Scientometrics (2015) 105:2005–2022
DOI 10.1007/s11192-015-1638-y

# A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'Big Data'

Ying Huang[1,2,3] · Jannik Schuehle[5] · Alan L. Porter[3,4] · Jan Youtie[6]

Received: 30 April 2015 / Published online: 15 July 2015
© Akadémiai Kiadó, Budapest, Hungary 2015

**Abstract** Bibliometric and "tech mining" studies depend on a crucial foundation—the search strategy used to retrieve relevant research publication records. Database searches for emerging technologies can be problematic in many respects, for example the rapid evolution of terminology, the use of common phraseology, or the extent of "legacy technology" terminology. Searching on such legacy terms may or may not pick up R&D pertaining to the emerging technology of interest. A challenge is to assess the relevance of legacy terminology in building an effective search model. Common-usage phraseology additionally confounds certain domains in which broader managerial, public interest, or other considerations are prominent. In contrast, searching for highly technical topics is relatively straightforward. In setting forth to analyze "Big Data," we confront all three challenges—emerging terminology, common usage phrasing, and intersecting legacy technologies. In response, we have devised a systematic methodology to help identify research relating to Big Data. This methodology uses complementary search approaches, starting with a Boolean search model and subsequently employs contingency term sets to further refine the selection. The four search approaches considered are: (1) core lexical query, (2) expanded lexical query, (3) specialized journal search, and (4) cited reference analysis. Of special note here is the use of a "Hit-Ratio" that helps distinguish Big Data elements from less relevant legacy technology terms. We believe that such a systematic

✉ Alan L. Porter
alan.porter@isye.gatech.edu

[1] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China

[2] Lab of Knowledge Management and Data Analysis (KMDA), Beijing Institute of Technology, Beijing 100081, China

[3] School of Public Policy, Georgia Institute of Technology, Atlanta, GA 30313, USA

[4] Search Technology, Inc., Atlanta, GA 30092, USA

[5] Department of Economics and Management, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

[6] Enterprise Innovation Institute, Georgia Institute of Technology, Atlanta, GA 30332, USA

search development positions us to do meaningful analyses of Big Data research patterns, connections, and trajectories. Moreover, we suggest that such a systematic search approach can help formulate more replicable searches with high recall and satisfactory precision for other emerging technology studies.

**Keywords** Search strategy · Lexical query · Citation analysis · Big Data

## Introduction

Publications are intrinsically rich information sources that enable measurement of important aspects of scientific research activities through various bibliometric and 'tech mining' (text analyses of science, technology and innovation information—c.f., Porter and Cunningham 2005; Guo et al. 2015) analyses. A challenging initial step is the ability to systematically delineate research on the emerging technology of interest from that on existing technologies. Some emerging domains are only incrementally separate from their legacy technologies. Often size is a part of this incremental departure—smaller in the case of nanotechnology, for example, or bigger, in the case of Big Data, which is the subject of this paper. Of course, size alone is not always sufficient to define an emerging technology, but it is often accompanied by other characteristics that lead to the disjunction of the domains. It is especially difficult to define the boundary of these types of technologies to harvest metadata about relevant publications and patents.

Tracking emergent sciences and technologies may be of great importance for researchers, social scientists and decision makers, but it often relies on poorly defined data. The dataset may be too large on the one hand (poor precision) and incomplete on the other (weak recall). Furthermore it may even reflect an incorrect balance among the various disciplines that shape the emerging field (Mogoutov and Kahane 2007).

We draw on several Boolean-based search methodologies used to locate relevant publications in research databases (Huang et al. 2011; Gorjiara and Baldock 2014). The first is core lexical query which usually applies a core of related keywords for a first search. It is the most common search method, but a major drawback is its susceptibility to subjectivity when experts are used to define the keyword set (Huang et al. 2011). The second method, expanded lexical query, aims to minimize the input of experts by systematically extracting a set of keywords with close relationships to the targeted topic from the core publications. The keywords are then ranked by their level of relevance to the field, based on the frequencies of their appearance in the core publication set. High-frequency keywords can be assessed variously—via a multi-stage, iterative process (Zucker et al. 2007), using a semi-automated method (noise ratio) (Arora et al. 2013), or, again, through expert opinion (Mogoutov and Kahane 2007). A hybrid lexical-citation method has been proposed by Zitt and Bassecoulard (2006); they harvested literature that cited the "core" literature through "seed" publications and tune threshold parameters that strike a balance between the specificity and the coverage of the publications. Another approach is based on the use of specialized journals in the field under analysis. For example, Leydesdorff and Zhou (2007) offered a methodology that began with a core set of journals and, through citation and network analysis (using betweenness centrality), expanded that core set to ten related journals with the highest impact factors.

Huang et al. (2011) described the strengths and weaknesses of those different search strategies. The strength of lexical queries lies in the ease of implementation, but problematic is their reliance on static keywords to measure a dynamic field. Expanded lexical queries try to minimize the input from experts but the keywords are still given to proficient third parties for validation. A hybrid lexical-citation approach holds appeal, but it demands capability to retrieve an identified core publication set's cited records. That can be an onerous task. Also, setting thresholds within a cited reference set is necessary, but somewhat arbitrary. For example, were one to start with 1000 core publication records, the cited references might well exceed 30,000, with considerable format variation.

Some bibliometric researchers have come to realize that greater attention should be applied to developing a more reasoned search strategy. Kable et al. (2012) presented a 12-step framework for documenting the search strategy prior to undertaking a critique and synthesis of a retrieved literature set. Arora et al. (2013), updating the nanotechnology search approach of Porter et al. (2008), employed feedback channels between the keyword identification process and elicitation of expert opinion to modify a list of keywords by: (1) systematic, semi-automated evaluations of high-occurrence keywords, and (2) interviews, surveys, and other data sources. However, they still heavily depended on the keywords and the method can be characterized as an expanded lexical query.

