommendation and evaluation, citation-based abstracting and review generation automatically, and domains knowledge metrics.

2 Corpus Construction for CCA

Before discussing the corpus construction for CCA, let us clarify the basic concepts and terminology related to it.

①Reference & Citation. These are two closely related concepts based on the citing and/or cited relationships among academic papers. They are also a pair of basic concepts in the entire field of citation analysis and many other concepts are defined or derived from them. Generally, "reference" refers to a cited paper, which usually appears in the reference list (at

the end or after the body of a paper) and is described or annotated in the standard format; "citation" refers to a phrase, clause or sentence where the citation marker (or reference anchor) is located when a reference is cited or mentioned in the body of a citing paper. There is a many-to-many relationship between them, that is, a reference has at least one or more citations in the body of a citing paper, and a citation will also correspond to or be associated with one or more references in its reference list. Due to this, for "reference", there are Unitary Mentioned Reference (UMR, a reference that be mentioned only once in a citing paper) and Multiple Mentioned Reference (MMR, a reference that be mentioned more than once in a citing paper); for "citation", there are Unitary Reference Citation (URC, a citation only includes one reference) and Multiple References Citation (MRC, a citation includes more than one reference), as shown in Figure 1 (Lin et al., 2019).

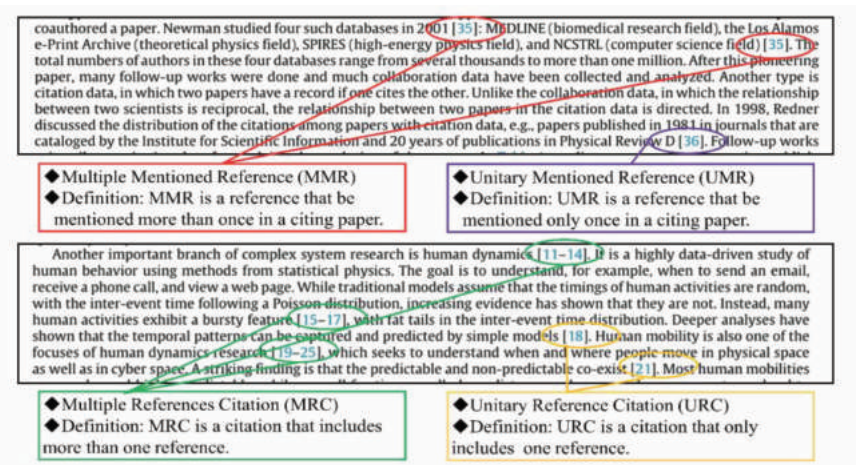


Figure 1 The Definition and example of MMR, UMR, MRC, URC

It should be noted that only bibliographic data can be used in early citation analysis while "citation" is often regarded as a synonym of "reference" (or cited paper). After entering the stage of full-text citation analysis (4.0), the differences (or semantic differences) between them gradually become clear. Many scholars have had in-depth discussions on "reference", "citation" and their semantic differences (Wouters, 1999), and the formation of the reference tradition and the establishment of the citation mechanism based on them have become the institutional guarantee for citation analysis. Robert K. Merton (1988), the founder of American sociology of science, emphasized that-- "We thus begin to see that the institutionalized practice of citations and references in the sphere of learning is not a trivial matter. While many a general reader--that is, the lay reader located outside the domain of science and scholarship--may regard the lowly footnote or the remote endnote or the bibliographic parenthesis as a dispensable nuisance, it can be argued that these are in truth central to the incentive system and an underlying sense of distributive justice that do much to energize the advancement of knowledge." Merton's view laid an important foundation for establishment of the Normative Theory of citation.

②Citation sentence. Nakov et al. (2004) first used a new term "citance" (abbreviation of "citation sentence") in 2004 to refer to the sentence surrounding the citation marker in a citing paper. "Citation sentence" can be understood in both narrow and broad sense, in which the narrow one is the sentence itself where the citation marker is located, and the broad can

be extended to the sentence where the citation marker is located and its surrounding text. In this paper, a complete sentence with a citation marker is taken as a citation sentence (or "explicit citation sentence"). Obviously, a citation sentence contains one or more citations, and each citation is associated with one or more references.

③Citation context. Small (2011) defined "citation context" as the text surrounding the references, which means the text content around the citation marker when a reference is cited in a paper (synonymous with the aforementioned generalized "citation sentence"). For the convenience of research, it is usually necessary to set a citation window to identify or extract citation context. Terms with the same meaning or similar to the "citation context" include "citation area", "citation site", "scope of influence of citation", and "citation statement", etc. In our paper, we use the term "citation context" uniformly and take the following understanding: the remainder of a specific citation window which removed the (explicit) "citation sentence" and some other sentences unrelated to this citation sentence. Ideally, the remainder has no sentences (or text) unrelated to this citation sentence, that is, all remaining sentences (or text) in the given window have a high semantic similarity with the specific citation sentence.

Normative Theory of citation believes that citation system is one of academic norms and basic professional standards that need to be consciously abided by scholarly community. Citation behavior can be regarded as an active way of communication (also known as "formal communication") in academic activities, which is important to maintain knowledge accumulation and discipline development from inheritance and transcendence diachronically to supplement and enrichment synchronically. From this theoretical perspective, citation, reference, citation sentence and citation context are all indispensable source of CCA corpus. Therefore, a fully functional CCA corpus should focus on the following three tasks for its construction: citation sentence extraction, citation-reference mapping and citation context extraction.

2.1 Citation Sentence Extraction

Citation sentence extraction is the primary task of CCA's corpus construction, and it is also the basis of the latter two. Generally speaking, the writing of peer-reviewed papers has strict requirements or description standards for the annotation of references (in or after the body of a paper). Therefore, it is not too difficult to identify and extract citation sentences from the full-text of citing papers in most cases, especially from those with structured full-text (such as XML format). Of course, if a paper with structured full-text cannot be obtained directly, extracting its citation sentences will be relatively difficult because such full-text needs to be parsed and preprocessed.

In fact, due to the different writing habits of authors, the diversity of description standards and differences in citation styles, etc. there are still many detailed problems to be solved in accurately identifying and extracting citation sentences from the full text, and the complexity of these details cannot be ignored. One of the most important problems is the identification of citation markers (or reference anchors), so we need to investigate citation styles first. The widely used citation styles are as follows: Numbered (which use numbers or other abbreviations to refer to an entry in the reference list) and Author-Date (which use an "author-year" pair to uniquely identify an entry in the reference list, also known as "Harvard Style"). Different journals or publishers make some adjustments for themselves based on these citation styles. For example, different conventions are involved in the Numbered style, such as whether numbers are superscripted? How to represent consecutive numbers? And how to

select the separators between discontinuous numbers? etc. Powley & Dale (2007) have made a more comprehensive investigation on the citation styles used in journal papers and found that there are five kinds of citation styles more representative (see Table1).

Table1 Some examples of citation styles

Citation styles	Examples
Textual Syntactic	**Levin (1993)** provides a classification of over 3000 verbs according to…
Textual Parenthetical	…are WordNet **(Miller et al., 1990)** and Levin classes **(Levin, 1993)**
Prosaic	**Levin** groups verbs based on…
Pronominal	**Her** approach reflects the assumption…
Numbered	…of behavior among verb groups [1]

Citation sentence extraction also frequently meets the problems of informal citations or implicit (non-explicit) citations (Powley & Dale, 2007; Qazvinian & Radev, 2010), which are more common especially when the Harvard style is used for marking (or annotating) citations. For example, two cases in Table 1 (the third and fourth), there are no obvious citation markers, instead of a person's name or personal pronoun used to describe or cite a specific reference. Many studies confirmed that such implicit citations often contain richer semantic information and have higher value in citation analysis (Athar & Teufel, 2012). Our paper adopts a narrow understanding of citation sentence, so here only focuses on extraction of explicit citation sentences, while the identification and extraction of implicit citation sentences will be discussed in Section 2.3.

After a citation sentence extracted, it can be stored in a database table and the fields of the table are as follows: citation sentence number (unique), citation sentence type, citation sentence content (text) and citation sentence source, etc. The "source" field can be further subdivided into several subfields to comprehensively record its number of citing paper, numbers of chapter (or section) position, paragraph and sentence where the citation sentence is located in the given citing paper.

2.2 Citation–Reference Mapping

The second construction task of CCA corpus is to scan each citation sentence stored in the database table and make mapping between each citation in the sentence and its corresponding reference (s) to form a citation-reference mapping record until all citation sentences are processed. Here the key problem is still the parsing of citation markers. Obviously, if a citation sentence contains only one URC citation, one citation-reference mapping record will be created; if a citation sentence contains more than one URC or MRC citations, multiple citation-reference records will be created. When all citation-reference mapping records of all citation sentences in a citing paper are stored in the database table, their metadata of citations for a citing paper can be aggregated. The main fields in citation-reference mapping table include: citation sentence number (unique), reference number (unique), reference author, reference title, reference publication year, reference source, etc. Furthermore, if reference (s) and citing paper are from the same literature database, more fields of reference(s), such as abstract, keywords, and author institutes, etc., can be considered to write into citation-reference mapping records.

Citation-reference mapping table is one of the important parts of CCA corpus, which can

provide effective supplementary or supporting information for accurately understanding their content and semantics of citation sentences, and evaluating rationality or interdisciplinary of citation behaviors, etc.

Normally, a citation appears in the body of a citing paper with reference anchor(s), and the corresponding reference (s) appears at the end of the citing paper (corresponding to the reference list). The common way to establish a citation-reference mapping record is to use regular expression to identify the various styles of citation markers located in the citation sentences. Figure 2 gives a simple example for the "Author-Date" citation style (Harvard style). It can be seen that how to accurately extract fields of each reference corresponding to citation (such as author, title, publishing year, journal, conference, etc.) from the reference list is quite complicated. When the Harvard style is used for citation marking, many NER problems will be involved due to the different abbreviations, translations and personal pronoun anaphora of author names, all which are fairly difficult to identify and parse accurately.

