Lundberg *Journal of Big Data* (2023) 10:112
https://doi.org/10.1186/s40537-023-00793-6

Journal of Big Data

RESEARCH Open Access

Bibliometric mining of research directions and trends for big data

Lars Lundberg[1*]

*Correspondence:
Lars Lundberg
lars.lundberg@bth.se
[1] Department of Computer Science, Blekinge Institute of Technology, SE-37179 Karlskrona, Sweden

Abstract
In this paper a program and methodology for bibliometric mining of research trends and directions is presented. The method is applied to the research area Big Data for the time period 2012 to 2022, using the Scopus database. It turns out that the 10 most important research directions in Big Data are Machine learning, Deep learning and neural networks, Internet of things, Data mining, Cloud computing, Artificial intelligence, Healthcare, Security and privacy, Review, and Manufacturing. The role of Big Data research in different fields of science and technology is also analysed. For four geographic regions (North America, European Union, China, and The Rest of the World) different activity levels in Big Data during different parts of the time period are analysed. North America was the most active region during the first part of the time period. During the last years China is the most active region. The citation scores for documents from different regions and from different research directions within Big Data are also compared. North America has the highest average citation score among the geographic regions and the research direction Review has the highest average citation score among the research directions. The program and methodology for bibliometric mining developed in this study can be used also for other large research areas. Now that the program and methodology have been developed, it is expected that one could perform a similar study in some other research area in a couple of days.

Keywords Bibliometrics, Research directions, Research trends, Fields of science and technology, Geographic regions, Scopus database

Introduction

The term Big Data has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term [1]. Today, the amount of data collected and managed in most applications is increasing at a staggering pace. In May 2018 Forbes noted that 2.5 quintillion (10^{18}) bytes of data are produced every day [2], and the data production rate is increasing. As can be expected, the high relevance of the Big Data area has triggered a lot of research. A recent study of the trends and research directions in Big Data showed that during the 10-year period from 2012 to 2021 there are more than 118 000 documents related to Big Data in the Scopus database alone, and that approximately 20 000 new documents are published every year [3] (i.e., more than 50 new documents every

day!). It is clearly very challenging for any researcher to stay up-to-date with the trends and directions in a research area with a production rate as high as that of Big Data.

There are standardized classification codes in some research areas. Two examples are the ACM Computing Classification System [4] and the system for Mathematics Subject Classification [5]. These systems are hierarchical; the top level corresponds to different research areas in Computing and Mathematics respectively, and the lower levels correspond to research directions within these areas. If documents are tagged with such classification codes one can search for documents belonging to different research directions and plot trends that show which directions that are growing/declining, have the most citations, are most active in different geographic regions etc. However, such standardized classification systems are static and do not easily adapt to new trends and research directions. Therefore, static classification systems are not widely used in fast-growing areas such as Big Data. In fast-growing research areas, one needs to dynamically identify research directions based on a corpus of documents in the area.

One common dynamic approach for identifying research directions is to form clusters of author-defined keywords that are extracted from a corpus of documents in the research area. Obviously, there is a need to express research directions in a way that is useful for researchers and humans in general. However, approaches where common keywords are automatically extracted often result in lists of keywords that are either too general to be useful when identifying research directions, e.g., 'model' [6], 'data' [7], and 'application' [8], or confusing when identifying research directions, e.g., 'big data analysis' different from 'data analysis' [7]; 'library' different from 'college library' [8]; and 'big data analytics' different from 'data analytics' [9].

The problems mentioned above mean that existing dynamic approaches for identifying research directions in large and fast-growing research areas are unsatisfactory in the sense that the identified research directions and trends tend to be too general or confusing. Therefore, additional research is needed. In this study the problems with too general and confusing keywords are handled by a blacklist with the keywords that are too general and an expert-defined thesaurus that defines research directions that are easy to understand and not confusing (see Sect. 3 for details).

In this paper the challenge of identifying research directions and trends in large and fast-growing research areas such as Big Data is attacked by developing a program that performs semi-automatic analysis of publication databases, in this case the Scopus database. The analysis is based on bibliometric data mining. The program does not perform a fully automatic analysis since expert domain knowledge is needed. However, the program provides support that makes it possible for experts to identify important research directions and trends with limited effort even for large research areas with hundreds of thousands of documents. One limitation of previous bibliometric studies in the field of Big Data is that the number of documents considered has been relatively small (in the range of 334 to 25,334 documents, see Sect. 2 for details). This study is based on 137,148 documents. This means that one of benefits of the unique approach presented here is a methodology and a program that makes it possible to handle large document corpuses. Another benefit with the approach presented here is that it, through proper tool support, provides efficient use of the research area experts' time when performing bibliometric mining (see Sect. 3 for details).

There are two research contributions in this paper:

1. Identification of important trends in Big Data. The parameters analyzed in this paper are research directions, research productivity, fields of science and technology, geographic distribution, and citations.
2. A tool (program) and a methodology that can be used for identifying trends and research directions in large and fast-growing research areas. By combining the knowledge of research area experts with proper tool support, the identified trends and research directions will be useful and not too general or confusing. The tool and methodology can be used for bibliometric data mining also for other research areas than Big Data.

The rest of this paper is structured as follows. Section 2 describes related works. In Sect. 3 the methodology and the program for doing bibliometric mining are presented. Section 4 presents the results from the bibliometric mining. These results and other aspects related to the study are discussed in Sect. 5. The limitations of the current study are summarized in Sect. 6. Section 7 contains the conclusions and suggestions for future work.

Related works

As discussed above, there are two research contributions in this paper. In Sect. 2.1 we discuss surveys and other studies related to the identification of important research directions and trends in Big Data, and in Sect. 2.2 we discuss works and tools related to bibliometric studies.

Studies of research trends and directions in big data

Table 1 shows a summary of the bibliometric studies for Big Data research. The table shows that the number of documents used in the studies vary between 334 and 25,334. The two most used databases are Web of Science (WoS) and Scopus. The work presented in this paper is based on 137,148 documents from Scopus for the years 2012 to 2022. The parameters analyzed in this paper are research directions and trends, research productivity, fields of science and technology, geographic distribution, and citations. The tools and graphs used here are a Python program for bibliometric mining, and line, bar, and pie charts.

Bibliometric analysis and tools

Bibliometrics can be used for two main purposes: performance assessment of scientific actors (countries, universities, departments, and researchers), or displaying the structural and dynamic aspects of scientific research, delimiting a research field, and quantifying and visualizing the detected sub-fields (the latter purpose is often referred to as science mapping) [19][20]. In [21] Jappe examined the performance assessment practice in Europe. One conclusion from that study was that bibliometric research assessment is most frequently performed in the Nordic countries, the Netherlands, Italy, and the United Kingdom. Another conclusion was that WoS is the dominating database used for public research assessment in Europe.

Campanario [22] discussed how bibliometrics can make it possible to plot and visualize the impact factor of different journals. The plots suggested do not require sophisticated statistical techniques, yet they can be very helpful.

Table 1 Bibliometric studies for Big Data research

Ref.	#Docs	Databases	Tools/Graphs	Parameters analyzed	Years
[9]	4524	Scopus	VOSviewer, line, pie and bar charts	Research productivity, subject categories, geographic distribution, and citations	2010–2019
[8]	16,016	WoS and CNKI	CiteSpace and bar charts	Citations, geographic distribution	2008–2017
[10]	7299	WoS	Line charts	Relation between Big Data and Data Science	2006–2019
[7]	25,334	Scopus	VOSviewer, line and bar charts	Research productivity, trends, publication sources, geographic distribution, and citations	2010–2016
[11]	24,662	ACM Dig. Library, IEEE Xplore, SAGE Journals, ScienceDirect and WoS	Line charts	Research productivity, journals, keywords, Big Data frameworks and research challenges	2000–2017
[6]	334	WoS	VOSviewer, CiteSpace, and line charts	Citations, research hotspots and trends	2010–2022
[12]	7274	Scopus	Line, pie and bar charts	Research productivity, journals, articles, authors, and geographic distribution	2009–2018
[13]	10,989	Science Citation Index and Social Science Citation Index	VOSviewer and bar charts	Journals, geographic distribution, keywords, research hotspots and trends	2009–2018
[14]	6572	WoS	Microsoft Excel, bar charts and world map	Trends, research areas, geographic distribution, journals, authors, keywords, and citations	1980–2015
[15]	5840	WoS	VOSviewer and bar charts	Citation analysis, geographic distribution, and evolutionary pathways	2000–2015
[16]	7520	WoS and Scopus	Line charts and word clouds	Citations and research themes	2001–2016
[17]	693	ScienceDirect	Kiviat charts and bar charts	Research productivity, trends, hot topics, authors, and journals	2006–2016
[18]	4070	Scopus	Line, bar and pie chart, and word cloud	Research productivity, geographic distribution, and research areas.	2013–2018

New and emerging research directions are usually identified through different forms of citation analysis [23]. One example is the annual report on Research Fronts [24] compiled by Clarivate and the Chinese Academy of Sciences (CAS). This report uses citation analysis to identify relatively small groups of articles (typically less than 50 articles) that form a research front. Research fronts are often connected to "hot topics", e.g., COVID-19.

Most bibliometric tools and methods are based on clustering related publications. The two most common approaches to determine the relatedness of publications are based on either citation relations or word relations [25]. One problem with using citations for identifying research fronts and directions is that reliable citation data can only be obtained a substantial time after the publication of an article or a cluster of articles. In order to compensate for such delays, various techniques for predicting citations bursts for articles have been used [26]. Machine learnings techniques have not only been used for predicting citation bursts; they have also been used for handling the problem that it is not always correct to link topics with the same label during different time intervals [27]. In the case of word relations, the relatedness of publications is based on shared words in the titles, abstracts, or user-defined lists of keywords, such approaches are often used for very large sets of documents [28]. In this paper, clustering based on word relations is

used. The word relations come from an expert-defined thesaurus (further described in Sect. 3).