These methods do not have to be applied in isolation. This paper argues that a framework can be developed to take advantage of the strengths of each method. Keyword

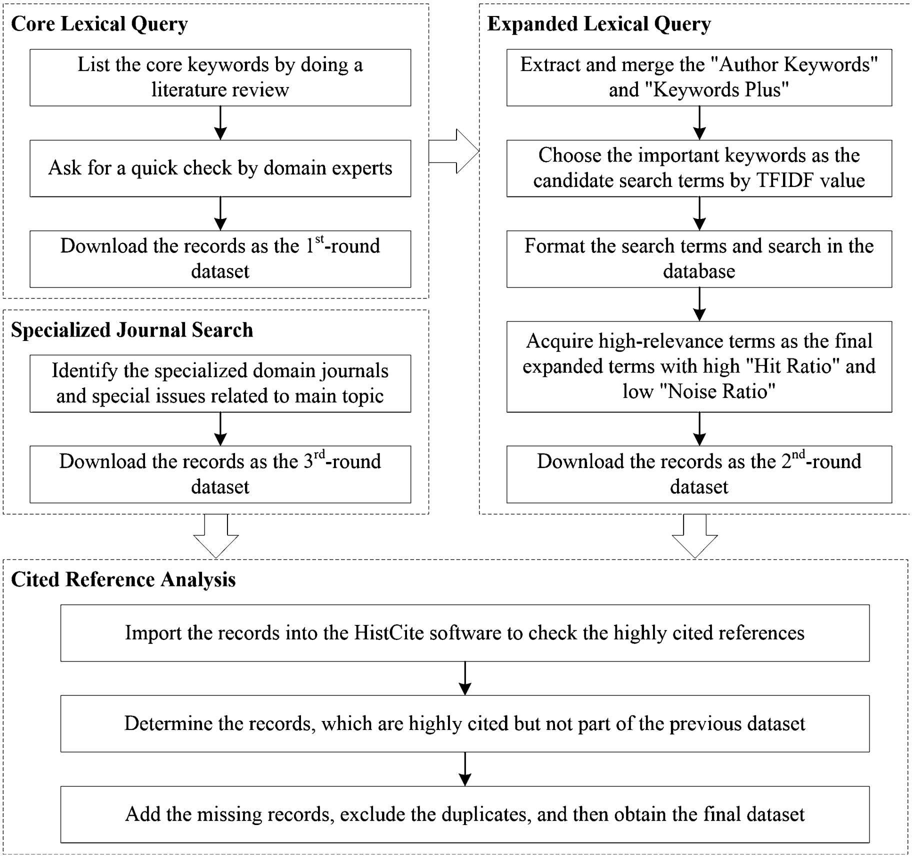
Core Lexical Query
List the core keywords by doing a literature review
Ask for a quick check by domain experts
Download the records as the 1st-round dataset
Specialized Journal Search
Identify the specialized domain journals and special issues related to main topic
Download the records as the 3rd-round dataset
Expanded Lexical Query
Extract and merge the "Author Keywords" and "Keywords Plus"
Choose the important keywords as the candidate search terms by TFIDF value
Format the search terms and search in the database
Acquire high-relevance terms as the final expanded terms with high "Hit Ratio" and low "Noise Ratio"
Download the records as the 2nd-round dataset
Cited Reference Analysis
Import the records into the HistCite software to check the highly cited references
Determine the records, which are highly cited but not part of the previous dataset
Add the missing records, exclude the duplicates, and then obtain the final dataset

**Fig. 1** The framework of our search strategy for science

combinations and classifications underlying search strategies are necessary for targeted scientific and technological domains (Thomas et al. 2011). Citation-based analysis, as a means to identify a corpus of literature embodied in electronic records, is effective in identifying communities of researchers and research streams (Zitt et al. 2011). At the same time, core journal search whose scope falls into the domain field should also be taken into consideration to help identify on-target research publications for a target emerging science or technology domain.

In the past few years, our team has worked with multiple search strategies. We have come to realize that searching in a domain with a prevalence of commonly used terms can be specially challenging. It becomes apparent that it is easier to identify relevant research on a technical topic like Dye-Sensitized Solar Cells (DSSCs), than on one involving a lot of broader computer science or managerial terminology like "Big Data." Our most effective approach is to devise contingency relationships to limit inclusion to papers relating to the domain [e.g., as done by Arora et al. (2013)].

This paper seeks to bring these methods together in a framework designed to distinguish Big Data research publications. The paper is organized as follows: Section "Framework and methodology" introduces the framework and methodology; the suggested approach is exemplified via a search strategy for Big Data, as presented in Section "Case study for Big Data"; Section "Comparative analysis" undertakes a comparative study to examine our results with different strategies proposed in previous research; in Section "Conclusions and discussion" we conclude with a summary, discussion, and ideas for further research.

## Framework and methodology

As indicated above, we propose a framework that combines elements of each of four main search strategies (Fig. 1). The combination begins with core lexical query operations, such as synoptic literature-based discovery of keywords, as checked by experts. It then proceeds to expanded lexical query through the application of a "Hit Ratio" and manual checking to obtain a balance between precision and recall. The next step identifies journals that specialize in the emerging technology and assesses them for possible inclusion in the search process, if they are not included through the first two steps already. The final stage analyzes papers that cite those downloaded in the first three stages to determine if additional keywords gleaned from the titles, abstracts, and authors, keywords of these papers should be considered for inclusion. It is a consecutive process and the order we propose here has proven reliable.

### Core lexical query

Defining the core lexical queries is a first, essential step for us in developing a search strategy. In this stage, reading some domain literature reviews can provide an initial understanding of the field and its major topical thrusts, and help compose an initial list of candidate search terms. Then the list is presented to domain experts to obtain judgments as to whether the terms are on target or not. One important factor is to consider synonyms of the search terms developed in this step and to control for different spellings and appropriate abbreviations. Often these synonyms concern size differentiators from legacy technologies, such as synonyms for "big" in Big Data. Core lexical query is a convenient but incomplete method if used alone, for distinguishing an emerging technology, especially for incrementally differentiated fields such as "Big Data."