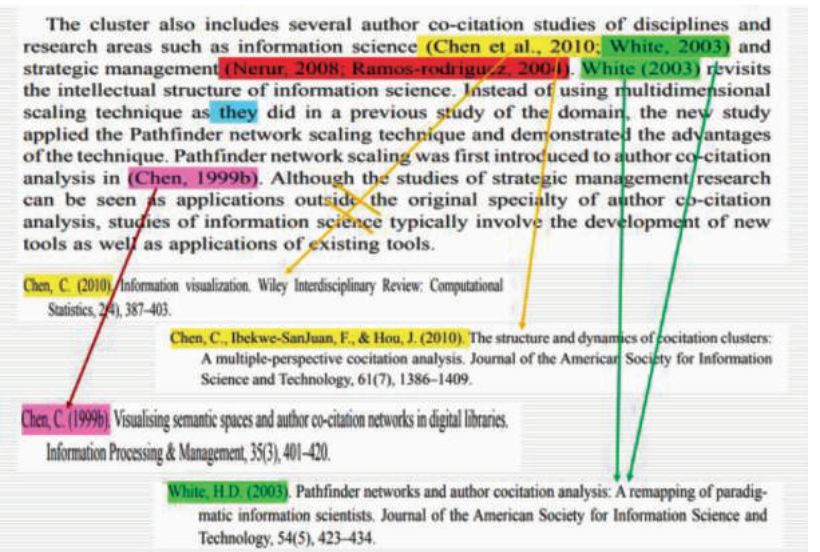


Figure 2 Examples of "Citation-Reference" Mapping

2.3 Citation Context Extraction

The third construction task of CCA corpus is to identify and extract its citation context around a citation sentence, or it can be understood as extraction of implicit citations. Current extraction strategies of citation context can be divided into the following three types: ①taking the whole paragraph where the citation sentence is located as its context; ②selecting a fixed number of sentences before and after the citation sentence as its context, or calculating the physical distances between them and the citation sentence, adding them different weights, and then considering whether to use them as context; ③calculating the semantic similarity between the citation sentence and its surrounding sentences (in given citation window), and selecting the sentences with high similarity as its context.

Among these strategies, the first two are simple and easy, while the third is the most ideal that requires a large computational cost and time consumption (deeply rely on NLP algorithms). Literature investigation found that there is no consensus on how to extract citation

context: the first two strategies are extensively used to resolve this problem; some directly extract citation sentences as an alternative; and a few researches also began to pay attention to the third strategy which tried to make elastic adjustments to the number of extracted sentences to improve extraction quality of citation context. For example, through calculating semantic similarity between each sentence (in the given citation window) and the citation sentence, the sentence(s) with high similarity (more than a certain threshold) but not necessarily adjacent to the citation sentence can be identified. Finally, the actual extraction result of context can contain different number of sentences by removing irrelevant text to the citation sentence and be divided into different types accordingly, such as No Context, Only Before (the citation sentence), Only After(the citation sentence), Both Before and After (the citation sentence).

Obviously, the challenge and complexity of accurately and completely extracting citation context far exceed the first two construction tasks of CCA corpus. Some researches specialized in the challenge tasks are in progress (Lei et al., 2016), and many other studies, such as citation function classification (Teufel et al., 2006a; Teufel et al., 2006b), citation summarization (Qazvinian & Radev, 2008; Qazvinian & Radev, 2010) and automatically generating review (Nanba et al., 1999; Nanba et al., 2011), also involved massively in it. Results of all these studies have further confirmed that CCA corpus can not only using (explicit) citation sentences. A lot of important information about author's attitude and comments on references (or cited papers) often appear somewhere around the citation sentences and their context information is very important for many CCA research topics.

2.4 Some Related Discussion

Though CCA has become active recently, related discussion focusing on its corpus construction is still rare. Three construction tasks discussed above are all important parts of a complete CCA corpus, among them extracting citation sentence is most critical, and the difficulty of citation-reference mapping and citation context extraction remains incrementally. They are all related to a large number of NLP technical issues, such as Named Entity Recognition (NER), anaphora resolution, definition and calculation of sentence similarity, dictionary building for clue words, and knowledge representment and extraction (for example semantic triple of SPO), etc., which require help of machine learning algorithms. Due to limited space, these issues can only be discussed in another paper.

We believe that the first choice for massive construction of CCA corpus is biomedical field in the current situation, some reasons list as follows: ①The full-text of academic papers in this field has a high degree of open access, especially the structured full-text. For example, only PubMed Central (PMC) has collected more than 3 million articles (in XML format), which is not only free and open, but also easy to analyze and use, which undoubtedly provides sufficient and high-quality sources for corpus construction. ②Research tools are abundant and relatively mature. There are many thesauri and their supporting tools, such as MeSH, UMLS and Medical Text Indexer (to extract medical subject terms), MetaMap (to extract UMLS concepts) (NLM, n.d.); Second, there are Semantic MEDLINE Database (SemMedDB) (Kilicoglu et al., 2012) which stored and represented in SPO triples and its supporting SPO extraction tool Batch SemRep (Kilicoglu et al., 2012); Third, there are tools of extracting citation sentences, such as Colil for extracting PMC citation sentences only and free AI tool Semantic Scholar for extracting beyond PMC. In addition, medical document feature modeling BioBERT (Lee et al., 2020) based on deep learning algorithm and the knowledge graph based on BioBERT (Xu et

al., 2005) are also in open and available. ③A large number of concepts, entities or knowledge objects (such as diseases, drugs, tissues/organs, genes, medical instruments, etc.) existing in this field can be used for bibliometric analysis, and semantic relations and their meanings among these concepts, entities or objects are also very rich and clear, which can provide enough research topics and application scenarios for CCA. Obviously, building its CCA corpus in biomedical field has significant benefits.

3 Corpus mining and utilization for CCA

Citation sentences and their context related to a specific research paper are very valuable because they are peers-reviewed text and have rich evaluation information among them. Especially for highly cited papers, such corpus of text continuously accumulated after a period of publication time has a huge amount of information and utilization value, which is worthy of in-depth mining and utilization. The corpus construction for CCA, on the one hand, can ensure the comprehensive extraction, analysis and storage of this kind of valuable text; on the other hand, it can effectively remove a large number of redundant fragments from full text and facilitate mining and efficient utilization.

After carefully considering and evaluating value of such citation corpus, five important research topics focusing on CCA corpus mining and utilization are listed and discussed as follows.

3.1 Classification of Citation Motivation (or Behavior) and Citation Sentiment

Accurate classification of citation motivation (or behavior) and citation sentiment is the primary research task of CCA, and it is also a difficult problem staying in the field of citation analysis for a long time. CCA corpus derived from full-text of papers makes it possible to solve above problems or achieve breakthroughs and plays an important role in reasonably differentiating weight or impact of every citation, modifying and improving research hypothesis and theory of citation analysis. At the same time, CCA corpus has also laid a solid foundation for a series of further researches, such as evaluation of quality, rationality and ecological environment of citations, as well as citation content-based academic evaluation, etc.

3.1.1 Classification of Citation Motivation (or Behavior)

Psychologists believe that motivation is an internal psychological process or impetus that leads to, inspires and maintains individual activities by goals or objects. The generation of motivation is mainly based on needs at various levels. Citation motivation is a kind of social motivation, mainly due to the needs of academic research. Furthermore, various motivational theories in psychology hold that motivations are the basis of most human behaviors, and there is a close relationship between motivations and behaviors. Therefore, citation behaviors can be regarded as an externalized expression of citation motivations. For the sake of convenience of discussion, the following does not make a strict distinction between them.

Citation motivations (or behaviors) is key to whether the research hypothesis of citation analysis is tenable and whether its theoretical basis is complete, so it has attracted great attention of scholars as early as 1960s. Through the investigation and observation, Garfield (1964), founder of citation analysis, first summarized citation motivations appearing in the writing of scientists' academic papers into the following 15 types: 1) paying homage to pioneers; 2) giving credit for related work (homage to peers); 3) identifying methodology, equipment, etc.; 4) providing background reading; 5) correcting one's own work; 6) correcting the work of others; 7) criticizing previous work; 8) substantiating claims; 9) alerting to forth-

coming work; 10) providing leads to poorly disseminated, poorly indexed, or uncited work; 11) authenticating data and classes of fact-physical constants, etc.; 12) identifying original publications in which an idea or concept was discussed; 13) identifying original publication or other work describing an eponymic concept or term as, e.g. Hodgkin's Disease, Pareto's Law, Friedel-Crafts Reaction, etc.; 14) disclaiming work or ideas of others (negative claims); 15) disputing priority claims of others (negative homage). Since then, many scholars have had more discussions on citation motivations (or behaviors) from different dimensions and different classification frameworks (Bornmann & Daniel, 2008; Brooks, 1985; Thorne, 1977; Weinstock, 1971). In recent years, domestic scholars have also introduced ecological perspective for classifying motivations of citations into different types, such as parasitism, mutualism, competitive symbiosis, commensalism, amensalism and irrelevant symbiosis (Li & Liang, 2012).

According to different research methods of citation motivations(or behaviors), existing related works or research outputs can be summarized into the following four categories: ①experience judgment of fields experts for simple classification of citation functions and motivations, most of early studies falls in this category; ②using questionnaire survey and interview to understand users' real citation motivations or behaviors; ③empirical research based on small-scale scientific literature data sets to verify or improve existing citation classification models; ④automatically identifying citation motivations (or behaviors). Among them, the fourth is one of the frontier topics in the CCA research which especially relies on the construction of large-scale citation corpus, and can effectively make up for the shortness and defects of the first three types of research.

Further literature investigation shows that the fourth type of research can also be roughly divided into rule-based, statistical-based and ontology-based methods. Among them, rule-based method is simple and effective, but it is time-consuming and laborious to build a base of rules manually by experts in advance, not easily shifting across different fields also leads to its poor flexibility; statistical-based method mainly uses various machine learning algorithms (such as Naïve Bayes, N-gram, Support Vector Machine, etc.) to train classifiers, which needs to build a manually annotated citation corpus in advance; ontology-based method needs to build an ontology using for citation classification description, there is now only Semantic Publishing and Ontologies (SPAR) can be compatible with ontology frameworks such as Citation Typing Ontology (CiTO) (Iorio, 2013; Shotton, 2010) and can be used for reference in practice.

In short, classification of citation motivations (or behaviors) is confronted with the difficulty of how to "enter the author's head", which a lot of problems of distinguishing psychological cognition and emotional attitude involves in. Now, various automatic recognition or classification methods are commonly encountered such corners as follows: inconsistent classification frameworks (no consensus or little agreement has not reached about it), poor corpus quality (small corpus, inaccurate and incomplete extraction of citation sentences and context, etc.), and algorithm limitations or heavy burden of manual annotation. For the future, it is necessary to strengthen the integrated utilization of different research methods, and how to construct a more comprehensive citation classification framework (or model) through extensive investigation or by reference to ontology tools such as CiTO is also key path in order to complete automatic classification of citation motivations (or behaviors) with higher accuracy according to some valuable cue words extracted from the citation sentence and its context.