There are a number of software tools for bibliometric analysis. Two of the most popular tools are CiteSpace (https://citespace.podia.com/) [29] and VOSviewer (https://www.vosviewer.com/) [30]. CiteSpace supports visualization and analysis of trends and patterns in scientific literature. The functionality and use of VOSviewer is similar to that of CiteSpace. However, VOSviewer also offers text mining functionality [31]. Markscheffel and Schröter have done a comparative study of CiteSpace and VOSviewer [32]. The conclusion from that study was that visualizations created with VOSviewer are clearer and more user-friendly. However, CiteSpace offered advantages in the evaluative analysis of network visualizations, e.g., by enabling analysis of the cluster nodes using a cluster explorer. Other similar, but less popular, software tools are BibExcel, Netdraw, Pajek, Sci2, PoP (Publish or Perish), CitNetExplorer, SciMAT and HistCite.

All of the tools mentioned above focus on visualization of bibliometric data, often in the form of (very) large graphs with keywords, authors, or countries as nodes. Table 1 shows that no bibliometric study in Big Data considered as many documents as the current study (137,148 documents). The graphs produced by the tools discussed above tend to become increasingly detailed and complex when the number of documents grows, and even for studies based on a considerably smaller number of documents than this study, the graphs become very hard to read and interpret [7][13][15]. Therefore, visualization using existing tools seems less suited for a study that involves hundreds of thousands of documents. There does not seem to be any bibliometric study or tool that is able to automatically or semi-automatically identify important and useful research directions (in the form of frequently used keywords) in large research areas. As discussed in the introduction, the current approaches for identifying research directions in large and fast-growing research areas are unsatisfactory in the sense that the identified research directions tend to be too general or confusing. This is a research gap that is addressed in this paper. The approach presented here is semi-automatic and makes it feasible for experts to identify important research directions through data mining of hundreds of thousands of documents in a research area with limited effort. One key advantage of the approach suggested here is that it benefits from the intellectual work done by the authors when they define their lists of keywords. Moreover, through the use of a blacklist and a thesaurus the approach identifies useful research directions, i.e., directions that are not too general or confusing. The approach suggested here provides efficient ways of using expert knowledge in the mining process (see Sect. 3 for details).

Methodology

The research question that provides the basis for this bibliometric study is "What are the trends in Big Data in terms of productivity, research directions, geographic regions, citations and Big Data research in different fields of science and technology?".

Section 3.1 provides an overview of the mining process used in this study, and in Sect. 3.2 the data generated by the mining process are explained. A description of, and some additional information about, the program used in the mining process are presented in Sect. 3.3.

Overview of the mining process

Figure 1 gives an overview of the mining process. We start by defining the research area and the time period that we would like to study (Step 1 in Fig. 1). In this case the research area was Big Data, and the time period was the 11-year period from 2012 to 2022. These two parameters are sent to the program for bibliometric mining.

Using the Scopus API, we collect all documents with "Big Data" in either the title, list of author-defined keywords or abstract for each of the 11 years (Step 2 in Fig. 1). This is done in 11 steps using the 11 search strings: "TITLE-ABS-KEY ({big data}) AND (PUBYEAR=2012)",..., "TITLE-ABS-KEY ({big data}) AND (PUBYEAR=2022)". These 11 search strings define the corpus of documents that are included in the study.

For each document that we find with the search strings shown above, we get a record containing the following (and other) fields (Step 3 in Fig. 1):

1. Title of the document.
2. Author-defined keywords for the document (if any).
3. Abstract.
4. Number of citations.
5. Affiliation country.

These records are processed and filtered by the Python program for bibliometric mining that we have developed. One important task for the mining program is to identify popular and fast-growing keywords that reflect important research directions within Big Data. A major challenge is the overwhelming amount of user defined keywords in the retrieved documents (in this case more than 178 000 unique keywords). Another challenge is that some keywords are very general (e.g., "data", "research" and "future"); such general keywords are in most cases not very useful when identifying research trends

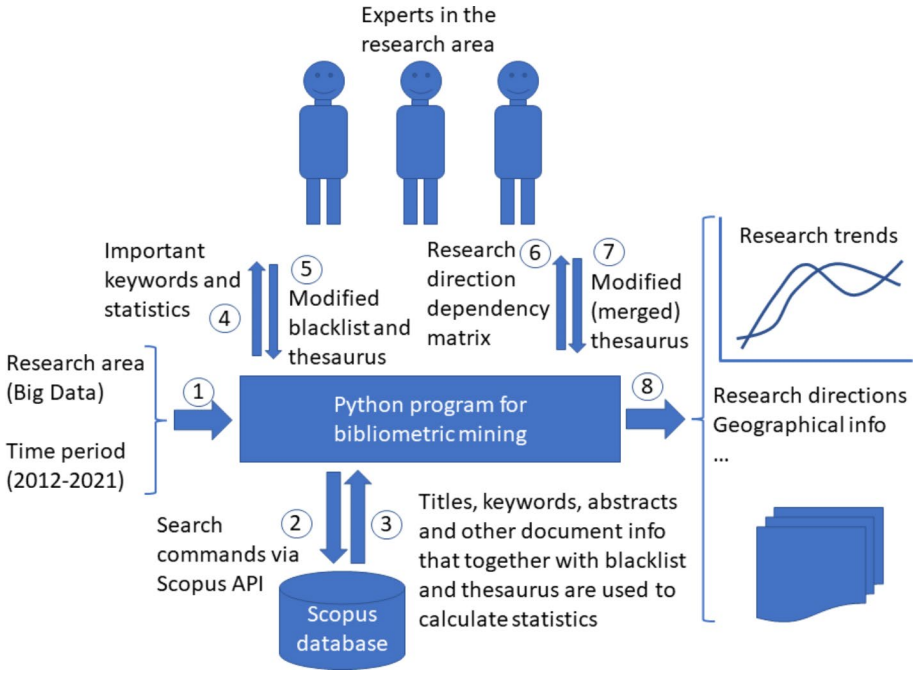

Fig. 1 Overview of the mining process

and directions. Two approaches have been used to address these challenges: one general automatic approach and one manual, research area specific, approach.

The automatic approach for reducing the number of keywords and removing keywords that are too general is based on only considering author-defined keywords that are present in at least a certain number of documents. When developing our methodology, we saw that keywords that only consist of one word (e.g., "data", "research" and "future") tend to be general and thus less useful compared to keywords that consist of two of more words (e.g., "deep learning", "convolutional neural networks" and "cloud computing"). Based on this observation and the need to reduce the number of keywords in general, and the number of very general keywords in particular, we developed a heuristic rule: For keywords consisting of two or more words, we automatically remove the keyword if it is present in less than a certain limit (MinimumOccurences) author-defined keyword lists. For keywords consisting of only one word, we automatically remove the keyword if it is present in less than a certain larger limit (OneWordFactor*MinimumOccurences) author-defined keyword lists. This means that to consider a keyword, we require OneWordFactor times more occurrences for a one-word keyword compared to a keyword consisting of two or more words. By experimenting with different values on OneWordFactor and MinimumOccurences and discussing with research area experts, it was decided that OneWordFactor=4 and MinimumOccurences=25 provided a useful list of keywords that could be presented to human experts (Step 4 in Fig. 1). For OneWordFactor=4 and MinimumOccurences=25 there were approximately 1600 remaining keywords, i.e., the initial number of keywords was reduced with more than a factor of 100.

After automatic filtering, which reduced the number of keywords from 178 000 to 1600, human experts developed a blacklist containing remaining keywords that are very general and thus less useful for determining research trends and directions (Step 5 in Fig. 1). Since many keywords are similar or related to the same research direction within the Big Data research area, the experts also created a thesaurus that clusters keywords into groups with similar meaning. These groups can be considered as research directions within Big Data. Some clustering is trivial and based on linguistic aspects, e.g., "neural network" and "neural networks" are put in the same group (all keywords in the documents are converted to lower case), and "health care" and "healthcare" are put in the same group. Some common abbreviations are also trivial to cluster, e.g., "internet of things" and "iot" and "artificial intelligence" and "ai". Clustering that requires the research area experts' knowledge and judgement are decisions such as putting "parallel processing" and "distributed processing" in the same group and putting "edge computing" and "fog computing" in the same group. Appendix A contains the blacklist and thesaurus used. Keywords that represent groups (we refer to such keywords as research directions) in the thesaurus are capitalize, e.g., "Deep learning" represents a group of keywords including "deep learning" (see Appendix A for details).

The output from the program to the experts in Step 4 in Fig. 1 is a list of keywords and numbers indicating the number of documents that contain the keyword, e.g. (<…> represent keywords that are omitted in this example):

<…>.

Security and privacy 9807.

<…>.

research 8009.

<…>.

privacy-preserving 1101.

<…>.

The fact that "Security and privacy" start with a capital "S" shows that this keyword (or research direction) is defined in the thesaurus by the experts. The other two keywords start with small letters and have been automatically extracted from the author-defined lists of keywords in the documents. Based on the list of keywords printed by the program, the experts may in Step 5 in Fig. 1 decide to put "research" on the blacklist (because this keyword is very general) and add "privacy-preserving" to the research direction "Security and privacy" in the thesaurus. When Step 4 is repeated the output from the program to the experts may look like this:

<…>.

Security and privacy 10,165.

<…>.

This means that "research" is removed from the keyword list and that "privacy-preserving" is included in the research direction "Security and privacy". N.B. since some documents may contain both "privacy-preserving" and some keyword that was already included in "Security and privacy" in the thesaurus the number of documents that contain a keyword associated with the research direction "Security and privacy" only increases with 358 from 9807 to 10,165 (10,165–9807 = 358), and not with 1101, which was the number of documents that contained "privacy-preserving" (see above).