## Expanded lexical query

Expanded lexical query, the second step, can help to identify records lying near the boundaries of emerging and legacy technologies. This step makes substantial use of precision and recall indicators rather than relying solely on expert judgment. Although Huang et al. (2011) emphasized the importance of expert judgment, Arora et al. (2013) indicated that experts had difficulty delineating boundary conditions. In this paper, we merge the "Author Keywords" and "KeyWords Plus" obtained from Thomson Reuters Web of Science (WoS) (accessed through their Web of Knowledge) and calculate their TF-IDF (Term Frequency–Inverse Document Frequency—a measure of how specialized the terms are) values to create a set of 100 top candidate terms. In order to decide whether or not to adopt candidate terms as search terms, we introduce the "Hit Ratio" and perform manual checking. We treat the results obtained through the core lexical approach as group A in the equation below; results obtained through combining "contingent" terms (terms that require being found together with another term to yield relevant records) as B; and results obtained from inclusion or exclusion or a candidate term that is on the border between the emerging and legacy technologies as C (Fig. 2). Contingency requirements are vital for Big Data in that many prevalent terms share common usages. The equations representing the "Hit Ratio" can be written as:

$$\text{Hit Ratio} = (\text{A\&B\&C})/(\text{B\&C}) = (\text{A\&B\&C})/((\text{A\&B\&C}) + (\text{B\&C Not A}))$$

The ratios can be used in a two-step process to decide whether the candidate terms should be adopted or not:

(1) Hit Ratio. If Hit Ratio $\geq 70\,\%$: Approve (The candidate term has been approved for inclusion in the search strategy); If Hit Ratio $\leq 30\,\%$: Exclude (The candidate term should be removed from the search strategy); If $30\,\% <$ Hit Ratio $< 70\,\%$: Manual Check required;

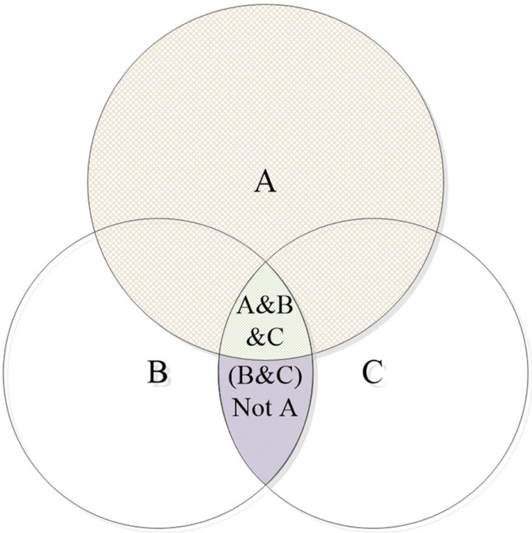

**Fig. 2** Visualization of the groups and the search strategy

(2) A Manual Check can be necessary for terms with a "Hit Ratio" in the range between 30 % and 70 %. We propose looking specifically at the records that lie in the ((B&C) NOT A) area individually. To do so we sample the results in WoS (e.g., sorting by author or other factor—preferably NOT using publication date as term usage evolves importantly over time). We randomly open at least 20 records from different time periods for a further check:

- If more than 50 % deal with the intended science area, we approve the term.
- If <50 % meet the criteria, we exclude the term from the candidate list.

The aim of the expanded lexical query is to obtain a balance between seeking high recall of "core" Big Data research, and moderate recall of "peripheral" research relating to the targeted technology. We are willing to accept a moderate amount of noise (i.e., moderate precision).

## Specialized journal search

The rise of journals specialized for an emerging technology is one of the hallmarks that distinguishes it from legacy technology counterparts. One might expect that the papers in these journals should be identified through the core lexical and expanded lexical strategies, but we have found that some papers appearing in these specialized journals can be missed. The downside is that journal-based methods, as proposed by Leydesdorff and Zhou (2007), can be difficult to implement because new specialized journals require a length of time to become established, indexed, and cited. Likewise, determining which journals to include or not include is a challenge. One reasonable approach is to judge whether or not to include the journal based on the aims and scope section of the journal's website. However, this approach requires considerable manual work and judgment. A reasonable compromise is to apply basic keywords to the title of the specialized journal. This method presumably could be generalized to special issues on the domain topic, although special issues are also difficult to identify through non-manual means (e.g., for "Big Data," a pivotal issue of *Nature*).

## Cited reference analysis

Cooper et al. (2009) showed that citation linkages could express the relevance of works of others to the topic of discussion. Therefore, if a set of records is more highly cited by other publications in a certain domain field, then these records have a greater possibility of

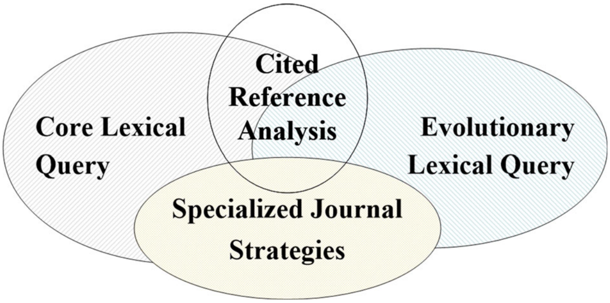

**Fig. 3** Coverage of different search methods

belonging to the same domain field. Based on this idea, we apply a cited reference analysis to determine if there are sets of papers that we have missed through application of the first three steps in the process (core lexical query, expanded lexical query, specialized journal perusal). In this paper, we use *HistCite* [http://interest.science.thomsonreuters.com/forms/HistCite/] to consider direct citation linkages among scientific papers (Garfield et al. 2006) to identify missing sets of papers. Our particular focus is on highly cited papers otherwise not detected in the aforementioned steps.

As previously indicated, we apply these four search strategies to identify papers about Big Data. Our results show that the four methods yield different record coverage. In general, the core lexical query and expanded lexical query comprise the bigger proportion of identified papers, followed by specialized journal search, and cited reference analysis (Fig. 3). Some records may appear in multiple search strategies; these duplicates are readily removed to obtain the final dataset [i.e., removing duplicate ISI Unique identifiers in *VantagePoint* software (www.theVantagePoint.com)].

## Case study for Big Data

The term “Big Data” refers to “analytical technologies that have existed for years but can now be applied faster, on a greater scale, and are accessible to more users” (Miller 2013). The promise of data-driven decision-making is now being recognized more widely, and there is growing enthusiasm for the notion of Big Data (Labrinidis and Jagadish 2012). Data constitute a fundamental resource and how to manage and utilize Big Data better has attracted much attention lately (McAfee and Brynjolfsson 2012). We are undertaking a Forecasting Innovation Pathways (“FIP”) (Robinson et al. 2013) project with a focus on the area of “Big Data & Analytics.” Therefore, we take Big Data as our target field for this case study.