Finally, it should be emphasized that explorations on citation motivations (or behaviors)

have always been as a central topic in field of citation analysis throughout all stages from 1.0 to 4.0. If summarizing all these studies theoretically, two schools or cliques for cognition about citation motivation gradually formed as below: one is the Normative Theory of citation, emphasizing that citation system is the academic norm abided by all scientists in practice, thus citations from peers represent obtaining recognition, and more citations represent more recognition. So, citation analysis as a method can be used to evaluate achievements and impacts of scientists and their works. The second is Social Construction of citation, which only regards citation as a rhetorical device (it has nothing to do with Merton's social norms theory). It holds that citations among papers are a kind of information utilization behaviors taken by individuals because of their perceived needs, and citation motivation is a complex, uncertain and private operation with certain propensity, so the usefulness of citation analysis is questionable. Many scholars have made a lot of theoretical discussions and experimental analyses around these topics (Baldi, 1998; Collins, 1999; Nicolaisen, 2003; Nicolaisen, 2007; Small, 1978; Small, 1998), some of them have tried to put forward new citation theories (Small, 2004; Nicolaisen & Frandsen, 2007). Up to now, although the two theories have their own achievements, their views or opinions about citation are in sharp opposition. How to eliminate or balance their disputes and seek their combination in the future has become an important research mission of CCA.

3.1.2 Classification of Citation Sentiment

Research for citation motivation affected by many subjective and objective factors is extremely complex, which has a natural difficulty in its accurate recognition. In contrast, the distinction of citation sentiment or emotions not only depends on the recognition of citation motivation, but also on the accurate extraction and correct understanding of the cue words representing various emotional attitudes in the citation corpus, which is also very challenging and difficult same as study for citation motivation.

Sentiment analysis is the task of identifying positive and negative opinions, sentiments, emotions and attitudes expressed in text. Although there has been a growing interest in this field in the past few years for different text genres such as newspaper text, reviews and narrative text, relatively less emphasis has been placed on extraction of opinions from scientific literature, more specifically, citations (Athar, 2011). Different from the highly personal emotional commentary texts on hot topics (or events) published by Web users, emotional expression is mostly considered to be relatively neutral when it comes to literature citation in academic papers, such as citing some facts or data, or objectively introducing the design idea and working principle of an algorithm. Only when it comes to the subjective evaluation of previous research, the more implicit and euphemistic emotional expression (positive or negative) will appear in the writing. CCA corpus (mainly involving citation sentences and their context) has rich evaluation information related to author's emotional expression for the cited paper. Therefore, the analysis or classification of citation sentiment can mainly use this part of the CCA corpus, but the difficulty and complexity of recognition of citation sentiment has been greatly increased undoubtedly due to the author's relatively cautious and careful wording.

The classification of citation sentiment has been involved in the research of citation context extraction and citation motivations (or behaviors) classification, etc. (Teufel et al., 2006a; Teufel et al., 2006b), and some studies adopted artificial methods directly (Yu, 2014).With rapid development of NLP technology and related machine learning algorithms, the research on automatic classification of citation sentiment is gradually developed. Its basic procedure

or main steps for experimental study can be described as follows:

①Construct or establish classification model (framework) of citation sentiment. One of the primary important challenges is how to define the implied sentiment in the citation corpus, including two dimensions of sentiment polarity and sentiment strength. Generally, the determinants of citation sentiment polarity are mostly related to citation motivation, and can be simply divided into three categories: positive, negative, and neutral; while citation sentiment strength can be divided into strong and weak. The elements combination of the two dimensions can form a preliminary citation sentiment classification model.

Similar to the classification of citation motivation belonging to the same category of psychological activities, there is no consensus classification framework for citation sentiment, especially the fine-grained framework. Therefore, a classification model that meets the requirements of citation sentiment recognition task can be constructed by referring to the relevant research results of citation motivation recognition.

②Based on the sentiment classification model and citation corpus, manually annotating citation sentiment and creating a training set by using annotation results. In fact, because most of the emotional expressions in citation corpus are implicit and euphemistic, the subjectivity of manual annotation is inevitable. How to ensure the quality and consistency of results of emotional annotation needs to be paid more attention to.

③Compile a dictionary of clue words for citation sentiment recognition. The compiling of the dictionary of clue words is a complicated NLP task, and there is no available one for citation sentiment recognition yet at present. HowNet (CNKI) is a universal emotional dictionary which is difficult to use directly because its coverage is not enough for fully covering specific fields. Therefore, it is necessary to start with the selection of seed words (mostly adjectives) and complete the compilation through continuous expansion of the set of seed words. It is worth noting that emotional words are not only adjectives, some nouns, adverbs, negative words and even transition conjunctions are all useful and important ones. For example, adverbs may be an important basis for judging sentiment strength, while transition conjunctions may directly determine or change the sentiment polarity of a sentence. So, it is very important for recognition task of citation sentiment in given field to building a fitting dictionary of clue words with high coverage (or wide range).

④Design or apply classification algorithm to complete the task of citation sentiment classification. The common machine learning algorithms which can be selected to use mainly include Naïve Bayes, Support Vector Machine (SVM), etc. According to the emotional score of each word in the emotional Dictionary (to be preset after compiling), emotional scores of all clue words existing in given citation sentence or its context can be used for calculating emotional score of the whole citation sentence or its context. Finally, all citation sentences and their context can be classified into different emotional categories by using their emotional scores respectively.

Deep learning algorithms have been developed and applied rapidly on various NLP tasks during recent years, but related research on citation sentiment recognition by these algorithms has not been founded easily. Relative lag of corpus construction for CCA is one of the major constraints because it is very difficult to estimate a large number of parameters that these deep learning algorithms demanded on existing small-scale citation corpus.

3.2 Indexing and Retrieval Based on Citation

Automatic indexing by using CCA corpus can optimize traditional indexing methods and

its results, and lay a foundation for citation retrieval, especially for citation context retrieval. Furthermore, the implementation of indexing and retrieval based on citation will accelerate the research of citation analysis.

3.2.1 Citation Indexing

The indexing value of citation corpus is mainly reflected in whether new keywords or subject words reflecting paper's content, theme and academic contribution can be extracted from such text or corpus, so as to add and enrich indexing terms obtained from its title, abstract and keywords in a given paper to a certain extent.

The research steps of citation indexing can be described as follows: ①selecting target papers (usually some highly cited papers) as a sample set of papers(D); ②for each paper dj in D, gathering all citation sentences and their contexts from each main body of its citing papers (set), and further extracting all keywords (set CKj) from these citation sentences and their contexts; ③comparing keywords in CKj with the original keywords of paper dj (set Kj, to be usually composed of keywords from title, abstract and some descriptors written by authors of dj), to find out whether or not new keywords are extracted in CKj? How many and their quality of these new keywords? Can they reveal the content or theme of the target paper as enriched indexing terms? ④Repeating steps② and③until all sample papers in D are finished.

It has been found that some new keywords with good indexing value can usually be extracted from citation corpus, and they are very useful for optimizing the characterization and disclosure of important contents of the target literature (Zhang et al., 2017). Generally speaking, the higher the cited times of a target paper, the more abundant citation sentences and contexts can be obtained, and correspondingly, the greater the possibility of finding additional index terms.

3.2.2 Citation Context Retrieval

The problem about citation context retrieval (or retrieval based on citation context)is described briefly below: according to the user's query or his/her search input of words, returning immediately online some citation sentences matched with this query or searching words in citing papers, and detailed information for each hit citation sentence in the answer set also includes its context (before or/and after the hit), position in the paper (for example IMR&D), all references occurred in the hit (with co-citation relationship), etc.

Objectively speaking, citation context retrieval does not require too many technical breakthroughs. It only needs to extract feature terms with index value (or significance) from citation sentences in CCA corpus and organize them into inverted files in order to matching user's queries. Hu (2016) has designed and implemented the SOS (search of sentence) system in his doctoral dissertation which search results can provide each hit citation and its following citation information: position (of chapter or section), context and all co-cited references. It is not only convenient for users to understand more citation details and realize the progress from "what to cite" to "how to cite", but also lay a research foundation for more fine granular co-citation analysis (for example, co-citation analysis at the level of single citation sentence, or single paragraph or section of papers, etc.).

Semantic Scholar, a free search tool launched by Allen Institute for AI (AI2) in 2016, also provides some citation retrieval services, such as References (referenced by this paper), Citations (citing this paper), and Figures, Tables and Topics (search for charts and their titles in given paper). According to recent visit and investigation, Semantic Scholar has collected nearly 200 million papers in its database, which can be used as a citation extraction tool in

addition to search service. Compared with Colil (DBCLS, n.d.), an OSS (open source software) that only aims at extracting citation sentences from abstract of articles in PMC, Semantic Scholar can be applied to more literature databases except for PMC, such as Microsoft Academic, Springer, Nature and ArXiv, etc.

3.3 Citation Recommendation and Evaluation

Citation recommendation is mainly devoted to providing timely and effective help for the author's academic paper writing and successful publication, including immediate recommendation and evaluation recommendation. The former serves for the author's academic paper writing, and provides a list of suitable citing reference(s) related to the current content written online by author, while the latter serves to comprehensively evaluate the quality of references cited in the paper after finishing it, including relevance, comprehensiveness, academic quality of these references; citation ecological environment of the written paper, and gives some comments (such as reservation/ deletion, ecological health/sub-health/pollution, etc.) or provides important references that failed to cite or omitted by author(s). Both the former and the latter need to be based on an accurate understanding of the content or semantic relationship between the cited papers (references) and citing papers, especially the former, which is more dependent on high-quality citation corpus, and is also an important source of high-quality citation corpus. At present, some service systems with citation recommendation function have appeared. For example, after uploading a paper in PDF format or its URL in Website, such system like Citeomatic, can return a list of references already cited by author (s) and a list of papers which are deserved to be cited by Citeomatic but not yet cited by author(s).

Citeomatic is a typical citation evaluation and recommendation tool. Although its function is limited (no immediate recommendation now), there is a large room for service enrichment or function expansion for expected citation evaluation in future, especially for academic ecological assessment based on CCA corpus. For example, given citing paper(s)(single or collective) and then obtaining all citation sentences and their positions occurring in the citing paper(s), the overall judgment about the single paper or macro monitoring of ecological environment about the collection of papers can be considered comprehensively by analysis of multi-dimensional influencing factors, including citation motivations (or behaviors), citation relevance(combined with using citation-reference mapping records in CCA corpus), quantity and quality of their citations, etc. Finally, some further comments on the citing paper(s) can be concluded, or evaluation results of classification can be given, like "Health", "Sub-Health", "Pollution", etc. This undoubtedly has a great influence and guidance on creating healthy academic environment (or atmosphere), developing academic research activities orderly and improving current mechanism of academic evaluation drastically.