Steps 4 and 5 are repeated until the research area experts are satisfied with the blacklist and thesaurus. An $n \times n$ research direction dependency matrix for the n most important research directions is then calculated (what makes a research direction important is defined in Sect. 3.2.1). The value of n is decided by the experts. However, to avoid an overwhelming number of research directions, n should be in the order of 10 or smaller. Let K_i be the set of all documents that contain a keyword, that (via the thesaurus) contains an author-defined keyword that is associated with research direction I, in either the title, the author defined list of keywords or the abstract. Entry $m_{i,j}$ in the research direction dependency matrix was obtained as

$$m_{i,j} = |K_i \cap K_j| \,/\, |K_i| \tag{1}$$

This means that, $m_{i,j}$ is a number between zero and one, and it is one when $i = j$. By looking at the research direction dependency matrix one can see if there is a large overlap between two research directions, i.e., if the number of documents that contain keywords associated with both research directions is relatively high (Step 6 in Fig. 1). If the values $m_{i,j}$ and $m_{j,I}$ (see Eq. 1) are high the research area experts may decide to merge research directions i and j by modifying the thesaurus (Step 7 in Fig. 1), i.e., the mining program supports the experts in their non-trivial task of creating a thesaurus that reflects important and reasonably non-overlapping directions within the research area. Based on data retrieved from the Scopus database, the thesaurus and the blacklist, the data mining program then generates data that describes important research directions, trends etc. in the research area (Step 8 in Fig. 1).

Data generated by the mining process

Research directions and research trends

The main result from the mining process is a list of the most important research directions found. The importance of a research direction is based on three criteria. To quantify these criteria, all keywords and research directions after the automatic and manual filtering are numbered in some arbitrary order. At this point author-defined keywords and expert-defined research directions are treated in the same way, and in the definition of the three criteria below we refer to an expert-defined research direction as a keyword. If keyword i is a research direction, the number of documents is calculated in the following way: The list of p_i keywords that correspond to the research direction is obtained from the thesaurus. For each of these p_i keywords we create a set A_j containing the documents that contain keyword j ($1 \leq j \leq p_i$) in the title, author-defined list of keywords or abstract. A set A_i is then created as

$$A_i = \bigcup_{j=1}^{p_i} A_j \tag{2}$$

The number of documents for research direction I, is the cardinality of A_i (see Eq. 2). The three criteria for selecting the most important research directions are:

- The total number of documents (ta_i) for keyword i ($1 \leq i \leq m$) for the entire period (2012–2022).

  The growth rate (gr_i) for keyword i during the time period. The idea is that if the number of documents that contain keyword i has increased rapidly during the time period, then keyword i is important. This metric is calculated in the following way: Let $k_{i,j}$ denote the number of documents published during year j that contain keyword i. Based on this we calculate gr_i in the following way (we set $k_{i,2011} = 0$, and the age factor $a = 1.5$).
  $$gr_i = \sum_{j=2012}^{2022} a^{j-2012}(k_{i,j} - k_{i,j-1}) \tag{3}$$

  The citation count (cc_i) for keyword i. Let K_i be the set of all documents that contain the keyword (the cardinality of K_i is ta_i). Also, let c_j be the number of citations of document j, then.
  $$cc_i = \sum_{document j in K_i} c_j \tag{4}$$

Three ranking lists with keywords were created: one based on ta_i, one based on gr_i (see Eq. 3), and one based on cc_i (see Eq. 4). The three ranks were added for each keyword and then the n keywords with the lowest sum were selected as the most important keywords. After some experimentation and discussions with research area experts it was decided that this way of selecting keywords and research directions made sense, since each of the three criteria reflects an important aspect that should affect the selection of important research directions within Big Data.

To visualize the trend in research direction I, the list of p_i keywords that correspond to research direction i is first obtained from the thesaurus. A document is counted as belonging to a research direction if it, in the title, list of author-defined keywords or abstract, contains at least one of the p_i keywords that is associated with the research direction. As a consequence, one document can belong to more than one research direction and some documents may not belong to any of the n most important research directions.

Documents that are published early will in general have more citations than documents published later, e.g., a document from 2012 will in general have more citations than a document published 2022. To be able to compare the citation counts from different years, a year-normalized citation score (NCS=Normalized Citation Score) was calculated for each document. The NCS for a document is the number of citations for the document divided with the average number of citations for documents in our dataset that are published the same year. By definition the average NCS for all documents in our dataset is 1.

The average NCS was calculated for each research direction i by considering the set A_i of all documents that contain at least one of the p_i keywords that, according to the thesaurus, are associated with the research direction. The average NCS for research direction i is the average of the NCS for the documents in set A_i.

Information about different fields of science and technology

Big Data research plays different roles in different fields of science and technology, and we therefore produce plots for each of the six top level fields in the Field of Science and Technology (FOS) classification (https://en.wikipedia.org/wiki/Fields_of_Science_and_Technology). Table 2 shows the top-level fields and the strings that are added to the Scopus search string, e.g., TITLE-ABS-KEY ({big data}) AND ("mathematics" OR "computer and information" OR "physics" OR "chemistry" OR "environmental science" OR "biology") AND (PUBYEAR=2012), for Big Data documents in Natural sciences from 2012, and TITLE-ABS-KEY ({big data}) AND ("medicine" OR "health" OR "health biotechnology") AND (PUBYEAR=2020) for Big Data documents in Medical and health sciences from 2020. Since the field specific search string is added to the original search string using the AND-operator, the set of documents considered for each field of science and technology is a subset of all the documents in this study. The field specific additions to the Scopus search strings are based on keywords from the second level of the FOS classification system.

The average NCS values for all documents from all years in each top-level field were calculated. In order to determine how the main research directions identified in Big Data were represented in each field of science and technology, a frequency factor $f_{i,x} = x_i/x_c$ was calculated, where x_c is the percentage of documents that belong to research direction X in the complete set of documents and x_i is the percentage of documents that belong to research direction X in the field specific subset of documents. If $f_{i,x} > 1$, then it is more common that documents belong to research direction X in field i of science and technology compared to how common it is that documents belong to research direction X in the complete set of documents. For instance, if $f_{\text{Engineering and technology, Internet of things}} = 1.27$, then the probability that a document in the subset corresponding to the research field Engineering and technology belongs to the research direction Internet of things is 27% higher compared to the probability that a document from the complete set of documents obtained without adding any of the search strings shown in Table 2 belongs to the research direction Internet of things.

Geographic information

For the important research directions, as well as for the research area Big Data as a whole, we plot the number of documents for the major geographic regions (based on

Table 2 Top level fields in science and technology and their associated search strings in Scopus

Top-level field	Addition to the Scopus search string
Natural sciences	("mathematics" OR "computer and information" OR "physics" OR "chemistry" OR "environmental science" OR "biology")
Engineering and technology	("civil engineering" OR "electrical engineering" OR "electronic engineering" OR "information engineering" OR "mechanical engineering" OR "chemical engineering" OR "materials engineering" OR "medical engineering" OR "environmental engineering" OR "environmental biotechnology" OR "industrial biotechnology" OR "nano technology")
Medical and health sciences	("medicine" OR "health" OR "health biotechnology")
Agricultural sciences	("agriculture" OR "forestry" OR "fishery" OR "animal" Or "diary" OR "veterinary" OR "agriculture biotechnology")
Social science	("psychology" OR "economics" OR "business" OR "educational science" OR "sociology" OR "law" OR "political science" OR "social and economic geography" OR "media and communications")
Humanities	("history" OR "archaeology" OR "languages" OR "literature" OR "philosophy" OR "ethics" OR "religion" OR "arts" OR "music")

affiliation country). We consider four geographic regions: North America (USA and Canada), European Union (taking Brexit into consideration by including UK until the end of 2019), China and The Rest of the World. A document that has affiliation countries from more than one geographic region will be counted proportionally in the corresponding geographic regions, e.g., a document with three authors with affiliation countries China, Sweden and Brazil will be counted 1/3 in the region China, 1/3 in the region European Union and 1/3 in the region The Rest of the World.

The average NCS was calculated for each region (considering all documents from that region), as well as for each combination of region and research direction. In order to calculate the average NCS for research direction i in a certain region the set of all documents that are from the region were put in a set A (documents are counted proportionally if there are authors from different regions). The list of p_i keywords that correspond to research direction i is then obtained from the thesaurus. For each of these p_i keywords we create a subset A_j of A such that A_j consists of the documents in A that contain keyword j ($1 \leq j \leq p_i$) in the title, author-defined list of keywords or abstract. A set A_i (see Eq. 5) is then created as

$$A_i = \bigcup_{j=1}^{p_i} A_j \tag{5}$$

The average NCS for research direction i for the region is the average NCS for the documents in A_i.

The mining program

The mining program is written in Python using the pybliometrics interface to Scopus [33]. Figure 2 gives an overview of how the program works. The program goes through all documents twice. During the first pass author-defined keywords are collected and for each author-defined keyword the number of documents that has the keyword in the list of author-defined keywords is counted. As described in Sect. 3.1, keywords are then filtered out based on one of two criteria: the keyword is present in less than a certain number (MinimumOccurences) of the author-defined lists of keywords, or the keyword is on the expert-defined blacklist (because the keyword is too general). Keywords that are in the expert-defined thesaurus are replaced with their research direction.

Pass one: Go through all documents to collect and count author-defined keywords.

Filter out keywords that are present in less than a certain number of documents and keywords on the blacklist and combine keywords that are in the thesaurus into expert-defined research directions.