Another dimension in delimiting the search is the choice of database(s). We are exploring multiple databases, but focus in this paper on WoS. In initial interviews, experts indicated that conference papers and proceedings represented key intellectual products in the Big Data research domain. We thus incorporate WoS proceedings citation datasets to enrich coverage beyond journals.

### Core lexical query for Big Data

As the first step of our search strategy, we determine core search terms. Unsurprisingly, “Big Data” itself is central. We directly used “Big Data” as a seed search string applied to the “Topic” field in WoS. and as a basic search in four other databases—*INSPEC, EI Compendex, Derwent Innovation Index* (DII), and *US National Science Foundation (NSF) Awards*. The most common terms that co-occurred with Big Data in the five databases became candidate search terms themselves—“MapReduce” and “Hadoop.” A literature review of bibliometric studies of Big Data surfaced the key work of Park and Leydesdorff (2013), which led us to consider 11 keywords from their search strategy: big data, big science, cloud computing, computational science, cyberinfrastructure, data integration, data mining, data warehouse, Hadoop, MapReduce and NoSQL as candidates for inclusion in our search string.

We next asked Big Data researchers at the Beijing Institute of Technology (BIT) to review these candidate terms. They warned that “data integration,” “data mining,” and “data warehouse” were commonly used in data science and had been discussed for a long time (i.e., legacy terms). They also suggested that “big science,” “cloud computing,” “computational science,” and “cyberinfrastructure” were important terms in Big Data research, but not specialized solely to that domain. These became candidate terms to use contingent on their co-occurrence with other indicator terms of involvement in Big Data research. Furthermore they pointed out that NewSQL provided the same scalable performance of NoSQL systems and thus should be included as a core Big Data term. Therefore, in this stage, we developed a preliminary search string consisting of: “Big Data,” “Hadoop,” “MapReduce,” “NoSQL,” and “NewSQL.” Considering the different writing formats, we applied TS[1] = (“Big Data” or Bigdata or “Map Reduce” or MapReduce or Hadoop or Hbase or Nosql or Newsql) to search in the:

- Science Citation Index Expanded (SCI-Expanded)
- Social Sciences Citation Index (SSCI)
- Arts & Humanities Citation Index (A&HCI)
- Conference Proceedings Citation Index- Science (CPCI-S) and Conference Proceedings Citation Index- Social Science & Humanities (CPCI-SSH)

from 2008 to 2014. [A special issue of *Nature* on Big Data examining “what Big Data sets mean for contemporary science” was published in September, 2008 (Campbell 2008); it is one candidate “starting point” for Big Data research, as distinguished from legacy information and data science.] Inclusion of the conference proceedings is a critical search call—explored further in Section “Comparative analyses”. This preliminary search resulted in 4738 records.

## Expanded lexical query for Big Data

Expanded lexical methods are applied to the preliminary, core Boolean search. First, these methods sought to expand the core term set appropriately. Second, we sought to identify contingency terms—to be used to retrieve records only if certain terms co-occur in those records with a set of “contingent” terms (themselves relating to Big Data interests). Our “contingent” terms are “Big Near/1 Data”, “Huge Near/1 Data”, “Massive Data”, “Data Lake”, “Massive Information”, “Huge Information”, “Big Information”, “Large-scale Data”, Petabyte, Exabyte, Zettabyte, “Semi-Structured Data”, “Semistructured Data” and “Unstructured Data.”

How did we identify contingency terms? We used TS = (“Big data” OR “Bigdata”) to search from 2008 to 2014, and got 2444 records (Search on April 6, 2015) based on a search of three types of publications: proceedings papers, articles, and reviews. The “Author Keywords” and “Keywords Plus” fields from these 2444 records were then merged to get a list of the top 100 terms based on the frequency of appearance of these terms in the records. After removing the terms already used in our core lexical query and some very general phases (like “systems,” “information,” etc.), we got 58 candidate terms. We then applied “Hit Ratio” to those (Table 1). Ten of these terms had Hit Ratios > 70 %. Nine of these terms were included in the final set; HDFS was not included because all of its records where already included in the initial dataset of group A. Forty-one of the candidate lexical terms with Hit Ratios between 30 % and 70 % were reviewed through manual processes; of these, 22 were included in the final set. All terms that were ultimately excluded had Hit Ratios below 60 % but some were included because of having a “Noise

[1] WoS Topical Search (in the Advanced Search feature) captures occurrences in titles, abstracts, authors’ keywords, and Keywords Plus fields. A narrower option considered is provided to search just within titles.

Ratio" (Arora et al. 2013) of 50 % or less, indicating that these terms were highly relevant. Based on this method, we composed our Group C of 31 candidate search terms.

Further review by a Big Data project manager at Ernst and Young suggested that the final search term list might be omitting some important peripheral/emerging technologies from the "Big Data Ecosystem." In particular, he was concerned about omissions of various Apache open source software projects, and some proprietary products, like Apache Spark. After manually checking these possibilities, we found that most of the terms in such a "Big Data Ecosystem" either were already captured by our existing search strategy (particularly by parent technologies such as Hadoop), did not return a significant additional number of records to warrant inclusion, or increased noise by yielding records that actually belonged to other research fields (noting the concern for relatively common usage terms—here, common to broader Computer Science interests).

## Specialized journal search for Big Data

After searching, we identified nine journals that focus on Big Data, including foundational aspects, as well as on specific related platforms and technologies (Table 2). Of these nine journals, three (*Data Science and Engineering, Open Journal of Big Data, American Journal of Big Data Research*) had not yet been published at the time of writing this paper (because they were founded in 2015), four (*Journal of Big Data, International Journal of Big Data Intelligence, Big Data & Society, Big Data Research*) began publishing in 2014, and one (*Big data*) began in 2013. However, these journals were not yet indexed by WoS, so they were not included in our data sources here. In addition to these nine journals, we identified other journals that specialize in "data science," like the *Data Science Journal, Journal of Data Science,* and *EPJ Data Science*. We considered adding these journals but they increased the "Noise Ratio" too much to qualify for inclusion. In future years, as more Big Data specialty journals become indexed by WoS, or as we turn attention to analyses of records from other databases (e.g., *INSPEC*), we would expect to incorporate these journals into an updated search strategy.