3.4 Citation–Based Abstracting and Review Generation Automatically

An abstract refers to a brief and accurate description of the content of a document (or a document unit), and usually does not include the supplement, explanation or comment to the original. There are various types of abstracts, two kinds of manually writing ones are informative abstract and indicative abstract, while automatically generating ones can be divided into more types according to different standards. For example, according to whether original text is used, it can be divided into Full text-Based Abstract (FBA) and Citation-Based Abstract (CBA); according to whether it is oriented to specific users, it can be divided into

Generic Abstract and Biased Abstract, and the latter can be further subdivided into different subtypes, such as biased document themes, biased user interests and biased user's queries; according to the number of documents processed, there are Single Document Abstract (SDA) and Multiple Document Abstract (MDA). In particular, when the number of documents to be automatically summarized reaches or exceeds a certain threshold level, MDA can be regarded as automatically generating survey or review. Finally, the following four kinds of abstracts based on different research strategies are often referred to, such as statistical-based abstract (or excerpt), NLP-based abstract, information extraction-based abstract and text structure-based abstract.

As one of the important research topics emerged in the field of NLP since 1950s, automatically abstracting has always been focused on the generation technology and method of FBAs for a long time. Although these FBAs (including the author's abstracts) can better reflect the content of the original text, their ability to summarize the influence of the original text is relatively limited, and they can't reflect the diachronic changes of the literature influence after publishing (Mei & Zhai, 2008). So, CBA emerged gradually as a new research and exploration direction since 2008 (Qazvinian & Radev, 2008). It mainly learns from theory of citation analysis in Bibliometrics and uses the citation corpus (especially citation sentences and their contexts) to form a generalization or understanding about main content of a paper to be abstracted which can reflect its academic influence or value from the views of other peer's researches.

All existing theoretical and empirical analyses show that CBA has many advantages compared with its FBA (Bradshaw, 2003; Elkiss et al., 2008; Kan et al., 2002; Mohammad et al., 2009). Especially for the highly cited papers, their citation abstracts are more obvious in objectivity and diversity. In fact, their citation corpus is more general, extended and critical after ever-increasing accumulation and processing by more and more academic peers which can better reflect the significant parts of the original text.

A typical research task for CBA in early time can be described as below: given a paper to be abstracted, gathering its citing papers completely and extracting all citation sentences and their contexts (and regard as a set of citations); then selecting some citation sentences from the set to generate its citation abstract, and ensure that these selected sentences as a subset has sufficient compression rate and good generalization ability. Its main research steps (or key issues involved in)are listed as follows: ① selecting an appropriate full-text database (for its coverage and availability); ②identifying and extracting citation sentences (in broad or narrow sense); ③ classifying and screening these citation sentences by identifying their types and purposes; ④organizing or ranking citation sentences to form an abstract (as a draft); ⑤ post-processing of abstract draft, such as de-duplication, coherence processing for sentences, etc.; ⑥evaluating abstract.

Obviously, CBA's study is closely related to the task of construction of CCA corpus. The key points and difficulties are mainly derived from the second, third and fourth steps. The existing problems are the lack of structured full-text corpus (except PMC), accurate identification or extraction of citations and contexts in non-fixed citation window, semantic de-duplication and reasonable ranking of citation sentences, and new design for suitable evaluation indicators and schemes (such as for evaluation of sentence coherence). Up to now, many kinds of CBAs still focus on the generation of single document abstract (SDA). It should be considered how to extend from SDA to multiple document abstract (MDA) and to generate literature review automatically, which will lead to all old problems, such as classification/cluster-

ing, de-duplication and ranking of citation sentence, becoming more difficult and challenging. Furthermore, the combination of CBA and FBA is also an important research topic because CBA will be greatly restricted or impracticable for mostly low cited papers in academic collections due to the lack of their citation corpus.

3.5 Domains Knowledge Metrics

Metrics and analysis of scientific knowledge has attracted wide attention in recent year, and CCA corpus has become an important guarantee for such researches which are promoting development of iMetrics to Knowledge metrics gradually. Two basic problems need to be solved in knowledge metrics: one is knowledge representation, which defines the knowledge entity (or object) and its representation method for quantitative analysis, that is, determining the basic knowledge unit; the other is meta-knowledge representation, which is used to represent its sources and cognitive states for a given knowledge entity (object) or basic knowledge unit, like known, unknown, etc.

Take an example of the field of medicine. To solve the problem of knowledge representation, there are now three representative research achievements: ①SemMedDB of National Library of Medicine (USA) (Kilicoglu et al., 2012), which transforms the free text describing medical knowledge that can be understood by human into "Subject-Predication-Object" (SPO) triples that can be understood easily by machine, and maps all concepts and semantic relations involved in SPOs to the UMLS (Universe Medical Language System). ②Nano publication model, proposed by The Netherlands Bioinformatics Centre (NBIC), and here "nano publication" refers to the smallest and machine-readable publication unit with scientific significance (Groth et al., 2010). ③ Micro publication model proposed by Harvard Medical School (Clark et al., 2014). Among the three representations (or models) of medical knowledge described above, SPO Triples of SemMedDB is the most influential and most concerned. SemMedDB implements extraction and storage of knowledge units (SPO triples) in very large scale (the latest version contains nearly 100 million SPO triples extracted from abstract of papers in database of PubMed), but has the disadvantage of ignoring or missing most of the meta-knowledge of each SPO triple. This involves the second basic problem, the representation of meta-knowledge. Here, meta-knowledge mainly refers to some scientific judgments related to the cognitive state for a given SPO triple, such as whether the knowledge represented by the SPO triple is pure research hypothesis, or the conclusion of new experiments, or even only objectively citing previous scientific assertions (or opinions), etc. Scientists usually express their academic opinions through carefully selected words in the text, or give a more rigorous explanation and conclusion for it. This type of information or clues expressed in the text, which can provide the certainty (or uncertainty) degree of the knowledge represented in a SPO triple, is abundant in citation sentences or context, especially in medical field (Chen et al., 2018; Murray et al., 2019). Therefore, the representation of meta-knowledge can be considered starting from some valuable clue words in citation sentences or context. By accurately identifying and extracting these clue words and adding them to their corresponding SPO triples, a complete representation of scientific knowledge can be formed and then used for quantitative analysis and evaluation of different knowledge units.

Most researches on the topic of knowledge metrics heavily focuses on the field of biomedicine, and uncertainty measurement of medical knowledge is particularly active among them. For example, a series of studies completed by H. Small et al. (Small & Klavans, 2011; Small, 2018; Small et al., 2019; Small, 2019), proposed a new method to identify scien-

tific breakthroughs in scientific literature by combining co-citation analysis and citation context, and made empirical analyses from various aspects by the indicator of hedging rate. Here, "hedging rate" is defined as the proportion of citation sentences containing words "may", "could" or "might" in all citation sentences in a paper. It is mainly used to quantify the uncertainty of scientific knowledge contained in different types of papers and their citation sentences.

Besides these, the team of Chen Chaomei has also made fruitful findings on the measurement of knowledge uncertainty in the medical field. At the end of 2017, they published their work——"Representing Scientific Knowledge: The Role of Uncertainty" (Chen & Song, 2017), proposed that uncertain information should be regarded as the meta-knowledge of scientific proposition (SPO triples), and that hedging words can't cover all aspects of knowledge uncertainty, emphasized that scientific knowledge in the state of inconsistency, contradiction or controversy is an important driving force for the emergence of new paradigm or scientific reform. H. Small commented that "uncertainty is key to understanding the development of scientific knowledge", "opens up a new area in the study in Scientometrics and Informetrics as well as information visualization, namely the study and measurement of uncertainty of scientific knowledge and how uncertainty is expressed in scientific texts". Domestic scholars such as Du Jian also regard knowledge uncertainty as an important research front (Du, 2019; Du, 2020), and are committed to promoting the development of medical knowledge metrics.

Taking "uncertainty" as the core of topic, our paper believes that CCA and its corpus can also make more researches on domains knowledge measurement or metrics, here are some examples:①analysis of spatial-temporal evolution of knowledge (or visualization), combined with the publication time and authors' affiliations (institutions or countries/regions belong to, etc.) of papers associated with SPO triples, we can observe the change (curve) of knowledge uncertainty from dimensions of time and/or space, and then make judgment on the maturity, development and evolution, and key turning points of scientific knowledge in given triples;②knowledge interaction /cross- analysis, exploring the transformation relationship between scientific knowledge (in papers) and technology (in patents), or between scientific knowledge (in papers) and clinical treatment (in patients' cases). ③construction and analysis of meta-knowledge graph. Here, the meta-knowledge graph (MKG) uses each SPO triple as its node and the relationship between triples of SPOi and SPOj as its edge, it can be understood as a new high-level knowledge graph established on the basis of underlying knowledge graph, in which each "S" or "O" represents different node and each "P" as an edge of nodes. In a specific MKG, the meaning of edge of (SPOi, SPOj) can be defined by using their co-occurrence or co-citation relationship while taking the frequency of co-occurrence or strength of co-citation as its weight of the edge. Furthermore, the underlying knowledge graph (composed by knowledge entities and their semantic relations) and its meta-knowledge graph (MKG) can be integrated to build a two-layer (or even multi-layer) heterogeneous network. It is very noteworthy for the feasibility and significance of information science for studying such two-layer (or multi-layer) heterogeneous network in the future.

4 Conclusion

CCA is booming and developing rapidly at home and abroad during recent years. On the one hand, the large-scale and high-quality corpus of citations is far from being built and completed now, and the existing tools of extracting citation sentences can only provide very

limited data support for CCA; on the other hand, some basic issues, such as the classification of citation motivation and sentiment, etc., have not yet reached a research consensus, more research topics of CCA are waiting to be explored and expanded, which obviously fails to promote and drive corpus construction at once. In fact, the corpus construction and mining for CCA are two intertwined and closely related issues. This paper only discussed them from a macro and holistic perspective. In the future, on the basis of clarifying basic concepts of CCA, it is necessary to make more pragmatic and technical discussion on more detailed levels in close combination with our research jobs supported by fund of NSFC in order to promote the continuous and in-depth development of citation analysis in open full-text era.

References

Abu–Jbara, A., Ezra, J., & Radev, D.(2013). Purpose and Polarity of Citation: Towards NLP–based Bibliometrics. In: *Proceedings of the Main Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 596–606.

Athar, A.(2011). Sentiment Analysis of Citations using Sentence Structure–Based Features. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 81–87.

Athar, A., & Teufel, S.(2012). Detection of implicit citations for sentiment detection. In: *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, 18–26.