Pass two: Go through all documents again and for each keyword, research direction and year, count the number of documents containing the keyword in the title, author-defined list of keywords or abstract. If the keyword is part of a research direction (defined by the thesaurus), each document is only counted once for each research direction even if the document contains many keywords that are included in the same research direction.

Generate three ranking lists based on (i) total number documents for each keyword and research direction, (ii) the growth rate for each keyword and research direction, and (iii) the citation rate for each keyword and research direction.

Calculate the combined ranking for each keyword and research direction (will expand the thesaurus so that research directions will dominate over single keywords). the experts

Calculate statistics in the form of NCS, number of documents per year etc. for top research directions, different fields of science and technology and different geographic regions.

Fig. 2 A brief description of the program used for bibliometric mining

During the second pass the program goes through all documents again and looks for the keywords that are not filtered out in the title, user-defined list of keywords, and abstract. There is a set associated with each research direction (and each keyword that is not filtered out and that does not belong to a research direction). If the keyword is part of a research direction (as defined by the thesaurus), the document containing the keyword is added to a set associated with the research direction. Since we are using sets, a document is never counted twice for a research direction, even if the document contains several keywords that are associated with a certain research direction.

The program then creates the ranking lists for the three criteria discussed in Sect. 3.2. These ranking lists are combined into a total ranking as described in Sect. 3.2 the top candidates from the combined ranking are selected as the most important research directions. Normally, the experts will continue to expand the thesaurus until the top candidates consist of research directions and no single keywords (see Sect. 4.1).

Finally, the statistics in the form of normalized citation score (NCS), number of documents, etc. for the top research directions are calculated. We do this for the entire field of Big Data research, as well as for each field of science and technology, and for each geographic region.

The Python code for the program can be found on https://github.com/Lars-Lundberg-bth/Bibliometric-mining.

Results

Directions in big data research

The Scopus search (Step 2 in Fig. 1) resulted in 137,148 documents out of which 109,986 had author-defined keywords. These documents contained a total of 558,684 author-defined keywords out of which 178,149 were unique keywords. After automatic filtering, the number of unique keywords is reduced to 1602. After some iterations of steps 4 and 5 in Fig. 1, the experts produced a blacklist and thesaurus (see Appendix A). The ranking list for the 12 top keywords is shown in Table 3. The reason for the gaps in the three ranking lists is that other keywords and research directions, that did not make it into the top 12, have these ranks. It was decided that the top 11 keywords should be used for

Table 3 First list of important research directions

Thesaurus keywords	Number of docs		Growth rate		Citation count		Sum of ranks
	ta_i	Rank	gr_i	Rank	cc_i	Rank	
1. Machine learning	13 716	0	37 469	1	189 329	0	1
2. Internet of things	9 045	4	24 434	3	154 725	1	8
3. Data mining	12 709	1	13 289	10	144 045	2	13
4. Cloud computing	11 323	2	12 634	12	142 712	3	17
5. Artificial intelligence	7 602	8	62 970	0	82 644	11	19
6. Healthcare	7 616	7	16 808	8	119 472	5	20
7. Deep learning	6 415	11	24 231	4	102 029	6	21
8. Security and privacy	10 165	3	12 300	13	101 813	7	23
9. Review	5 177	12	16 654	9	139 208	4	25
10. Neural networks	7 403	9	17 831	7	88 398	10	26
11. Manufacturing	4 707	13	18 734	6	92 269	9	28
12. Smart cities	3 053	16	5 402	31	51 501	14	61

Table 4 The research direction dependency matrix. The two largest values (except the trivial 1.00 along the diagonal) are highlighted

	1	2	3	4	5	6	7	8	9	10	11
1	1.00	0.08	0.17	0.06	0.17	0.09	0.16	0.09	0.07	0.15	0.05
2	0.13	1.00	0.07	0.28	0.16	0.08	0.06	0.15	0.08	0.04	0.15
3	0.19	0.05	1.00	0.07	0.06	0.07	0.05	0.07	0.06	0.06	0.03
4	0.08	0.22	0.08	1.00	0.08	0.06	0.03	0.16	0.05	0.03	0.06
5	0.30	0.18	0.10	0.12	1.00	0.12	0.15	0.08	0.08	0.12	0.10
6	0.17	0.10	0.12	0.08	0.12	1.00	0.06	0.09	0.07	0.05	0.02
7	0.33	0.08	0.09	0.06	0.18	0.08	1.00	0.08	0.06	**0.47**	0.03
8	0.13	0.13	0.09	0.18	0.06	0.07	0.05	1.00	0.05	0.05	0.03
9	0.18	0.14	0.14	0.11	0.12	0.10	0.08	0.10	1.00	0.05	0.09
10	0.28	0.05	0.10	0.05	0.12	0.05	**0.40**	0.07	0.03	1.00	0.03
11	0.14	0.29	0.07	0.16	0.16	0.03	0.04	0.07	0.10	0.05	1.00

further analysis, since there is a large gap from 28 to 61 in the sum of the ranks between Keyword 11 and Keyword 12 (see Table 3). The keywords in Table 3 all come from the thesaurus (they all start with a capital letter). Each keyword in Table 3 thus represents a research direction consisting of a group of author-defined keywords.

The research direction dependency matrix for the 11 most important research directions is shown in Table 4. The two highest values are $m_{7,10} = 0.47$ and $m_{10,7} = 0.40$. From Table 3 we see that Keyword 7 is "Deep learning" and Keywords 10 is "Neural networks". The research direction dependency matrix thus shows that the number of documents in the intersection between "Deep learning" and "Neural networks" is 47% of the total number of documents that has a keyword associated with "Deep learning". From Table 3 we see that the total number of documents that has a keyword associated with "Deep learning" is 6415. Consequently, the number of documents in the intersection is $0.47*6415 \approx 3000$. Also, the research direction dependency matrix shows that the number of documents in the intersection between "Deep learning" and "Neural networks" is 40% of the total number of documents that has a keyword associated with "Neural networks", i.e., $3000 \approx 0.40*7403$ ($ta_{10}=7409$, see Table 3).

The research direction dependency matrix also contains other interesting information, e.g., by looking at $m_{1,5}$ (0.17) and $m_{5,1}$ (0.30) one can see that the research directions "Machine learning" and "Artificial intelligence" are relatively strongly related. By looking

Table 5 s and final list of important research directions

Thesaurus keywords	Number of docs		Growth rate		Citation count		Sum of ranks
	ta_i	Rank	gr_i	Rank	cc_i	Rank	
1. Machine learning	13 716	0	37 469	1	189 329	0	1
2. DL and neural networks	10 820	3	34 059	2	138 018	5	10
3. Internet of things	9 045	5	24 434	4	154 725	1	10
4. Data mining	12 709	1	13 289	9	144 045	2	12
5. Cloud computing	11 323	2	12 634	11	142 712	3	16
6. Artificial intelligence	7 602	9	62 970	0	82 644	10	19
7. Healthcare	7 616	8	16 808	7	119 472	6	21
8. Security and privacy	10 165	4	12 300	12	101 813	7	23
9. Review	5 177	11	16 654	8	139 208	4	23
10. Manufacturing	4 707	12	18 734	6	92 269	9	27

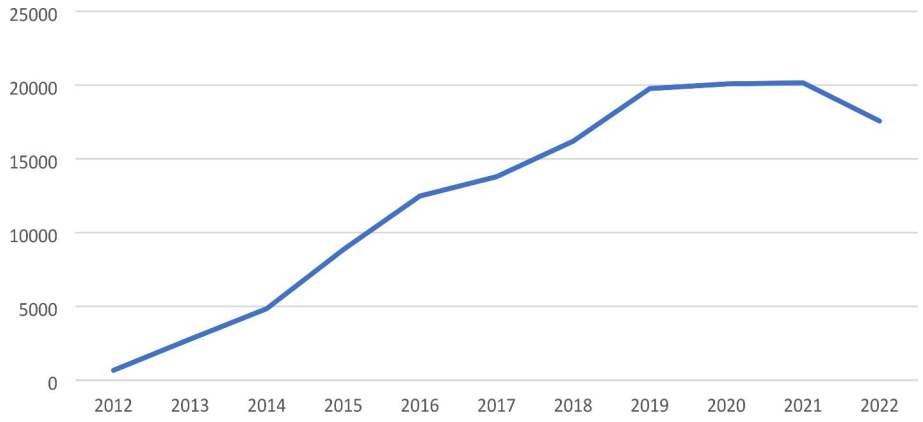

Fig. 3 Total number of documents per year in Big Data for the time period 2012 to 2022

at $m_{6,11}$ (0.02) and $m_{11,6}$ (0.03) it is clear that the research directions "Healthcare" and "Manufacturing" are almost totally independent.

Since there was a large overlap between the research directions "Deep learning" and "Neural networks", it was decided that these two research directions should be merged into a new research direction "DL and neural networks" in the thesaurus. Table 5 shows the 10 most important research directions after this merge. It can be noted that the number of documents for the new research direction is not the sum of the numbers of documents for the two previous keywords. The reason for this is, as discussed above, that the intersection between the two merged research directions was relatively large.

Trends in big data research

Figure 3 shows the total number of documents in Scopus for Big Data for the time period 2012 to 2022. The figure shows that there has been a rapid growth from 2012 to 2019. The number of documents were still growing for 2020 and 2021. However, the number of documents for 2022 is decreasing compared to 2021. The data from Scopus were collected April 2023. Some part, but not all, of the decrease could be explained by the fact that new documents are added to 2022 also after April 2023. The values visualized in Figs. 3, 4, 5, 6, 7, 8, 9 and 10 can be found in Appendix B.