## Cited reference analysis for Big Data

After we combined the records obtained from the aforementioned strategies, we imported the data into *HistCite*. We found 171 papers cited more than 20 times. Only 35 % of these were located in WoS and only 35 of those papers were published in the period of 2008–2014 and included in the WoS databases we searched. Of these 35, half were already in our sample. Those records not in our sample tended to be in the biosciences and genomics areas (Table 3). Although those were important application areas for Big Data, we judged them somewhat out of our domain to warrant inclusion.

The final search strategy for this paper is presented in Table 4; it just uses the core lexical query and expanded lexical query. We introduce the specialized journal and cited reference parts of the selection model (Fig. 1) for general interest, and we anticipate incorporating those in future extensions of study of Big Data. But for the reasons just discussed, they don't warrant incorporation in this dataset for the current case analysis.

We assessed the search results to check for desired high recall and satisfactory precision. We randomly chose multiple 10-record samples for each year to assess their relevance to the topic by reading title, abstract, and keywords. We also presented results to topical experts for further review to estimate precision. The results showed that on average nearly 90 % of the retrieved records from 2008 to 2014 had a close relationship with Big

**Table 1** Hit Ratio and Noise Ratio analysis of candidate terms

| No | Keywords | Candidate terms | B&C | A&B&C | Hit Ratio (%) | Preliminary conclusions | Noise Ratio | Final conclusions |
|---|---|---|---|---|---|---|---|---|
| 1 | Cloud computing | Cloud comput* | 406 | 326 | 80.30 | Approved | | Approve |
| 2 | Data mining | Data min* | 528 | 272 | 51.52 | MC | HR | Approve |
| 3 | Model | Model* | 1579 | 817 | 51.74 | MC | LR | Exclude |
| 4 | Algorithm | Algorithm* | 1366 | 626 | 45.83 | MC | LR | Exclude |
| 5 | Classification | Classif* | 490 | 199 | 40.61 | MC | LR | Exclude |
| 6 | Analytics | Analytic* | 637 | 527 | 82.73 | Approved | | Approve |
| 7 | Privacy | Privacy | 192 | 147 | 76.56 | Approved | | Approve |
| 8 | Management | Data manag* | 239 | 144 | 60.25 | MC | HR | Approve |
| 9 | Social media | Social media* | 129 | 112 | 86.82 | Approved | | Approve |
| 10 | Machine learning | Machine learning | 194 | 128 | 65.98 | MC | HR | Approve |
| 11 | Visualization | Visualiz* | 319 | 161 | 50.47 | MC | LR | Exclude |
| 12 | Design | Design* | 1047 | 538 | 51.38 | MC | LR | Exclude |
| 13 | Social networks | Social network* | 186 | 136 | 73.12 | Approved | | Approve |
| 14 | Clustering | Cluster* | 645 | 374 | 57.98 | MC | LR | Exclude |
| 15 | Bioinformatics | Bioinformatics | 122 | 70 | 57.38 | MC | LR | Exclude |
| 16 | Security | Security | 267 | 168 | 62.92 | MC | LR | Approve |
| 17 | Twitter | Twitter* | 89 | 69 | 77.53 | Approved | | Approve |
| 18 | Optimization | Optimiz* | 484 | 264 | 54.55 | MC | LR | Exclude |
| 19 | Ontology | Ontology | 140 | 58 | 41.43 | MC | LR | Exclude |
| 20 | Data analysis | Included | | | | | | |
| 21 | Identification | Identif* | 685 | 322 | 47.01 | MC | LR | Exclude |
| 22 | Prediction | Predict* | 492 | 287 | 58.33 | MC | HR | Approve |
| 23 | Recognition | Recogni* | 259 | 123 | 47.49 | MC | LR | Exclude |
| 24 | Data stream | Stream* | 310 | 153 | 49.35 | MC | HR | Approve |
| 25 | Architecture | Architect* | 541 | 307 | 56.75 | MC | HR | Approve |
| 26 | Distributed computing | Distributed comput* | 115 | 81 | 70.43 | Approved | | Approve |
| 27 | Regression | Regress* | 142 | 75 | 52.82 | MC | LR | Exclude |
| 28 | Selection | Select* | 437 | 190 | 43.48 | MC | LR | Exclude |
| 29 | Semantic web | Semantic* | 345 | 136 | 39.42 | MC | LR | Exclude |
| 30 | Business intelligence | Business intelligence | 73 | 49 | 67.12 | MC | HR | Approve |
| 31 | Data analytics | Included | | | | | | |
| 32 | Data integration | Data integrat* | 88 | 41 | 46.59 | MC | | Exclude |
| 33 | GPU | GPU | 76 | 38 | 50.00 | MC | HR | Approve |
| 34 | Innovation | Innovat* | 216 | 145 | 67.13 | MC | HR | Approve |
| 35 | Text mining | Text min* | 78 | 25 | 32.05 | MC | | Exclude |
| 36 | GIS | GIS | 79 | 27 | 34.18 | MC | HR | Approve |
| 37 | HDFS | HDFS | 51 | 51 | 100.00 | Approved | | |