Baldi, S.(1998). Normative versus social constructivist processes in the allocation of citations: A network–analytic model. *American Sociological Review*, 829–846.

Bradshaw, S.(2003). Reference directed indexing: Redeeming relevance for subject search in citation indexes. *Research and Advanced Technology for Digital Libraries*, 499–510.

Brooks, T. A.(1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36* (4), 223–229.

Bornmann, L., & Daniel, H.(2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64* (1), 45–80.

Chen, C., Song, M., & Heo, GE.(2018). A scalable and adaptive method for finding semantically equivalent cue words of uncertainty. *Journal of Informetrics, 12* (1), 158–80.

Chen, C., & Song, M.(2017). *Representing Scientific Knowledge: The Role of Uncertainty*. Springer.

Clark, T., Ciccarese, P. N., & Goble, C. A.(2013). Micropublications: A semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics, 5* (1), 28.

Collins, H. M.(1999). Tantalus and the aliens: Publications, audiences and the search for gravitational waves. *Social Studies of Science, 29* (2), 163–197.

Du, J.(2019). An Automated Approach for Extracting Uncertain Clinical Knowledge from Published Medical Documents. In: *Proceedings of the 2019 Tianfu International Forum on Scientometrics and Research Evaluation*, Chengdu, China.

Du, J.(2020). Measuring Uncertainty of Medical Knowledge: A Literature Review. *Data Analysis and Knowledge Discovery, 46*, 14–27.

Elkiss, A., Shen S., & Fader, A.(2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology, 59* (1), 51–62.

Garfield, E.(1964). Can citation indexing be automated? *Statistical association methods for mechanized documentation, symposium proceedings*. National Bureau of Standards, Miscellaneous Publication 269, Washington DC, 189–192.

Groth, P., Gibson, A., & Velterop, J.(2010). The anatomy of a nanopublication. *Information Services & Use, 30*, 51–56.

Hu, Z.(2016). F*ull–text Citation Analysis: Theory, Method and Application*. China Science Publishing & Media Ltd.

Iorio, A. D., Nuzzolese, A. G., & Peroni, S.(2013). Towards the Automatic Identification of the Nature of Citations. In: *Proceedings of 3rd Workshop on Semantic Publishing*, 63–74.

Kan, M.Y., Klavans, J. L., & Mckeown, K. R.(2002). Using the annotated bibliography as a resource for indicative summarization. In: *Proceedings of LREC*, 1746–1752.

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Shin, D.(2020). Broad–Coverage Biomedical Relation Extraction with SemRep. *BMC Bioinformatics, 21* (1), Article No.188.

Kilicoglu H, Shin D, Fiszman M., & Shin, D.(2012). SemMedDB: a PubMed–scale repository of biomedical semantic predications. *Bioinformatics, 28* (23), 3158–3160.

Lee, J., Yoon, W., & Kim, S.(2020). BioBERT: a pre–trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36* (4), 1234–1240.

Lei, S., Chen H., Huang, Y., & Lu, W.(2016). Research on Automatic Recognition of Academic Citation Context. *Library and Information Service, 60* (17), 78–87.

Li, Z., & Liang, Y.(2012). The ecology explanation on citation motivation. *Studies in Science of Science, 30* (4), 487–494.

Lin, G., Hou, H., & Hu, Z.(2019). Understanding Multiple References Citation. In: *Proceedings of 17th International Conference on Scientometrics & Informetrics*, 2347–2357.

Liu, X., Zhang, J., & Guo, C.(2013). Full–text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology, 64* (9), 1852–1863.

Mei, Q., & Zhai, C.(2008). Generating Impact–Based Summaries for Scientific Literature. *Association for Computational Linguistics*, 816–824.

Merton, R. K.(1988). The Matthew Effect in Science II. Cumulative Advantage and the Symbolism of Intellectual Property. *ISIS, 79* (299), 606–623.

Milojecvic, S., & Leydesdorff, L.(2013). Information metrics(iMetrics): A research specialty with a socio–cognitive identify? *Scientometrics, 95* (1), 141–157.

Mohammad, S., Dorr, B., & Egan, M.(2009). Using citations to generate surveys of scientific paradigms. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 584–592.

Murray, D., Lamers, W., Boyack, K., Larivière, V., Sugimoto, C. R., & Van Eck N. J. (2019). Measuring disagreement in science. *17th International Conference on Scientometrics and Informetrics*, 2370–2375.

Nakov, P. I., Schwartz, A. S., & Hearst, M.(2004). Citances: Citation sentences for semantic analysis of bioscience text. In: *Proceedings of the SIGIR´04 workshop on Search and Discovery in Bioinformatics*, 81–88.

Nanba, H., & Okumura, M.(1999). Towards multi–paper summarization using reference information. *International Joint Conference on Artificial Intelligence*, 926–931.

Nanba, H., Kando, N., & Okumura, M.(2011). Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online, 11* (1), 117–134.

Nicolaisen, J.(2003). The social act of citing: Towards new horizons in citation theory. In: *Proceedings of the American Society for Information Science and Technology, 40* (1), 12–20.

Nicolaisen, J.(2007). Citation analysis. *Annual review of information science and technology, 41* (1), 609–641.

Nicolaisen, J., & Frandsen, T. F.(2007). The handicap principle: a new perspective for library and information science research. *Information Research, 12* (4), 12–14.

Powley, B., & Dale, R.(2007). Evidence–based information extraction for high accuracy citation and author name identification. In: *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering*, 618–632.

Qazvinian, V., & Radev, D. R.(2008). Scientific paper summarization using citation summary networks. In: *Proceedings of the 22nd International Conference on Computational Linguistics*, 689–696.

Qazvinian, V., & Radev, D. R.(2010). Identifying non–explicit citing sentences for citation–based summarization. In: *Proceedings of the 48th annual meeting of the association for computational linguistics,* 555–564.

Shotton, D.(2010). CiTO, the citation typing ontology. *Journal of Biomedical Semantics, 1* (Suppl 1), S6.

Small, H.(1978). Cited documents as concept symbols. *Social Studies of Science, 8* (3), 327–340.

Small, H.(1998). Citations and consilience in science–Comments on theories of citation?. *Scientometrics, 43* (1), 143–148.

Small, H.(2004). On the shoulders of Robert Merton: towards a normative theory of citation. *Scientometrics, 60* (1), 71–79.

Small, H.(2010). Referencing through history: how the analysis of landmark scholarly texts can inform citation theory. *Research Evaluation, 19* (3), 185–193.

Small, H.(2018). Characterizing highly cited method and non–method papers using citation contexts: The role of uncertainty. *Journal of Informetrics, 12* (2), 461–480.

Small, H.(2019). What makes some scientific findings more certain than others? A study of citing sentences for low–hedged papers. In: *Proceedings of the 17th International Conference of the International Society for Scientometrics and Informetrics*, 554–560.

Small, H., Boyack, K. W., & Klavans, R.(2019). Citations and certainty: a new interpretation of citation counts. *Scientometrics, 118* (3), 1079–1092.

Small, H., & Klavans, R.(2011). Identifying scientific breakthroughs by combining co–citation analysis and citation context. In: *Proceedings of 13th International Conference of the International Society for Scientometrics and Informetrics*, 783–793.

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

Teufel, S., Siddharthan, A., & Tidhar, D.(2006b). Automatic Classification of Citation Function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103–110.

Thorne, F. C.(1977). Citation index–another case of spurious validity. *Journal of Clinical Psychology, 33* (4), 1157–1161.

Weinstock, M.(1971). *Citation indexes, Encyclopedia of Library and Information Science*. New York: Marcel Dekker.

Wouters, P. F.(1999). The citation culture. [Doctoral dissertation, University of Amsterdam]. UvA–DARE(Digital Academic Repository). https://garfield.library.upenn.edu/wouters/wouters.pdf

Xu, J., Kim S., Song M., & Jeong, M.(2020). *Building a PubMed knowledge graph*. Retrieved from https://arxiv.org/abs/2005.04308

Yu, B.(2014). Automated Citation Sentiment Analysis: What Can We Learn From Biomedical Researchers. In: *Proceedings of the American Society for Information Science and Technology*. https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/meet.14505001084

Zhang, G., Ding, Y., & Milojevic, S.(2013). Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. *Journal of the American Society for Information Science and Technology, 64* (7), 1490–1503.

Zhang, S., Liang, M., & Cao, G.(2017). Research on Subject Extraction of Scientific and Technical Documents Based on Citation. *Information studies: Theory & Application, 40* (6), 122–127.

DSI DATA SCIENCE AND INFORMETRICS
VOL.1, NO.1, FEB. 2021

The Logarithmic Eigenfactor: Solving the Problems with the Normalized Eigenfactor

Liping Yu[a], Xinwen Long[b*]

a School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China
b Library, Southeast University, Nanjing, China

ABSTRACT

Since the introduction of the Normalized Eigenfactor as a journal influence factor in 2009, there has been little research into potential problems with this measure. In order to explore and resolve drawbacks associated with the Normalized Eigenfactor, this paper begins by proving that the discriminability realized by this method can be improved upon. By using the JCR2016 mathematics journals as an example, an analysis from the perspective of the discriminative degree and data distribution is performed to compare the Eigenfactor Score with that of the Normalized Eigenfactor. This is done using the Median Maximum Value Ratio, High Score Ratio, Low Score Ratio, Passing Rate, Discrete Coefficient, HHI and Jarque-Bera Test values. The results of the study show that the Normalized Eigenfactor had little effect on the discrimination and data distribution over the Eigenfactor. As such, the published accuracy of Eigenfactor Scores is misleading in claims that the Normalized Eigenfactor can improve the discriminating degree. In reality, it only becomes significant when magnifying the mean value of the Eigenfactor Score by more than 100 times. The Normalized Eigenfactor is a nonlinear transformation, and it will slightly degrade the information captured by the Eigenfactor Score. When the Normalized Eigenfactor is converted, the numerator is the journal's Eigenfactor Score, and the denominator is derived from other journals' Eigenfactors; therefore, the scale of the measurement is not fixed and is at odds with the basic principle of measurement. Furthermore, the divergence between indicator values and evaluation attributes of the Normalized Eigenfactor manifests as data distribution bias, a low Pass Rate, and low sub-area data congestion. On this basis, this paper proposes to replace the Normalized Eigenfactor with the Logarithmic Eigenfactor.