The number of documents for the time period 2012 to 2022 for the 10 important research directions are shown in Fig. 4 (five research directions in the upper part of the figure and the other five in the lower part). Figure 4 shows that the fastest growing

research direction is "Artificial intelligence", which is consistent with the fact that this research direction has the highest growth rate (see Table 5). The figure also shows that the research directions "Data mining" and "Cloud computing" grew fast during the first half of the time period. However, during the second half of the period these two research directions grew rather slow. Figure 4 also shows that the research direction "Security and privacy" has stagnated the last years. This is consistent with the ranking of the growth rate of "Security and privacy" in Table 5.

Figure 5 shows the average NCS for the 10 important research directions. There is a significant difference between the average NCS for the research directions with the smallest average number of (year-normalized) citations per document ("Security and privacy" with NCS=1.01) and the research direction with the highest average NCS ("Review" with NCS=2.78). This means that on average there will be almost three times as many citations of a document related to the research direction "Review" compared to a document related to the research direction "Security and privacy" (2.78/1.01=2.75). It seems that systematic literature reviews and similar topics that the thesaurus has clustered under the research direction "Review" have high citation scores. The same is true for the keywords clustered under the research direction "Manufacturing". In Fig. 5 it may seem that the average NCS for all documents is larger than one. This is, however, not the case; the average NCS for all documents is by definition one. However, one of the three selection criteria for the 10 most important research directions within Big Data is the citation score, which is positively corelated with the NCS. This means that the research directions that end up as the top 10 have relatively high NCS.

When comparing the NCS values in Fig. 5 with the curves in Fig. 4 it seems that the research directions that has a low growth also have a low average NCS, i.e., "Security and privacy", "Data mining" and "Cloud computing" have relatively low average NCS, and all of these directions seem to be stagnating. For research directions with high average NCS, the connection between the average NCS and the growth rate is less clear.

Big data research in different fields of science and technology

Figure 6 shows the number of documents in Big Data per year for the six top-level fields of science and technology for the time period 2012 to 2022. The figure shows that three fields are growing (Medical and health sciences, Engineering and technology, and Agricultural sciences). The number of documents in other three fields has decreased from 2021 to 2022. This is particularly visible for Social science and Natural sciences. The decrease for Humanities is very marginal.

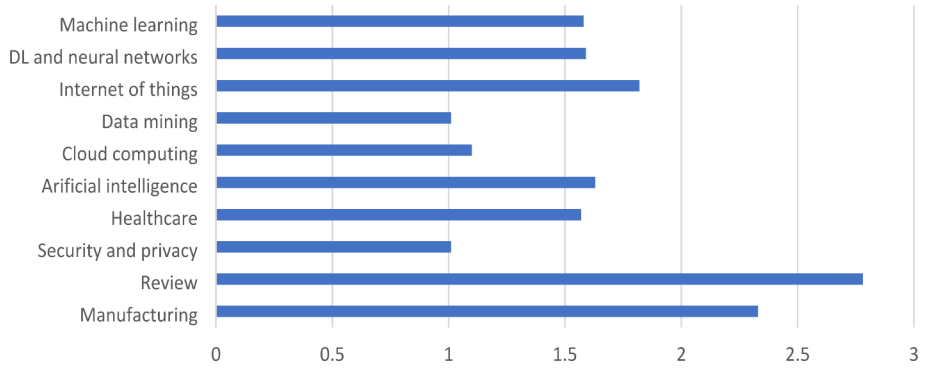

Fig. 5 The average Normalized Citation Score (NCS) for the identified research directions

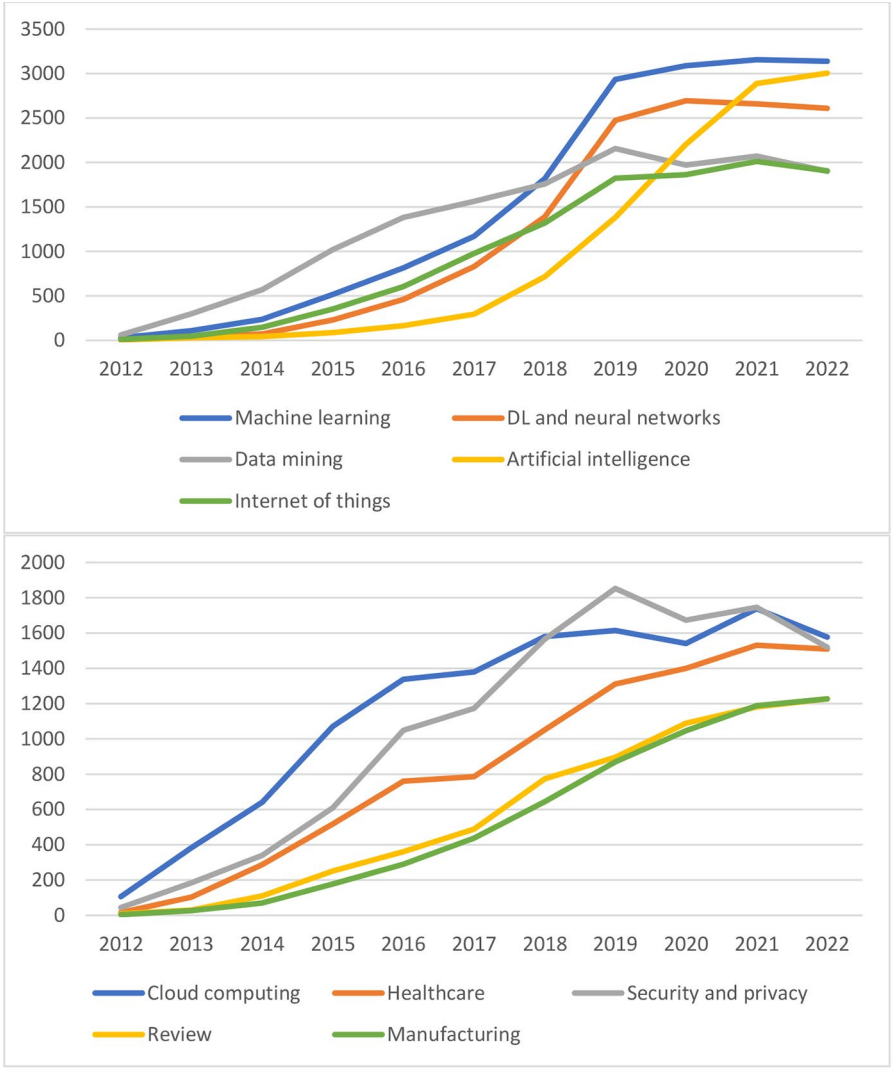

Fig. 4 The number of documents per year for the 10 identified research directions within Big Data for the time period 2012 to 2022

Figure 7 shows the average NCS for Big Data documents in the six top-level fields of science and technology. The figure shows that the average number of citations per Big Data document in Social science is the smallest and that the average number of citations is the largest in Agricultural sciences. This difference is, however, not very large.

Figure 8 shows the frequency factor $f_{i,x}$ for the 60 combinations of field in science and technology and identified research directions. The figure shows that documents related to the research direction Healthcare are more than a factor 3.5 more common than average in the field Medical and health sciences. This is not surprising. Figure 8 also shows that the percentage of Big Data documents in the research direction Healthcare is more than 50% higher in the field Natural sciences than the percentage of documents in Healthcare for the entire dataset consisting of 137,148 documents. The figure also shows

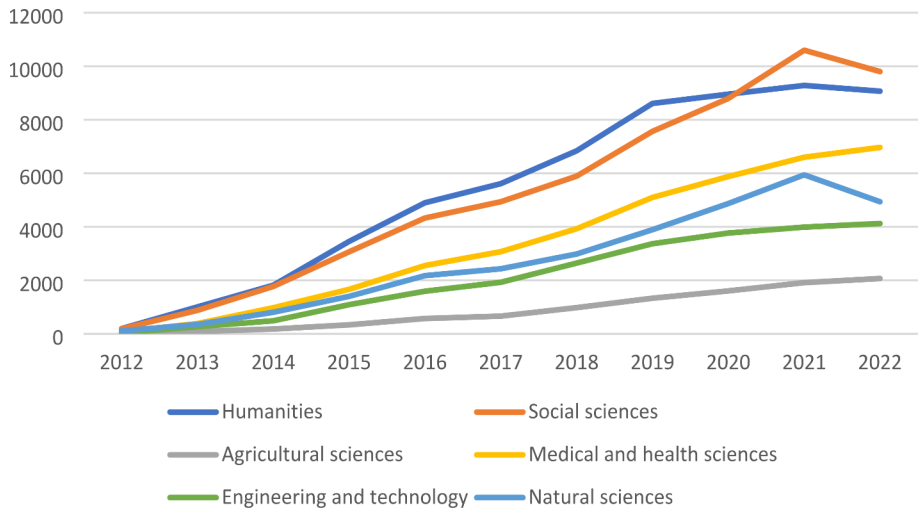

Fig. 6 The number of documents in Big Data per year for the six top-level fields of science and technology for the time period 2012 to 2022

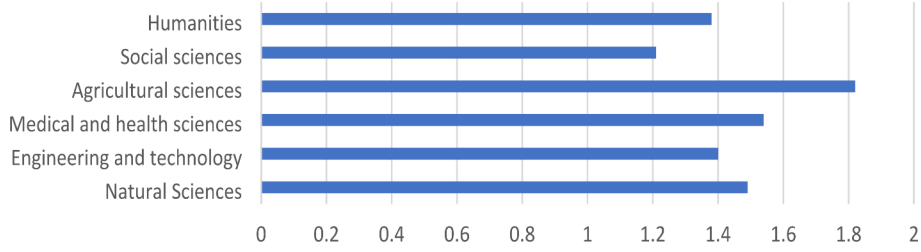

Fig. 7 The average NCS for the six top-level fields of science and technology

that the number of Big Data documents in the research direction Review is significantly higher in Humanities compared to entire set of Big Data documents.