**Table 1** continued

| No | Keywords | Candidate terms | B&C | A&B&C | Hit Ratio (%) | Preliminary conclusions | Noise Ratio | Final conclusions |
|---|---|---|---|---|---|---|---|---|
| 38 | Real-time | Real-time | 384 | 216 | 56.25 | MC | HR | Approve |
| 39 | Sensor networks | Sensor network* | 92 | 54 | 58.70 | MC | HR | Approve |
| 40 | Data extraction | Data extract* | 40 | 20 | 50.00 | MC | | Exclude |
| 41 | Data management | Included | | | | | | |
| 42 | Integration | Included | | | | | | |
| 43 | Smart grid | Smart grid* | 39 | 27 | 69.23 | MC | HR | Approve |
| 44 | Complex networks | Complex network* | 30 | 20 | 66.67 | MC | HR | Approve |
| 45 | Genomics | Genomics | 75 | 49 | 65.33 | MC | HR | Approve |
| 46 | Parallel computing | Parallel comput* | 98 | 64 | 65.31 | MC | HR | Approve |
| 47 | Support vector machine | Support vector machine OR SVM | 61 | 26 | 42.62 | MC | HR | Approve |
| 48 | Data fusion | Data fus* | 38 | 16 | 42.11 | MC | LR | Exclude |
| 49 | Distributed systems | Distributed | 702 | 444 | 63.25 | MC | HR | Approve |
| 50 | Reliability | Reliab* | 233 | 119 | 51.07 | MC | LR | Exclude |
| 51 | Scalability | Scalab* | 529 | 338 | 63.89 | MC | HR | Approve |
| 52 | Time series | Time serie* | 97 | 53 | 54.64 | MC | HR | Approve |
| 53 | Visual analytics | Included | | | | | | |
| 54 | Data science | Data science | 39 | 39 | 100.00 | Approved | | Approve |
| 55 | Informatics | Informatics* | 114 | 88 | 77.19 | Approved | | Approve |
| 56 | OLAP | OLAP | 44 | 26 | 59.09 | MC | HR | Approve |
| 57 | Predictive analytics | Included | | | | | | |
| 58 | Sentiment analysis | Included | | | | | | |

The asterisk (*) represents any group of characters, including no character

*MC* means “Manual Check”, *HR* presents “High Relevance” with a noise ratio of 50 % or less, *LR* indicates “Low Relevance” with a noise ratio of >50 %

Data, especially if we considered the journal papers. Conference proceedings papers address Big Data techniques that tend to have a close relationship with computer science.

## Comparative analyses

Even through there are several thousand publications that present research on Big Data, just a handful explicitly spell out a Big Data focus. Halevi and Moed (2012) examined the development of research related to Big Data by using the Scopus database from 1970 to

**Table 2** Specialized “Big Data” journals

| Journal | Publisher | Foundation year | Website | Publication periods |
|---|---|---|---|---|
| Big Data | Liebert | 2013 | http://www.liebertpub.com/big | 4 issues per year |
| Journal of Big Data | Springer Open | 2014 | http://www.journalofbigdata.com/ | Irregular |
| International Journal of Big Data Intelligence | Inderscience | 2014 | http://www.inderscience.com/jhome.php?jcode=ijbdi | 4 issues per year |
| Big Data & Society | Sage | 2014 | http://bds.sagepub.com/ | 2 issues per year |
| Big Data Research | Elsevier | 2014 | http://www.journals.elsevier.com/big-data-research/ | Irregular |
| Data Science and Engineering (DSE) | Springer | 2015 | – | – |
| Open Journal of Big Data (OJ”Big Data”) | RonPub UG | 2015 | – | – |
| American Journal of Big Data Research (AJ”Big Data”R) | Ivy Union Publishing (IUP) | 2015 | – | – |

early 2012. Rousseau (2012) performed similar search in WoS based on a TS = “Big Data” search strategy. Park and Leydesdorff (2013) collected papers (journal articles, letters, and reviews) by applying 11 keywords. For a diffuse domain like Big Data there is no absolute standard to gauge the validity of our Big Data search strategy. In this section, we explore the extent to which these different strategies affect the robustness of the analysis results and compare search strategies.

1. TS = (“Big Data”) (from now on called “**ROUSSEAU**”)
2. TI = (“Big Data” OR “Big Science” OR “Cloud Computing” OR “Computational Science” OR “Cyberinfrastructure” OR “Data Integration” OR “Data Mining” OR “Data Warehouse” OR “Hadoop” OR “Mapreduce” OR “Nosql”) (from now on called “**PARK**”)
3. Our search strategy (from now on called “**HUANG**”)

We have interpreted the search strategies from Rousseau, Park (and Leydesdorff), and Huang (and colleagues), and applied them to WoS. Treatment of Rousseau’s search strategy was fairly straightforward in that we applied it to the topic field in the WoS search function. Park and Leydesdorff (2013) applied their search string more narrowly, only to “titles,” “author keywords,” and “Keywords Plus” to retrieve 406 relevant documents from the DVD version of the SCI 2011 database and the SCI(E) 2011 database. To enable us to implement the Park search strategy, we applied the search terms to the Title field alone because WoS does not allow separate searches of keyword fields. We used the three search strategies in the WoS SCI-Expanded database from 2008 to 2014 to facilitate comparability, and we limited the results to English language (for ease of processing). Table 5 presents the results.

**Table 3** Papers highly cited by Big Data papers, but missed by our search strategy

| No | Title | Times cited by WoS | Times cited by our dataset |
|---|---|---|---|
| 1 | Ultrafast And Memory-Efficient Alignment of Short Dna Sequences to The Human Genome | 3597 | 75 |
| 2 | A View of Cloud Computing | 857 | 58 |
| 3 | Cloud Computing and Emerging It Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility | 465 | 52 |
| 4 | Computational Social Science | 406 | 45 |
| 5 | Detecting Influenza Epidemics Using Search Engine Query Data | 499 | 34 |
| 6 | Linked Data—The Story So Far | 237 | 33 |
| 7 | The Sequence Alignment/Map Format and Samtools | 3748 | 31 |
| 8 | Galaxy: a Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in The Life Sciences | 776 | 30 |
| 9 | Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform | 3660 | 29 |
| 10 | Understanding Individual Human Mobility Patterns | 864 | 26 |
| 11 | The Case for Cloud Computing in Genome Informatics | 138 | 26 |
| 12 | LIBVSM: a Library for Support Vector Machines | 776 | 23 |
| 13 | Cloud-Scale RNA-Sequencing Differential Expression Analysis with Myrna | 89 | 23 |
| 14 | Cloud Computing and the DNA Data Race | 77 | 23 |
| 15 | Top 10 Algorithms in Data Mining | 439 | 22 |
| 16 | An Integrated Encyclopedia of DNA Elements in the Human Genome | 1864 | 21 |
| 17 | The Internet of Things: a Survey | 560 | 20 |
| 18 | A Scalable, Commodity Data Center Network Architecture | 163 | 20 |

ROUSSEAU's use of the term "Big Data" alone yields less than half of the records of the multi-term Boolean searches from PARK and HUANG. On the other hand, the broader categories and research areas associated with the simple ROUSSEAU search are consistent with those of PARK and HUANG, although the individual journal titles differ. All three searches emphasize computer science, followed by engineering. It seems that the straightforward ROUSSEAU approach gives greater emphasis to publications by authors based in the US than do the other two. The ROUSSEAU search indicates that half of Big Data publications have authors based in the US compared with 35 % under the PARK definition and 43 % under the HUANG definition. This difference suggests that US publications may well be more apt to use the term "Big Data" because it has become more of a standard term of art in the US than in other countries.