KEYWORDS

Eigenfactor Score; Normalized Eigenfactor; Logarithmic Eigenfactor; Discrimination; Data Distribution; Journal Influence Factor

1 Introduction and Motivation

The Normalized Eigenfactor and the Eigenfactor are very close in nature. The latter was introduced as a bibliometric indicator by Bergstrom et al. (2008). It works to compare the weights of scholarly journals by constructing a citation network and drawing on the PageRank algorithm to evaluate the influence of each. The calculation spans up to 5 years and avoids bias by excluding the influence of self-citing. In 2009, Journal Citation Reports maintained by Thomson Reuters began to use the two indicators: Eigenfactor Score (ES) and Article Influence Score (AIS). These are collectively referred to as the Eigenfactor, and are two important bibliometric indicators after the Impact Factor (IF). In 2015, JCR released the Nor

*Correspondence Author

malized Eigenfactor (NE), which expresses a relationship between the Eigenfactor and the average value for other journals in the same discipline. For example, an EF score of 2 for a journal means that it is twice as influential as the average journal in the JCR.

It is of great significance to study the characteristics of the Normalized Eigenfactor and its possible problems. Since the publication of the Normalized Eigenfactor, the academic community has been focused on its relationship with other bibliometric indicators, and has done less research on the relationship between the Normalized Eigenfactor and Eigenfactor Score. Further work in this area will help to deepen the understanding and application of the Eigenfactor Score in academic journal evaluation. Specifically, comparing their discriminative ability and data distribution, analyzing their advantages and disadvantages, and working towards resolving problems.

The characteristics of the Eigenfactor score are of great academic research value. Franceschet (2010) proposed 10 reasons in advocating the use of the Eigenfactor method. He believes that it has a solid mathematical background, a well-founded basis of axioms, and an interesting probabilistic interpretation. In summary, Massimo believes that it provides a compelling measure of journal status and shares meaningful relationships with other bibliometric indicators. Ernesto et al. (2018) studied the Radiology, Nuclear Medicine and Medical Imaging journals and found that the Eigenfactor Score (ES), the Article Influence Score (AIS), the cited half-life, and the 5-year impact factor were four significant predictors of 2-year-ahead total citations. Rousseau & Stimulate (2009) studied 165medical journals and concluded that the correlation between the h-index and Eigenfactor score was strong. This was judged on the Pearson coefficient, which between them reached as high as 0.951. The authors went on discuss the feasibility of using ES as an alternative to the Web of Science for evaluating scientific journals. Shideler & Araú jo (2016) examined three different scientific fields (aquatic science, sociology, and immunology), and believed that the Eigenfactor score was the best indicator for the annual advertised subscription price for sociology journals. Of interest, they also felt that differences may vary according to disciplines. The empirical research showed that there was a high correlation between the journal's Eigenfactor, the Article Influence score, and the total citations.

Although the Eigenfactor score is in widespread use, there exist certain problems with the measure. Davis (2008) found that for periodical groups with relatively low overall influence, the difference in Eigenfactor between consecutively ranked journals is smaller, and the degree of dispersion is lower. Because the data for Eigenfactor calculations is often restricted, the accuracy of the calculation is difficult to test. Based on the relationship between the Eigenfactor indicator, the audience factor, and the influence weight indicator, Waltman & Eck (2010) pointed out that the three indicators are insensitive to field differences with low impact. In essence, this is because the Eigenfactor Score indicator does not adequately discriminate appropriately for journals with lower participation. Ren (2009) has pointed out that a low Eigenfactor Score and a low degree of dispersion are common for low-impact journals. Measured in 13 low-influence journals among 76 SCI journals in China, the ES begins to show a difference in the 4th digit after the decimal point. Also, there exist many journals that appear to have heavy values.

From extant research we see that the principle and the characteristics of the Eigenfactor Score is relatively mature, although the primary focus has been on the relationship between ES and other bibliometric indicators, as well as on the application of ES. Given that the Normalized Eigenfactor was only launched in 2015, it is a relatively new bibliometric indicator;

the corresponding research is still in its infancy. This article focuses on the following aspects: Firstly, does the Normalized Eigenfactor improve discrimination of the Eigenfactor Score, in particular for cases of low-partition journals? If so, can we provide a proof?

Secondly, since the Normalized Eigenfactor is based on the Eigenfactor Score, it is important to understand the characteristics and consequences of this conversion. Specifically, what is the relationship between the Normalized Eigenfactor and the Eigenfactor Score?

Thirdly, we seek to establish the difference in discriminative ability and data distribution between the Normalized Eigenfactor and the Eigenfactor score. Is this improvement significant? How does this affect the evaluation of scholarly journals?

Fourthly, given that the Normalized Eigenfactor is an indicator that reflects the influence of an academic journal on a deep level, then, can it accurately suggest the journals' natural influence and gap? If not, how should it be optimized?

Based on the theoretical analysis, this article takes the mathematics journals in JCR2016 as an example. We focus on the comparison between the Normalized Eigenfactor and Eigenfactor in terms of correlation, discrimination, and data distribution. In addition, we make an effort to further analyze problems and possible solutions with respect to the former.

2 Research methods

2.1 Calculation of the Normalized Eigenfactor

Assume that there are n academic journals in a subject, and x_m is the Eigenfactor score for the m^{th} periodical. The Normalized Eigenfactor y_m is:

$$y_m = \frac{x_m}{(\sum_{i=1}^{m-1} x_i + \sum_{i=m+1}^{n} x_i)/(n-1)} \tag{1}$$

The Normalized Eigenfactor ym is equal to the mean value of Eigenfactor xm, divided by the scores of other journals in the same discipline.

2.2 Proof of increasing the discrimination of the Normalized Eigenfactor

Assume that p and q represent two academic journals having Eigenfactor scores of x_p and x_q, and Normalized Eigenfactor scores of y_p and y_q, respectively. For convenience, we assume the Eigenfactor includes all journals except for p and q. The sum of scores is A.

Assuming that $x_p > x_q$, the ratio between the scores of the two journals' Eigenfactors is $x_p/x_q > 1$. If we can prove that the ratio of Normalized Eigenfactor is greater than the ratio of Eigenfactor scores, that is, $(y_p/y_q - x_p/x_q) > 0$, then we can be confident that the discrimination has improved after the Eigenfactor score is converted to a Normalized Eigenfactor.

According to formula (1):

$$y_p = \frac{x_p}{(x_q + A)/(n-1)} \tag{2}$$

$$y_q = \frac{x_q}{(x_p + A)/(n-1)} \tag{3}$$

$$\begin{aligned} \frac{y_p}{y_q} - \frac{x_p}{x_q} &= \frac{\frac{x_p}{(x_q + A)/(n-1)}}{\frac{x_q}{(x_p + A)/(n-1)}} - \frac{x_p}{x_q} = \frac{x_p(x_p + A)}{x_q(x_q + A)} - \frac{x_p}{x_q} \\ &= \frac{x_p(x_p + A) - x_p(x_q + A)}{x_q(x_q + A)} = \frac{x_p(x_p - x_q)}{x_q(x_q + A)} > 0 \end{aligned} \tag{4}$$

Thus, we have proved that Normalized Eigenfactor can improve the discrimination of the Eigenfactor. Furthermore, we draw the corollary that the ranking results of the Normalized Eigenfactor are consistent with the ranking of the Eigenfactor.

2.3 The impact of Normalized Eigenfactor conversion in journal influence evaluation

To begin with, the conversion of Eigenfactor Score into Normalized Eigenfactor improves the discrimination, consequently widening the gap of Eigenfactors. This has been proved theoretically, although empirical testing is required in order to determine the degree of effect this has, and importantly, whether the discrimination is significant.

It is important to remember that converting the Eigenfactor Score to the Normalized Eigenfactor is a nonlinear transformation that will destroy the linear relationship between the original data and the target data. As such, there is a certain degree of information loss and this needs to be evaluated. The most commonly used method is to use scatter plots to examine and compare the correlation coefficients.

Furthermore, the nonlinear transformation destroys the original data distribution. This requires that it also be analyzed from this perspective to assess the repercussions.

2.4 Research Methods

1) Methods of comparing the discrimination

Two common discrimination methods are the discrete coefficient and the median maximum ratio. We introduce these methods for use in the comprehensive evaluation of the following attributes: High score ratio, Low score ratio, and HHI (Herfindahl-Hirschman Index).
The High score ratio is the proportion of the total of the Normalized Eigenfactors among the top 20% of journals, to that of all journals' Normalized Eigenfactors. Similarly, the low score ratio is proportion of the sum of Normalized Eigenfactors among those in the lowest 20% of journals.

The HHI is a measure of concentration created by Hirschman (1968) to detect monopolies by means of determining market competitiveness. In the domain of journal influence factors, it is used indicate the degree of discrimination, where a larger value means it is less balanced and has a lower degree. The HHI of a Normalized Eigenfactor is the sum of the squares of all journals' Normalized Eigenfactors. The formula is as follows:

$$HHI = \sum_{i=1}^{n} \left[\frac{y_i}{\sum_{i=1}^{n} y_i} \right]^2 \qquad (5)$$

2) Methods of comparing data distribution

Many researchers have found inherent bias in the methods used to evaluate scholarly journals. Vinkler (2009) proved the right-skewed nature of the distribution of citations. The author believed that papers published in journals with a higher impact factor merely provide the possibility of obtaining many citations. It is unreasonable to use this as a measure of a journal's influence, and thus results in a large bias in terms of determining its actual influence. Seglen (1992) found that the distribution of citation analysis is a typical skewed distribution, which does not obey the normal distribution and has a power law characteristic. Adler et al. (2009) believed that the sole reliance on citation data provides and incomplete picture. They outline and provide examples where journals that contain longer papers get more citations. According to the power law, the distribution of citation data is usually right-skewed.

A data distribution test, sometimes referred to as a goodness-of-fit test, checks to see if the data matches a normal distribution. A common method for this is the Jarque-Bera test. Although many bibliometric indicators do not follow a normal distribution, the p-value is usually very low so it is difficult to compare. Alternatively, the size of the Jarque-Bera test value can be used to determine the skew in data distribution is more severe or be optimized when the Eigenfactor Score convert to Normalized Eigenfactor.

3 Empirical research results

3.1 Research data

This paper uses the mathematics journals listed in JCR 2016 as a basis for study. This includes 12 main indicators: Total Cites, Journal Impact Factor, Impact Factor without Journal Self Cites, Impact factor, Immediacy Index, Average Journal Impact Factor Percentile, 5-Year Impact Factor, Eigenfactor Score, Normalized Eigenfactor, Article Influence Score, Cited Half-Life, and Citing Half-life. This paper focuses on the analysis of the relationship between the Eigenfactor Score (ES) and the Normalized Eigenfactor (NE). There are a total of 310JCR 2016 mathematics journals as the data source.