Geographic regions in big data research

Table 6 shows the 20 major contributing countries in Big Data for the time period 2012 to 2022. The table shows that China and USA are the dominating countries. In fact, these two countries have more documents related to Big Data during the time period than the all the other 18 countries together. This imbalance is the main reason why regions and not countries have been considered here.

Figure 9 shows the number of documents in Scopus for Big Data for the time period 2012 to 2022 for the four geographic regions considered here. The figure shows that North America was the most active region during the first part of the time period. However, during the last part of the period China is the most active region in Big Data.

The upper left part of Fig. 10 shows the average NCS for the four geographic regions considered. The average NCS for all documents in all regions is by definition one. This part of the figure shows that the average NCS is more than twice as high for a document written by someone from North America compared to a document written by someone from China, i.e., on average there will be more than twice as many citations to a document from North America compared to a document from China written the same year. The difference in citation scores between North America and China has been observed

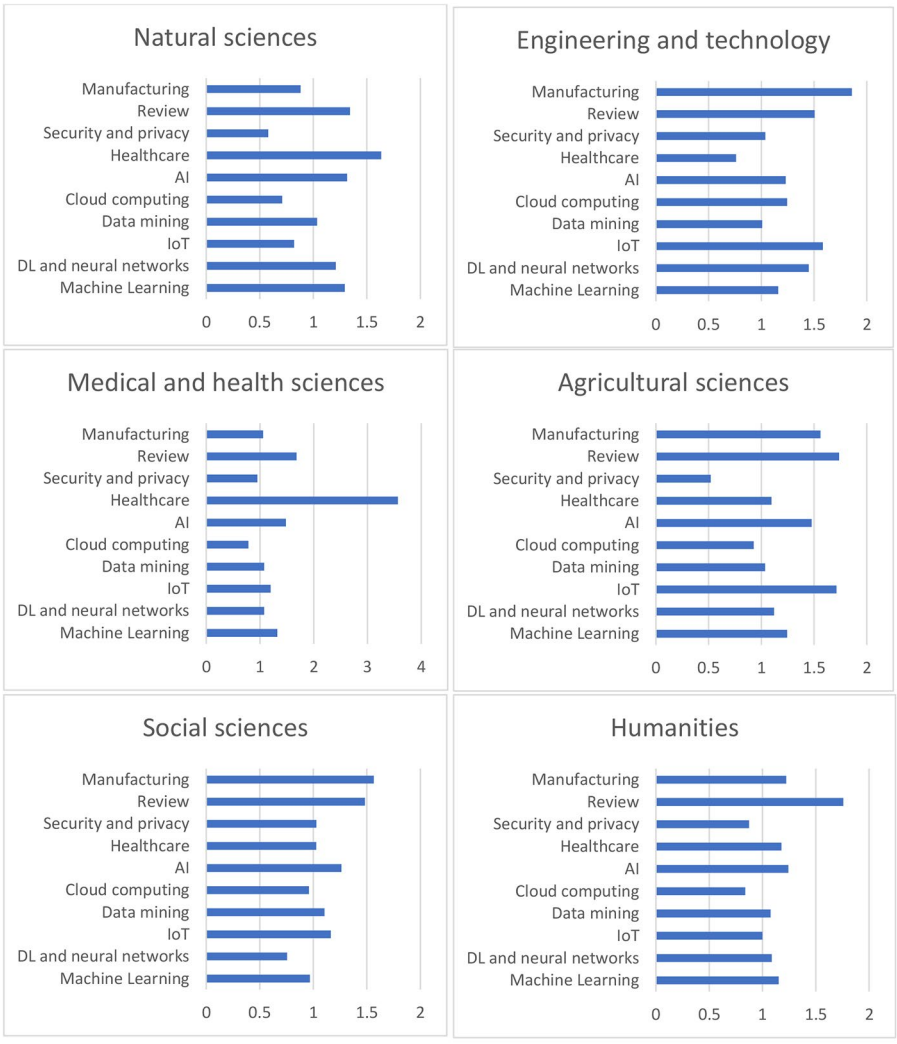

Fig. 8 The frequency factor $f_{i,x}$ for i = Natural sciences, Engineering and technology, Medical and health sciences, Agricultural sciences, Social sciences, and Humanities, and x = Manufacturing, Review, Security and privacy, Healthcare, AI, Cloud computing, Data mining, IoT, Deep learning and neural networks, Machine learning

Table 6 The 20 major contributing countries in Big Data for the time period 2012 to 2022

Country	Nr documents	Country	Nr documents
China	46,560	Spain	3266
United States of America	25,545	France	3238
India	10,993	Russian Federation	2400
United Kingdom	7444	Taiwan	2204
Germany	5146	Netherlands	1818
Italy	4447	Malaysia	1593
South Korea	4410	Greece	1516
Australia	4209	Hong Kong	1499
Canada	3775	Brazil	1503
Japan	3465	Saudi Arabia	1481

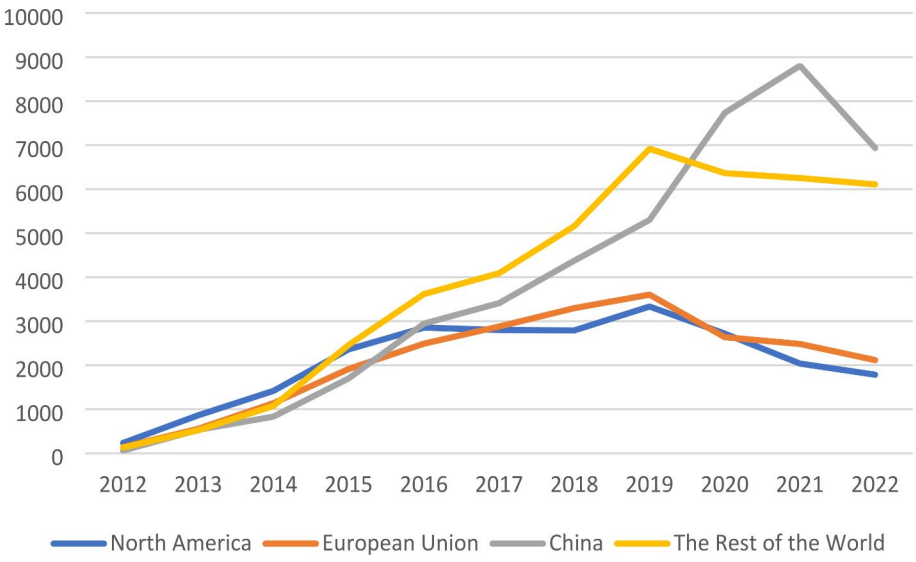

Fig. 9 The number of documents per year for the four regions North America, European Union, China, and The Rest of the World

previously [34]. One possible explanation for this difference could be that some Chinese authors may prefer to publish their results in Chinese journals [35].

The upper right part of Fig. 10 contains a pie chart that shows the distribution of the total number of documents in the four geographic regions that we consider. This part of the figure shows that China and The Rest of the World are the regions with the largest number of documents (the exact numbers can be found in Appendix B). N.B. the same colour code is used in all of the diagrams in Fig. 10, i.e., North America is blue, European Union is red etc.

The middle and lower parts of Fig. 10 show the same information as the upper part for each of the 10 identified research directions. These diagrams show that documents from North America have the largest average NCS, and that documents from China have the smallest average NCS for all of the 10 identified research directions within Big Data. For each of the 10 identified research directions the average NCS for all documents in all regions is shown in Fig. 4. As can be seen in Fig. 10 and Appendix B, a document written by authors from North America and related to the research direction "Review" (NCS=4.85) can expect more than 6.5 times more citations compared to a document written the same year by authors from China and related to the research direction "Security and privacy" (NCS=0.72).

By comparing the pie charts in the middle and lower parts of Fig. 10 with the big pie chart in the upper part of the figure, one can see that some research directions are more popular in some regions, e.g., "Healthcare" is popular in North America, "Manufacturing" is popular in the European Union and "DL and neural networks" is popular in China.

Discussion

The proposed semi-automatic method strikes a good balance between the need for automatic support to make it possible to investigate large research areas with hundreds of thousands of documents, and the need for expressing research directions in a way that is useful for researchers and humans in general. Fully automatic approaches result in a list

Fig. 10 The average NCS and distribution of documents for the different regions and for combinations of region and research direction

of common keywords that are either to general to be useful when identifying research directions (e.g., "model" [6], "data" [7], "application" [8], "massive amount" [15] and "big data" as a research trend in Big Data [14]) or confusing when identifying research directions (e.g., "big data analysis" different from "data analysis" [7]; "library" different from "college library" [8]; "smart city" different from "smart cities" [13]; "social science" different from "social scientist" [15]; and "big data analytics" different from "data analytics" [9]). In the approach presented here, these two problems are handled by the blacklist and thesaurus, respectively.

The approach presented in this paper makes it possible for research area experts, with efficient support of a program, to mine out important research directions in large research areas. The mining program and the methodology was developed in a research group consisting of more than 15 researchers. This group has a decade of experience of working with Big Data research together with international partners and partners from industry and society. The group has published more than 150 research articles related

to Big Data, some of these aimed at identifying trends and research directions [3][36] (more information about one of the large research projects done in the group can be found on https://a.bth.se/bigdata/).The group has done similar efforts before, but in that case without the support of a mining program but instead based on interviews with industry experts and senior international researchers in Big Data [3].