Comparing the PARK and HUANG searches, we see that the PARK search is more than 1.5 times larger than the HUANG search. We found that the use of the term "data mining" accounts for the difference between these two searches. The PARK search included the term "data mining" in a straightforward manner whereas the HUANG search paired it with a size contingency term. If we omit the term "data mining" from the PARK search, the two searches produce roughly the same number of records. When comparing both the PARK and HUANG searches by topical area coverage, the results are very similar. By journal,

**Table 4** The final search strategy for Big Data

| No | Search strategy | Search terms |
|---|---|---|
| 1 | Core lexical query | TS = ("Big Data" or Bigdata or "Map Reduce" or MapReduce or Hadoop or Hbase or Nosql or Newsql) |
| 2 | Expanded lexical query | TS = ((Big Near/1 Data or Huge Near/1 Data) or "Massive Data" or "Data Lake" or "Massive Information" or "Huge Information" or "Big Information" or "Large-scale Data" or Petabyte or Exabyte or Zettabyte or "Semi-Structured Data" or "Semistructured Data" or "Unstructured Data") |
| | | TS = ("Cloud Comput*" or "Data Min*" or "Analytic*" or "Privacy" or "Data Manag*" or "Social Media*" or "Machine Learning" or "Social Network*" or "Security" or "Twitter*" or "Predict*" or "Stream*" or "Architect*" or "Distributed Comput*" or "Business Intelligence" or "GPU" or "Innovat*" or "GIS" or "Real-Time" or "Sensor Network*" or "Smart Grid*" or "Complex Network*" or "Genomics" or "Parallel Comput*" or "Support Vector Machine" or "SVM" or "Distributed" or "Scalab*" or "Time Serie*" or "Data Science" or "Informatics*" or "OLAP") |
| 3 | Specialized journal | The papers published in these specialized journals were not indexed by WoS |
| 4 | Cited reference | The publications, which were cited more than 20 times did not fulfill the criteria for inclusion (see paragraph "Cited Reference Analysis") |

both the PARK and HUANG searches include Future Gen. Comp. Sys., Plos One, and BMC Bioinformatics among their five most prevalent journals; however Expert Systems Applications and Journal of Supercomputing are more prevalent in the PARK search.

A key feature of the final HUANG search (rightmost column of Table 5) is the inclusion of conference papers and proceedings. Indeed, nearly 60 % of the papers in the Big Data domain under the final HUANG search are comprised of conference papers and proceedings. Authors based at institutions in China, in particular, appear more prominent once conference papers and proceedings are included. The HUANG search without conference papers indicates that US-based authors account for 43 % of these papers compared to 21 % based in institutions in China. However, if we include conference papers, this percentage changes to 32 % for US-based authors versus 25 % for authors based in China.

## Conclusions and discussion

In the paper, we propose a four-stage search strategy method that considers core lexical query, expanded lexical query, specialized journal search, and cited reference analysis. We apply our framework to Big Data in the SCI-Expanded database and additional WoS databases, and compare them to previous search strategies used in other bibliometric analyses of "Big Data."

During the process of developing a search strategy, we first obtain core terms that are highly relevant to the domain field, and then ask for expert input to validate the relevance of these terms. Secondly, we extract candidate terms from the keywords of the merged "Author Keywords" and "KeyWords Plus" fields, whose TF-IDF values rank in the top 100, to harvest the 2nd round dataset. In order to filter the final search terms, we apply "Hit Ratios" and Manual Checking ("Noise Ratios") to indicate the relevance of the terms to the domain topic. To complement the search strategies based on core terms and extended terms, we search for specialized journals that focus on the field that we are analyzing.

**Table 5** Comparing different Big Data search strategies

| | ROUSSEAU | PARK | HUANG (for comparison) | HUANG (full search) |
|---|---|---|---|---|
| Database | SCI-Expanded | SCI-Expanded | SCI-Expanded | SCI-Expanded, SSCI, A&HCI, CPCI-S, CPCI-SSH |
| Records | 1162 | 4230 (2320)–Without “Data Mining” in search term | 2681 | 6928 |
| Categories | CS-IS (17.0 %); CS-SE (13.0 %); EEE (9.5 %); CS-TM (9.0 %); MDS (8.8 %); | CS-IS (16.9 %); EEE (13.2 %); CS-AI (12.7 %); CS-TM (11.6 %); CS-SE (11.2 %); | CS-IS (20.7 %); CS-TM (15.5 %); CS-SE (15.3 %); EEE (12.6 %); CS-AI (11.0 %); | CS-TM (32.8 %); EEE (27.1 %); CS-IS (26.6 %); CS-AI (12.8 %); CS-HA (12.6 %); |
| Research Areas | CS (38.1 %); EN (13.8 %); STOT (9.2 %); TE (7.5 %); HCSS (5.2 %); | CS (46.0 %); EN (23.0 %); ORMS (6.8 %); TE (6.5 %); BMB (6.1 %); | CS (50.1 %); EN (17.2 %); BMB (7.3 %); TE (7.1 %); STOT (6.0 %); | CS (63.1 %); EN (31.8 %); TE (7.6 %); BMB (3.0 %); MA (2.8 %); |
| Types | Article (64.2 %); Editorial Mtl (19.1 %); Review (7.5 %); Meeting Abs (3.2 %); News Item (3.0 %); | Article (73.2 %); Editorial Mtl (12.3 %); Meeting Abs (6.4 %); Review (4.0 %); Proc Paper (2.4 %); | Article (80.9 %); Editorial Mtl (9.1 %); Review (5.7 %); Proc Paper (2.6 %); Meeting Abs (1.5 %); | Proc Paper (58.1 %); Article (34.7 %); Editorial Mtl (4.1 %); Review (2.3 %); Meeting Abs (0.6 %); |
| Countries/ Territories | USA (50.0 %); China (17.0 %); England (7.8 %); Germany (6.0 %); Australia (5.0); | USA (35.0 %); China (14.5 %); England (6.6 %); Taiwan (6.1 %); Germany (4.8 %); | USA (42.8 %); China (21.0 %); England (6.9 %); Germany (6.4 %); South Korea (4.9 %); | USA (31.7 %); China (24.9 %); Germany (5.1 %); England (4.4 %); India (3.5 %); |
| Institutions | Chinese Acad Sci (3.2 %); Harvard Univ (2.5 %); Stanford Univ (2.2 %); MIT (2.0 %); UCLA (1.7 %); | Chinese Acad Sci (1.7 %); Univ Illinois (1.064 %); Harvard Univ (1.1 %); Tsinghua Univ (0.9 %); Univ Melbourne (0.9 %); | Chinese Acad Sci (3.3 %); Harvard Univ (1.7 %); Stanford Univ (1.6 %); MIT (1.4 %); UC Berkeley (1.4 %); | Chinese Acad Sci (2.4 %); Tsinghua Univ (1.1 %); MIT (0.9 %); Harvard Univ (0.8 %); UC Berkeley (0.8 %); |
| Source Titles | Computer (2.2 %); Nature (2.1 %); Plos One (1.7 %); Future Gen. Comp. Sys.(1.5 %); Health Affairs (1.5); | Expert Syst. Appl. (3.5 %); J. Supercomput (1.4 %); Future Gen. Comp. Sys. (1.3 %); BMC Bioinformatics (1.3 %); Plos One (1.1 %); | Future Gen. Comp. Sys. (1.8 %); Plos One (1.8 %); BMC Bioinformatics (1.6 %); Concur. Comp.- Practice E (1.3 %); IEEE T. Parallel Distr. (1.2 %); | LNCS (6.2 %); BigData 2013 (2.3 %); Applied Mechanics and Materials (1.9 %); CCIS (1.5 %); Proc. SPIE (1.5 %); |