3.2 Comparison of discrimination and data distribution

Comparison between Eigenfactor and Normalized Eigenfactor is shown in Table 1. The Median Maximum Ratio, High Score Ratio, Low Score Ratio, Passing rate, Discrete Coefficient, and HHI of Eigenfactors and Normalized Eigenfactors are all relatively close. In fact, they are all the same after the decimal point, which indicates that the Normalized Eigenfactor has little effect on improving data discrimination and distribution. On average, the mean value of Normalized Eigenfactor is 0.264285, and the mean value of Eigenfactor Score is 0.002305. By observation we see that the Normalized Eigenfactor is simply the Eigenfactor score amplified by 114.66.

Table 1 Comparison between Eigenfactor and Normalized Eigenfactor

Comparative indicators	Eigenfactor Score	Normalized Eigenfactor
Mean	0.004449	0.509824
Median	0.002305	0.264285
Maximum	0.049840	5.711670
Minimum	0.000060	0.007160
Std. Dev.	0.006515	0.746535
Median / Maximum	0.046248	0.046271
High score ratio	0.619839	0.619837
Low scoreratio	0.030801	0.030806
Passing Rate	1.91%	1.91%
Discrete Coefficient	1.464374	1.464299
HHI	0.010120	0.010120
Skewness	3.753054	3.753124
Kurtosis	20.249680	20.250300
Jarque–Bera	4571.119000	4571.423000
Probability	0.000000	0.000000

The scatter plots of NE and ES are shown in Fig. 1. It is clear that the plot depicts almost a straight line, where the correlation coefficient is 1.000000. The means they are highly correlated and almost homogenous.

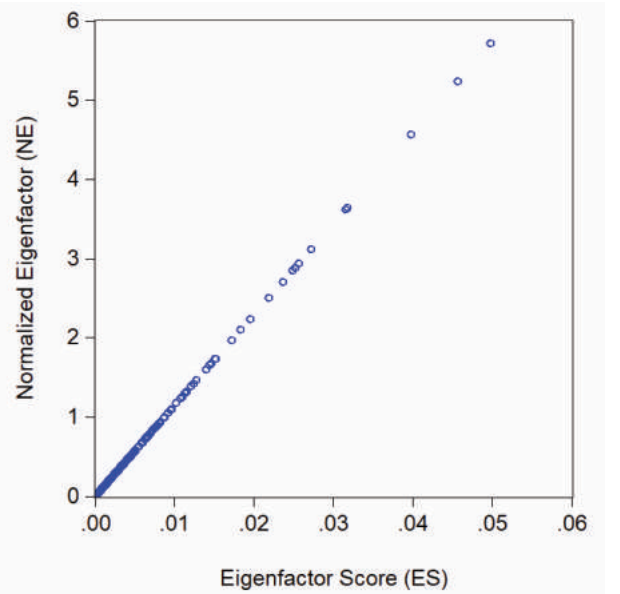


Figure 1 Scatter plots of Normalized versus standard Eigenfactor Scores

From the results we see that when the ES values are accurate to 5 decimal places, there are 69 cases where the values of journals are the same. Comparatively, when the NE values are accurate to 5 decimal places, the values of all journals are different. This suggests that the degree of discrimination has increased significantly.

The mechanism for improving the low segment discrimination of the Normalized Eigenfactor should be made clear by the following three aspects: First, when the Eigenfactor Scores are not equal, the Normalized Eigenfactors can definitely improve the discrimination. This point has been proven within this paper. Second, when the Eigenfactor Scores are equal, the Normalized Eigenfactors must be equal. Third, when the Eigenfactor Scores are equivalent and the Normalized Eigenfactor Scores are not equal, the true reason for the discrepancy is the increase in accuracy of the published score. Simply put, the precision relied upon by the JCR is insufficient. At present, only five digits after the decimal point are presented. If instead this were increased to perhaps the 6th or 7th decimal place, the Eigenfactor Score will reflect the difference. Ultimately, the degree of discrimination will increase.

To summarize, the Normalized Eigenfactor transforms the Eigenfactor to enhance the degree of discrimination. The mean value of the Normalized Eigenfactor is 114.66 times the mean of Eigenfactor score, and thus the difference is reflected in 5 decimal places. Since the score of the Eigenfactor is only published to the 5th decimal place, the discriminative degree cannot be realized. If the score of the low-partition periodical is equal, in fact, it would be necessary to publish up to and beyond the 7th decimal point.

A very important reason for the proposal of the Normalized Eigenfactor is to increase the discrimination for journals with a low partition. However, in summary analysis, the increase to this degree of discrimination is not borne by the contribution of Normalized Eigenfactor.

Rather, simply multiplying the Eigenfactor Scores by 100x, or even 1000x would achieve at least the same results.

4 Problems with Possible Improvements with the Normalized Eigenfactor

4.1 Further discussion on the Normalized Eigenfactor

The Normalized Eigenfactor is debatable, after all, since its conversion is a nonlinear transformation. Although the information conveyed by the Eigenfactor Score is generally not sacrificed, there is actually no improvement, or effective gain, when using the new measure. This is the case not only in discriminative capability, but also with respect to the data distribution. Therefore, it is our opinion that converting the Eigenfactor Score into a Normalized Eigenfactor is of little practical use.

In addition, since the denominator of the Normalized Eigenfactor conversion is the "average impact factor of all other journals", for each journal, the average value is not the same. Given that the Normalized Eigenfactor is supposed to be a measure of journal impact, does it makes sense that the scale can change? Is it still a good ruler? We do not think so.

Due to the reasons above, we feel that there is insufficient reason to convert the Eigenfactor Score into a Normalized Eigenfactor.

Research indicates that the data distribution for the Normalized Eigenfactor is heavily biased. In mathematics journals, such as "ADVANCES IN MATHEMATICS", the highest Normalized Eigenfactor is 5.71167. This is significant given that a perfect score is 100 points, and a passing score is 60. In our investigation we found that there are only 5 journals of the 310, which account for 1.61% of all journals. A result of this type tends to suggest that there is something drastically wrong with the approach. From the perspective of median, the median maximum ratio is 0.046271. That is, in the case of a full score of 100, half of the journals, i.e. 155, scored below 4.63 points. It seems the most likely explanation is due to a biased data distribution.

The Normalized Eigenfactor suffers from bias related to the difference between the evaluation value and its attribute. We suggest that this bias is only a superficial phenomenon. The Normalized Eigenfactor essentially reflects the influence attribute of academic journals. In the evaluation of this material, influence indicators should be close to a normal distribution except in cases of major or original innovations. Clearly, in situations like this, there should be a serious bias distribution. At present, there are few attributes that reflect major or original innovations in the bibliometric indicators. More often, there are indicators to reflect influence and timeliness. Consider the following example: When the Normalized Eigenfactor of the Journal A is 300 times of the Journal B, it does not mean that the influence of the Journal A is 300 times that of the second journal. This is drastically exaggerated. From our research it seems that editors can accept that the Normalized Eigenfactor of their own journal is 1/300 of that of a good journal, but they cannot accept that their journal is 1/300 of the impact of a good journal. It simply does not translate accordingly. Furthermore, the distribution of data based on the Normalized Eigenfactor should be much closer to a normal distribution.

The divergence between the evaluation value and the evaluation attribute is not unique to the Normalized Eigenfactor; the Eigenfactor Score and other bibliometric indicators also suffer from similar drawbacks.

4.2 Further Improvements to the Normalized Eigenfactor

According to the above analysis, the advantage of the Normalized Eigenfactor is that it can increase the degree of discrimination, albeit the level of improvement is almost negligible. There are four disadvantages. Firstly, the conversion of the Eigenfactor score to Normalized Eigenfactor is a nonlinear transformation, and consequently the information will degrade slightly. Secondly, also during the transformation, there is a phenomenon that the "measurement scale" is not uniform. Thirdly, the increase in Eigenfactor score to Normalized Eigenfactor is an illusion. Essentially, it is possible to achieve the same result simply by increasing the accuracy of published Eigenfactor scores. Finally, there is a notable difference between the evaluation value and the evaluation attribute of Normalized Eigenfactor.

4.2.1 Introducing the Logarithmic Eigenfactor

Due to the above problems, we feel that it is necessary to modify or replace the Normalized Eigenfactor to better reflect a journal's influence. In this paper, we introduce the Logarithmic Eigenfactor, which uses the natural Logarithmic scale to convert the Eigenfactor Score. This method has precedent and is used to calculate the population development index in the United Nations Development Programme. In this domain, the national income index reflects the diminishing marginal utility per each dollar increase in income and human development. The linear, dimensionless method is then used to obtain the national income index (UNDP, 2014). Taking the natural Logarithm is a nonlinear transformation, which can effectively reduce the gap between the evaluation objects. Furthermore, it will result in a data distribution closer to a normal distribution.

Since ES is usually very small and ln() is negative, in this paper, we use the following modified positive number:

$$LE = | \text{int}\{\min[\ln(X)]\} | + \ln(X) \qquad (6)$$

Where is X is the Eigenfactor Score, int(), ln() and min(), denote the integer component, the Logarithmic function and the minimum of X respectively. We call LE the Logarithmic Eigenfactor.

4.3 Logarithmic Eigenfactor analysis

The Notice that here, the indentation is inconsistent. Please correct this and standardize the formatting before making your final submission to the journal discrimination and data distribution between the Eigenfactor and the Logarithmic Eigenfactor is shown in Table 2. The median maximal value ratio of the LE is 0.561, indicating that the median is still at a slightly lower position. This represents a significant improvement over the ES (0.046). From the perspective of the Discrete Coefficient, the ES is 1.464, while the LE is the greatly reduced value of 0.261. The lower value indicates that the data is more uniform and the degree of discrimination has improved. The High Score Ratio of ES is 0.619, and that of the LE is 0.277. This means that the LE decreases the high score journal value and further unifies the distribution. The low score of ES is 0.031, while the low score of the LE shows improvement at 0.131. With respect to the Passing Rate, the ES is only 1.61%, while the LE reports 40.19%. This is an important advancement that more accurately represents the true influence of the journal.

Table 2 Comparison of Eigenfactor versus Logarithmic Eigenfactor

Comparison Indicators	Eigenfactor Score	Logarithmic Eigenfactor
Mean	0.004449	3.999
Median	0.002305	3.927
Maximum	0.049840	7.001
Minimum	0.000060	0.279
Std. Dev.	0.006515	1.042
Median / Maximum	0.046248	0.561
High Score Ratio	0.619839	0.277
Low Score Ratio	0.030801	0.131
Passing Rate	1.61%	40.19%
Discrete Coefficient	1.464374	0.261
HHI	0.010120	0.003
Skewness	3.753054	0.219
Kurtosis	20.249680	3.427
Jarque–Bera	4571.119000	4.831
Probability	0.000000	0.089

As we have discussed, the Eigenfactor Score does not obey the normal distribution. The Jarque-Bera test, however, calculates the value of the LE at 4.831, and the p value as 0.089, which rejects the null hypothesis of a non-normal distribution and shows that the LE does, in fact, obey one. Figures 2 and 3 illustrate more vividly the improvement of the distribution after the conversion from ES to LE.