When comparing the research directions obtained with the help of the mining program with previous similar efforts in the same group, one can see that some of the identified research directions are the same or similar, (e.g., "Deep learning" and "Cloud computing"), but some research directions identified in this study were overlooked in the previous study (e.g., "Healthcare" and "Security and privacy"). The conclusion from this observation is that in a large research area such as Big Data, interviews and discussions with industry experts and international researchers will result in a limited and subjective sample of the research in the area, even if one has a large network of international and industrial research experts. One problem with identifying important research directions through interviews with industrial and academic experts and then using these directions as the basis for different forms of bibliometric trend analysis is that it is hard to predict which keywords different authors use to describe a research direction, e.g., "internet of things", "the internet of things" and "iot" have all been used to describe the same research direction (see Appendix A), and some of these ways of describing the research direction may be overlooked, thus providing incomplete statistics. In Step 4 in Fig. 1, the experts see a list of the keywords that actually have been used, so the risk of overlooking a certain keyword is minimized. This means that, compared to interviews with experts, the bibliometric mining approach presented here provides a more objective and complete way of identifying research directions.

One important aspect in this study was the three criteria for determining if a research direction is important. Two of the criteria are simple to quantify (Number of documents and Citation count). It is less obvious how the third criterion (Growth rate) should be quantified. However, by comparing the growth rate ranking in Table 5 with the actual growth curves in Fig. 4, one can see that the mathematically calculated growth rate factors correspond reasonably well to the subjective impression concerning which research directions that grows fastest, e.g., when looking at Fig. 4 most people would probably agree that "Artificial intelligence" grows the fastest and that "Security and privacy" grows the slowest. This means that the mathematical definition of the growth rate factor captures and quantifies the intuitive opinion about different growth rates in a good way.

The research direction dependency matrix was useful when deciding how independent/overlapping different research directions are. This kind of information is also useful when deciding if different research directions should be merged. In this case it was decided that, due to a significant overlap, the research directions "Deep learning" and "Neural networks" should be merged to a new research direction "DL and neural networks".

When studying how Big Data research is represented in different fields of science and technology, it turned out that Big Data research is growing in three fields (Medical and health sciences, Engineering and technology, and Agricultural sciences). In order to compare how different research directions are represented in different research fields a new metric called frequency factor was defined. This metric makes it possible to monitor the relative importance of different research directions in different fields. When looking

at the frequency factors it is clear that the Big Data research direction with largest relative importance in Natural sciences is Healthcare, the Big Data research direction with largest relative importance in Engineering and technology is Manufacturing, the Big Data research direction with largest relative importance in Medical and health sciences is Healthcare, the Big Data research directions with largest relative importance in Agricultural sciences are Review and Internet of things (IoT), the Big Data research direction with largest relative importance in Social science is Manufacturing, and the Big Data research direction with largest relative importance in Humanities is Review. By looking at the frequency factor, the results obtained in this study make it possible to identify hot topics related to Big Data research in different fields of science and technology. This is a source of inspiration for future research and research questions, particularly in the fields where Big Data research is growing.

The fact that a computer program is used in the mining process makes it easy to calculate metrics such as growth rate, research direction dependency matrix, frequency factor, average NCS per geographic region, average NCS for the different research directions, and average NCS for combinations of geographic regions and research directions. Such metrics would have been more complicated to obtain without the support of a computer program.

As discussed in Sect. 3, all keywords considered are obtained from author-defined lists of keywords (in this case there was 178,149 unique author-defined keywords). Since we are not doing any general text mining, we do not need to do any advanced weighting such as TF-IDF (Term Frequency – Inverse Document Frequency) to find the relevant keywords. Instead, we are benefitting from the intellectual work done by the authors when they define their lists of keywords. The purpose of the IDF component is to filter out frequently used words that in most cases do not carry a lot of interesting semantic information. The mining process described in this paper solves this problem through the expert-defined blacklist. This is the reason only the TF component (i.e., the number of documents containing the keyword) is used in the initial automatic reduction of the number of keywords (see Sect. 3). The time to do a TF-IDF analysis of the 137,148 documents considered in this study would to be excessively long (the complexity of TF-IDF is $O(x \log(x))$, where x is the total number of words in the dataset).

Limitations

As mentioned in the introduction, there are two types of research contributions in this paper: identification of important research directions and trends in Big Data and a methodology and tool that can be used for bibliometric mining also in other research areas.

One limitation related to the first research contribution is that only the Scopus database has been used. Another limitation related to the first research contribution is that only documents from 2012 to 2022 have been considered.

One limitation related to the second research contribution is that the tool and methodology have only been tested on one research area (Big Data).

A limitation related to both research contributions is that the growth rate criterion used for determining which research directions are most important has only been implemented and evaluated in one way (see Sect. 3.2.1). There could be other ways to implement and evaluate the growth rate of a research direction.

Conclusions and future work

Big Data is an active research area that plays an important role in many fields of science and technology. Using a mining program and research area experts, 10 important research directions within Big Data have been identified. These research directions are Machine learning, Deep learning and neural networks, Internet of things, Data mining, Cloud computing, Artificial intelligence, Healthcare, Security and privacy, Review, and Manufacturing. By looking at Big Data research in different fields of science and technology it becomes clear that Big Data research plays an important and growing role in many areas, and that the relative importance of the identified research directions varies between the different fields.

When looking at the Big Data activities in different geographic regions one can see that China is the most active region during the last years. The number of citations to documents in Big Data from China are, however, on average less than half of the citations to documents in Big Data from North America published the same year. Documents from North America have the most citations, and documents from China have least citations for all 10 important research directions within Big Data.

By comparing the current study with a previous study based mainly on interviews with international well-known researchers and experts, it is clear that the bibliometric mining approach presented in this paper provides a more objective and complete way of identifying research directions.

The tool (program) and a methodology developed in this study are not specific to Big Data. As a next study, the same approach will be used for other research areas, and a new study using the same methodology and tool has already been initiated in the research area Machine Learning. Based on the experiences from the current study, the time required for defining the blacklist and the thesaurus, i.e., steps 4 to 7 in Fig. 1, is estimated to 3–4 days. This means that we expect that the tool and method described in this paper will make it possible to do similar studies of other areas within a limited time frame, which opens up new and exciting possibilities for doing bibliometric research.

As mentioned above, the potential of the tool and methodology seems promising, and the research direction dependency matrix defined in this study provides useful support for the experts when defining non-overlapping research directions. However, more research is needed to streamline and minimize the work of the experts even further. In particular the task of creating the thesaurus could benefit from additional tool support. One way of providing such support could be to start with a small expert-defined thesaurus and then automatically expand, or at least suggest ways to expand, the thesaurus based on set operations. For instance, a keyword that has not yet been inserted into the thesaurus by the experts could automatically be inserted into the thesaurus based on the number of documents in the intersections of the set of documents corresponding to the keyword and the sets of documents corresponding to the research directions provided by the experts. This kind of support would reduce the effort for the experts. This is an approach that will be investigated in future studies.

As stated in Sect. 6, one limitation in this study is that the tool is specific to the Scopus. However, the methodology and approach are general and could be applied also to other databases, such as Web of Science. One direction of future research could therefore be to develop and evaluate similar tools also for other databases.

Appendix A: Blacklist and Thesaurus.

The blacklist and thesaurus used are shown below. The keywords in the blacklist and the keywords in the thesaurus that start with a small letter all come from the author-defined keyword lists in the documents retrieved from the Scopus database. The keywords in the thesaurus that start with a capital letter are referred to as research directions. The research directions are defined by the experts shown in Fig. 1.

Blacklist.

['research', 'data', 'big data', 'big data technology', 'big data management', 'it', 'its', 'ict', 'analysis', 'information', 'system', 'processing', 'computing', 'analytics', 'challenges', 'model', 'models', 'technology',

'development', 'methods', 'technologies', 'learning', 'framework', 'algorithm', 'algorithms', 'process', 'future', 'time', 'management', 'intelligence', 'design', 'knowledge', 'service', 'intelligent', 'digital', 'quality', 'accuracy', 'environment', 'platform', 'prediction', 'industry', 'services', 'science', 'application', 'applications', 'internet', 'web', 'software', 'business', 'resources', 'value', 'evaluation', 'impact', 'control', 'smart', 'detection', 'cost', 'volume', 'context', 'data analytics', 'data analysis', 'big data analysis', 'big data analytics', 'the era of big data', 'data processing', 'information technology', 'big data era', 'construction', 'network', 'networks', 'data sets', 'data set', 'data collection', 'data management', 'big data processing', 'large data', 'data sources', 'large scale', 'bigdata', 'big-data', 'big data applications', 'big data technologies', 'massive data', 'modeling', 'metadata',

'big data analytics capability', 'big data analytics capabilities', 'monitoring', 'performance', 'efficiency', 'optimization', 'methodology', 'benchmark', 'classification', 'clustering', 'semantics', 'measurement']

Thesaurus.

Thesaurus before merging Deep learning and Neural networks.