In comparing the Logarithmic Eigenfactor to the Normalized Eigenfactor, it is clear that LE is superior in evaluating a journal's influence. In terms of linearity, LE is, indeed, also a non-linear transformation. However, LE is fundamentally different from NE. The purpose of the nonlinear transformation for LE is to make ES more representative of the journal's influence. Ideally, the relationship between the evaluation indicator value and the evaluation attribute is consistent. While the goal of NE is to improve the degree of discrimination, in reality it does not do so adequately. According to our research it would be easier to simply magnify ES by 100 times, and enjoy similar results.

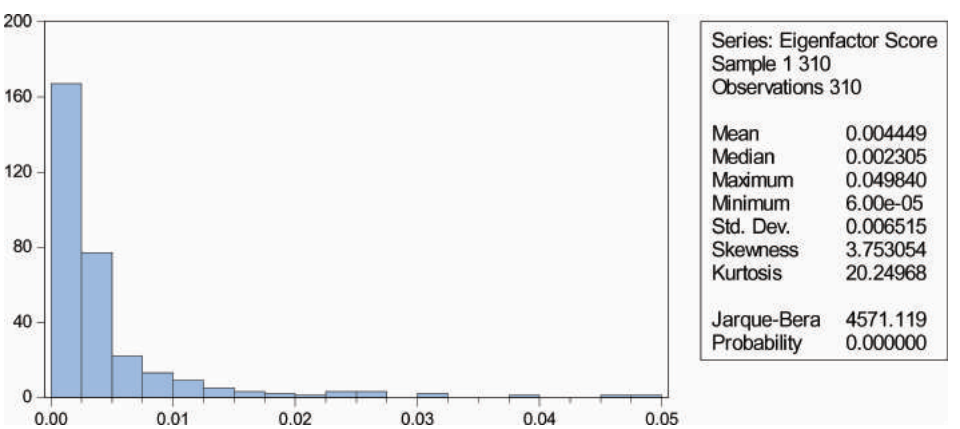


Figure 2 Data distribution of Eigenfactor Scores

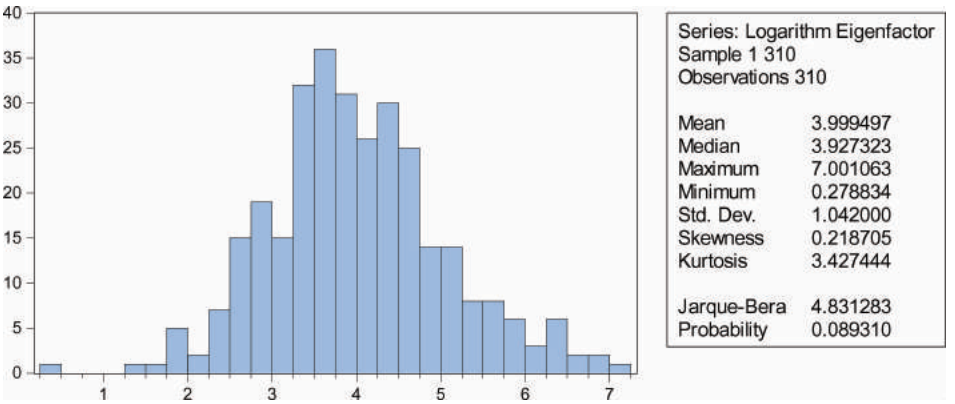


Figure 3 Data distribution of Logarithmic Eigenfactors

5 Research Conclusions

1) The Normalized Eigenfactor has little effect on the discrimination and data distribution of the Eigenfactor

This paper uses the mathematics journals in JCR2016 as an example to comprehensively compare Eigenfactor Scores and the Normalized Eigenfactor using the Median Maximum Ratio, High Score Ratio, Low Score Ratio, Passing Rate, Discrete Coefficient, HHI, and Jarque-Bera Test. The results of the study show that Normalized Eigenfactor had little effect on either the discriminate ability or the distribution of data of the Eigenfactor.

2) The published accuracy of the Normalized Eigenfactor is misleading in that it may cause people to believe that it will improve the discrimination

This paper proves that the Normalized Eigenfactor can improve the discrimination of Eigenfactor Scores in the low-partition of journals, but empirical research shows that it has little improvement. The main reason is that the Normalized Eigenfactor is equivalent to magnifying the mean value of Eigenfactor Score by more than 100 times. Therefore, when the decimal point is accurate to 5 digits, it can reflect the difference in less influential journals. At the same time, however, the Eigenfactor Score cannot highlight the gaps in the low-partition regions. This turns one's attention to the low degree of discrimination offered by NE. It is important to remember that the same discriminative ability can be obtained by increasing the precision of the Eigenfactor Score to7 decimal places, or simply by amplifying the Eigenfactor Score by 100x or 1000x.

3) There are theoretical and practical defects in the Normalized Eigenfactor

There are two theoretical flaws in the Normalized Eigenfactor. First, the conversion of the Eigenfactor Score to the Normalized Eigenfactor is a nonlinear transformation. This causes the loss of some information implicit in ES, although it is relatively minor. Second, when ES is converted into NE, the denominator is the mean of other journals' Eigenfactors. Because this changes dynamically - that is, the scale of the measurement is not fixed - it violates the basic principle of measurement.

From a practical point of view, there are also two major disadvantages. The first is that the Normalized Eigenfactor has neither effectively improved the discrimination of Eigenfactor Scores in the low partitions, nor changed the data distribution. What it has actually done is performed a "data amplification". As such, rather than use the Normalized Eigenfactor, it is

more straightforward to simply amplify the Eigenfactor Score accordingly. Secondly, the Normalized Eigenfactor reflects citations of journals on the surface. Essentially, it reports the influence of journals, but fails with respect to identifying gaps because it does not obey a normal distribution of data. In other words, there is significant difference between the indicator values and the evaluation attributes when using the Normalized Eigenfactor.

4) The Logarithmic Eigenfactor is a better indicator

Logarithmic conversion of the Eigenfactor Score into a Logarithmic Eigenfactor greatly improves the discriminative ability, as well as data bias. This study found that the Logarithmic Eigenfactors for JCR2016 mathematics journals are normally distributed and more consistent with the public perception pertaining to the journal in question. Additionally, the distribution of journal influence is closer to normal, which intuitively makes more sense. All things considered, we feel it is necessary to replace the Normalized Eigenfactor with the Logarithmic Eigenfactor. To study the bibliometric indicators, we must analyze not only the data of the indicators themselves, but also to keep the larger picture in mind and consider the results from the perspective of the nature of the indicators.

5) The Logarithmic Eigenfactors of journals in other disciplines need further study

As noted, above, our empirical study was completed using a subset of journals in JCR2016. Our source was comprised solely of mathematical journals, and we are cognizant of the fact that differences in discipline may change LE's degree of discriminant ability, or the distribution of data from normal, or both. This recognition warrants further study, and as such, we feel that exploration through the empirical testing using journals in other fields is an important topic of future research.

Acknowledgment

This paper is supported by the Humanities and Social Sciences projects of the Ministry of Education (17YJA630125), the Philosophy and Social Science Foundation of Zhejiang Province (17NDJC107YB), and China's National Natural Science Foundation Project (71663058). The authors thank them wholeheartedly for supporting this research.

Data Availability

The data used to support the findings of this study have been deposited in the mathematics journals listed in JCR 2016.

References

Adler, R., Ewing, J., & Taylor, P.(2009). Citation statistics: A report from the international mathematical union (IMU) in cooperation with the international council of industrial and applied mathematics(ICIAM) and the institute of mathematical statistics(IMS). *Statistical Science, 24* (1), 1–14.

Bergstrom, C. T., West J. D., & Wiseman M. A.(2008). The Eigenfactor metrics. *Journal of Neuroscience, 28* (45), 11433–11434.

Davis, P. M.(2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts?. *Journal of the American Society for Information Science & Technology, 59* (13), 2186–2188.

Roldan–Valadez, E., Orbe–Arteaga, U., & Rios, C.(2018). Eigenfactor score and alternative bibliometrics surpass the impact factor in a 2–years ahead annual–citation calculation: a linear mixed design model analysis of radiology, nuclear medicine and medical imaging journals. *La Radiologia Medica, 123* (7), 524–534.

Franceschet, M.(2010). Ten good reasons to use the Eigenfactor? metrics. *Information Processing & Manage–*

ment, 46 (5), 555–558.

Hirschman, A. O.(1968). The political economy of import–substituting industrialization in Latin America. *Quarterly Journal of Economics, 82* (1), 1–32.

Ren, S. L.(2009). Eigenfactor: The importance of analyzing journals and papers based on citation network(in Chinese). *Chinese Journal of Scientific and Technical Periodicals, 20* (3), 415–418.

Roldan–Valadez, E., Orbe–Arteaga, U., & Rios, C.(2018). Eigenfactor score and alternative bibliometrics sur–pass the impact factor in a 2–years ahead annual–citation calculation: A linear mixed design model analysis of radiology, nuclear medicine and medical imaging journals. *La Radiologia Medica, 123* (7), 524–534.

Rousseau, R., & Stimulate, G.(2009). On the relation between the WoS impact factor, the Eigenfactor, the SCImago Journal Rank, the Article Influence Score and the journal h–index. This article has been presented during a conference held at Nanjing University on 24–25 April 2009.

Seglen, P. O.(1992). The skewness of science. *Journal of the Association for Information Science & Technolo–gy, 43* (9), 628–638.

Shideler, G. S., & Araújo, R. J.(2016). Measures of scholarly journal quality are not universally applicable to de–termining value of advertised annual subscription price. *Scientometrics, 107* (3), 963–973.

UNDP.(2014). Human Development Report Technical Notes 2014. Retrieved from http://hdr.undp.org/sites/de–fault/files/hdr14_technical_notes.pdf

Vinkler, P.(2009). Introducing the current contribution index for characterizing the recent, relevant impact of journals. *Scientometrics,79* (2), 409–420.

Waltman, L., & van Eck, N. J.(2010). The relation between Eigenfactor, audience factor, and influence weight. *Journal of the American Society for Information Science and Technology, 61* (7), 1476–1486.