Research direction	List of keywords associated with the research direction
'Artificial intelligence'	['artificial intelligence', 'ai', 'artificial intelligence ai']
'Cloud computing'	['cloud', 'cloud computing', 'cloud services', 'cloud platform', 'cloud storage']
'Data mining'	['data mining', 'mining', 'educational data mining', 'text mining', 'big data mining', 'data science', 'process mining']
'Deep learning'	['deep learning', 'deep neural network', 'deep neural networks', 'deep learning dl', 'deep reinforcement learning']
'Edge and fog computing'	['edge computing', 'fog computing']
'Healthcare'	['health', 'healthcare', 'medical informatics', 'mhealth', 'electronic health records', 'health care', 'public health', 'digital health', 'telemedicine', 'personalized medicine', 'precision medicine', 'health informatics', 'medical big data', 'epidemiology', 'medical imaging', 'e-health', 'diabetes', 'breast cancer', 'genetics']
'Internet of things'	['iot', 'internet of things', 'the internet of things', 'internet of things iot', 'internet-of-things', 'industrial internet of things']
'Machine learning'	['machine learning', 'machine learning algorithms']
'Manufacturing'	['industry 4 0', 'manufacturing', 'smart manufacturing', 'digital twin', 'smart factory', 'automation', 'robotics', 'industrial big data', 'supply chain management', 'supply chain', 'predictive maintenance']
'Neural networks'	['neural network', 'neural networks', 'artificial neural networks', 'artificial neural network', 'convolutional neural networks', 'convolutional neural network', 'convolution neural network', 'cnn', 'recurrent neural network', 'recurrent neural networks', 'bp neural network']
'Parallel and distributed computing'	['parallel', 'distributed', 'gpu', 'cuda', 'parallel processing', 'parallel algorithms', 'parallel algorithm', 'distributed processing', 'parallel computing', 'distributed computing', 'load balancing', 'distributed systems', 'distributed system', 'high performance computing', 'high-performance computing', 'hpc']

Research direction	List of keywords associated with the research direction
'Review'	['review', 'literature review', 'survey', 'systematic literature review', 'systematic review', 'bibliometrics', 'bibliometric analysis']
'Security and privacy'	['security', 'authentication', 'data security', 'data protection', 'privacy', 'cloud security', 'privacy preservation', 'privacy preserving', 'privacy-preserving', 'data privacy', 'differential privacy', 'big data security', 'intrusion detection', 'anomaly detection', 'encryption', 'homomorphic encryption', 'cryptography', 'cyber security', 'trust', 'information security', 'cybersecurity', 'privacy protection', 'network security']
'Smart cities'	['smart city', 'urban computing', 'smart cities', 'citizen science', 'smart home']
'Social media'	['social media', 'social big data', 'social network', 'social networks', 'social network analysis']
'Storage'	['storage', 'database', 'databases', 'data lake', 'nosql', 'data storage', 'mongodb', 'file system', 'database systems', 'hbase', 'cassandra']
'Teaching and education'	['teaching', 'education', 'higher education', 'e-learning']

Thesaurus after merging Deep learning and Neural networks to DL and neural networks.

The research directions Deep learning and Neural networks were replaced by the research direction DL and neural networks.

Research direction	List of keywords associated with the research direction
'DL and neural networks'	['deep learning', 'deep neural network', 'deep neural networks', 'deep learning dl', 'deep reinforcement learning', 'neural network', 'neural networks', 'artificial neural networks', 'artificial neural network', 'convolutional neural networks", convolutional neural network', 'convolution neural network', 'cnn', 'recurrent neural network', 'recurrent neural networks', 'bp neural network']

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-023-00793-6.

Supplementary Material 1

Acknowledgements
The author would like to thank the researchers and teachers at the department of Computer Science at Blekinge Institute of Technology for good discussions and providing a good research environment for this study.

Authors' contributions
There is only one author, Prof. Lars Lundberg. Lars has written the paper and done the work.

Funding
This research was funded by Blekinge Institute of Technology, Sweden. No external research grant.
Open access funding provided by Blekinge Institute of Technology.

Data Availability
The data used are available in Appendix B. The data in Appendix B have been extracted from the Scopus database in April 2023.
Competing interests.
No competing interests.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
There is only one author (Lars Lundberg), and he gives his consent for publication.

Received: 30 September 2022 / Accepted: 21 June 2023
Published online: 01 July 2023

References

1. Lohr S. (1 February 2013), "The Origins of 'Big Data': An Etymological Detective Story". The New York Times. Archived from the original on 6 March 2016. https://archive.nytimes.com/bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/, Retrieved 26 April 2023.
2. Marr B. "How much data do we create every day? The mind-blowing stats everyone should read," https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=661e274e60ba, 2018.
3. Lundberg L, Grahn H. "Research Trends, Enabling Technologies and Application Areas for Big Data," *Algorithms*, vol. 15, no. 8, p. 280, 2022, DOI: https://doi.org/10.3390/a15080280.
4. Speretta M, Gauch S, Lakkaraju P. "Using CiteSeer to analyze trends in the ACM's computing classification system," in 2010,. DOI: https://doi.org/10.1109/HSI.2010.5514510.
5. Dong Y. "NLP-Based Detection of Mathematics subject classification," In: Davenport J, Kauers M, Labahn G, Urban J, editors Mathematical Software – ICMS 2018. ICMS 2018. Lecture notes in Computer Science(), vol 10931. Springer, Cham. https://doi.org/10.1007/978-3-319-96418-8_18.
6. Wang C, Dai J, Xu L. "Big data and data mining in education: A bibliometrics study from 2010 to 2022," 7th International Conference on Cloud Computing and Big Data Analytics (2022), DOI: https://doi.org/10.1109/ICCCBDA55098.2022.9778874.
7. Gupta V, et al. A quantitative and text-based characterization of big data research. J Intell Fuzzy Syst. 2019;36(5):4659–75.
8. Wang W, Lu C. "Visualization analysis of big data research based on Citespace," Soft Comput (Berlin Germany), vol. 24, (*11*), pp. 8173–86, 2019;2020.
9. Rawat KS, Sood SK. "Emerging trends and global scope of big data analytics: a scientometric analysis," Qual Quant, vol. 55, (*4*), pp. 1371–96, 2020;2021.
10. Raban DR, Gordon A. The evolution of data science and big data research: a bibliometric analysis. Scientometrics. 2020;122(3):1563–81.
11. Gupta D, Rani R. A study of big data evolution and research challenges. J Inform Sci. 2019;45(3):322–40.
12. Parlina A, Ramli K, Murfi H. Theme mapping and bibliometrics analysis of one decade of big data research in the scopus database. Inform (Basel). 2020;11(2):69.
13. Xu Z, Yu D. A Bibliometrics analysis on big data research (2009–2018). J Data Inform Manage. 2019;1(1–2):3–15.
14. Kalantari A, et al. A bibliometric approach to tracking big data research trends. J Big Data. 2017;4(1):1–18.
15. Zhang Y, et al. Discovering and forecasting interactions in big data research: a learning-enhanced bibliometric study. Technological Forecast Social Change. 2019;146:795–807.
16. Lu LYY, Liu JS. "The major research themes of big data literature: From 2001 to 2016," in 2016 IEEE International Conference on Computer and Information Technology.
17. Akoka J, Comyn-Wattiau I, Laoufi N. Research on Big Data – A systematic mapping study. Comput Stand Interfaces. 2017;54(Part 2):105–15.
18. Liu X, et al. The research landscape of big data: a bibliometric analysis. Libr Hi Tech. 2020;38(2):367–84.
19. Herrera-Viedma E, Martinez MA, Herrera M. "Bibliometric tools for discovering information in database," Lecture notes in Computer Science (including Subseries lecture notes in Artificial Intelligence and Lecture Notes in Bioinformatics), H. Fujita Eds. Cham: Springer International Publishing, 2016, 193–203.
20. Gutiérrez-Salcedo M, et al. Some bibliometric procedures for analyzing and evaluating research fields. Appl Intell (Dordrecht Netherlands). 2018;48(5):1275–87.
21. Jappe A. Professional standards in bibliometric research evaluation? A meta-evaluation of european assessment practice 2005–2019. PLoS ONE. 2020;15(4):e0231735.
22. Campanario JM. JIF-Plots: using plots of citations versus citable items as a tool to study journals and subject categories and discover new scientometric relationships. Scientometrics. 2017;113(2):1141–54.
23. Mazov NA, Gureev VN, Glinskikh VN. The methodological basis of defining Research Trends and Fronts. Sci Tech Inform Process. 2020;47(4):221–31.
24. Analytics C. "Research Fronts 2021," https://discover.clarivate.com/ResearchFronts2021_EN, 2022. Visited April 29, 2023.
25. van Eck NJ, Waltman L. "Visualizing bibliometric networks," Measuring Scholarly Impact, Springer International Publishing, 2014, 285–320.
26. Amjad T et al. "Citation burst prediction in a bibliometric network," *Scientometrics*, vol. 127, *(5)*, pp. 2773–2790, 2022.
27. Zhang Y, et al. Scientific evolutionary pathways: identifying and visualizing relationships for scientific topics. J Association Inform Sci Technol. 2017;68(8):1925–39.
28. Boyack KW, et al. Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. PLoS ONE. 2011;6(3):e18029–9.
29. Guzmán Sánchez MV, "CHEN, CHAOMEI, CiteSpace: A Practical Guide for Mapping Scientific Literature., Hauppauge NY, Nova Science. 2016, 169 pp. ISBN: 978-1-53610-280-2: eBook: 978-1-53610- 295-6 [CiteSpace: una guía práctica para el mapeo de la literatura científica]," *Investigación Bibliotecológica*, vol. 31, *(nesp1)*, pp. 293–295, 2018;2017.
30. Wong D. VOSviewer. Tech Serv Q. 2018;35(2):219–20.
31. van Eck NJ, Waltman L. "Text mining and visualization using VOSviewer," https://arxiv.org/abs/1109.2058, 2011.
32. Markscheffel B, Schröter F. Comparison of two science mapping tools based on software technical evaluation and bibliometric case studies. Collnet J Scientometrics Inform Manage. 2021;15(2):365–96.
33. Rose ME, Kitchin JR. Pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. Softwarex. 2019;10:100263.
34. Zhu J et al. "Measuring recent research performance for Chinese universities using bibliometric methods," *Scientometrics*, vol. 101, *(1)*, pp. 429–443, 2014.
35. Shu F, Julien C, Larivière V. Does the web of science accurately represent chinese scientific performance? J Association Inform Sci Technol. 2019;70(10):1138–52.

36. Lundberg L, et al. Editorial to the special issue on Big Data in Industrial and Commercial Applications. Big Data Research. 2021;26:100244.

Publisher's Note