In the Categories field: *CS-AI* computer science artificial intelligence, *CS-HA* computer science hardware architecture, *CS-IS* computer science information systems, *CS-SE* computer science software engineering, *CS-TM* computer science theory methods, *EEE*-engineering electrical electronic, *MDS* multidisciplinary sciences

In the Research Areas field: BMB—Biochemistry Molecular Biology; CS—Computer Science; EN—Engineering; HCSS—Health Care Sciences Services; MA—Mathematics; ORMS—Operations Research Management Science; STOT—Science Technology Other Topics; TE—Telecommunications

In the Source Titles field: LNCS—Lecture Notes in Computer Science; CCIS—Communications in Computer and Information Science

While we have not identified any pertinent Big Data specialty journal records to include in our current analysis, because the relevant specialty journals are new and have yet to be indexed in WoS, we anticipate that such journals will become a useful addition to the search in future years. And fourth, we consider papers highly cited by Big Data publications as another promising source of relevant research. However, for the present WoS Big Data search, these just add few papers, so are not incorporated in the case analysis. At this point in the evolution of "Big Data," the first two lexical methods appear more suitable for delineating the domain than are named journal and citation based approaches.

We evaluate the resulting search strategy by comparing the outcomes to two other Big Data searches. Big Data is an important emerging technology, not only because of its distinctive evolution from conventional information technologies, but also because Big Data has become a key basis of competition, potentially underpinning new waves of productivity growth, innovation, and consumer surplus (Manyika et al. 2011). The results show that when we search in the same database, we achieve a reasonable balance between high precision and plentiful records compared to the methods proposed by Rousseau (2012) and Park and Leydesdorff (2013). In addition, we find that there are abundant publications in conference papers and proceedings that we feel should be included.

Further work to improve search efficiency and assess search results is certainly warranted. The "Hit Ratio" provides a good method to assess terms' suitability, but our 30 %, 50 %, and 70 % rules are heavily based on our trial and error; they are not rooted in previous literature. Therefore, we suggest further exploration of suitable threshold values to use in tuning lexical queries.

We have performed preliminary searches in several additional databases. We plan to adapt our present Big Data search strategy to each as we pursue particular analyses. Of special note, patent databases (e.g., *DII*) provide categorical information (namely, International Patent Classes, Manual Codes) that can well complement lexical searching. Also, research publication databases (e.g., *INSPEC; Compendex*) offer indexing/categorization information. Combining lexical parameters (term searching) with categorical information enriches the contingency possibilities to generate high precision searches.

Another big challenge is how to assess search results, especially for incrementally distinctive emerging technologies such as "Big Data." Indicators like "recall" and "precision" from the fields of pattern recognition and information retrieval cannot be mechanically applied to judge suitability for a field such as Big Data. The reason is that we don't exactly know how many records should be included in a "complete" set and identifying unrelated records involves considerable work with high residual uncertainty.

We are beginning to explore Big Data's footprint in additional databases. It is worth noting at this point that there is extensive business attention, even as the research thrust remains young (with an arguable initiation point in 2008). We recognize that a search strategy tuned to research literature does not routinely transfer to less technical domains. For instance, in searching Lexis Nexis Newspapers database, the technical terms of Table 4 are largely out of place. We tend to revert to a simple search on the key term—"Big Data"—but realize that other terms and possible categorical information can enrich the simple search, somewhat analogously to Table 4 enriching the technical search.

While we like to think linearly of (1) tuning the search strategy; (2) cleaning the data: and (3) performing the analyses, it is worth noting that iteration and revision are important. Our search model (Fig. 1) explicitly works from initial results to enrich the search via additional terms and heavily cited papers. Our present search provides the basis of a relatively straightforward depiction of Big Data research emphases, with one interesting finding being that two countries dominate the global research (China and the US) (Porter et al. 2015). As noted, our search include conference papers from WoS, so provides a different profile than searches limited to WoS journals (Table 5). But we could have extended to other technical data and/or more contextual treatments (business, popular, policy sorts of information), and that would change national prominence, etc. Our search strategy requires likely revision to be on target in retrieving data from different sources.

Furthermore, the search strategy needs tuning to assure suitability to meet study aims. Different criteria warrant somewhat different cuts on the data. Just to illustrate, a bibliometric or tech mining study might be done, variously, to do research profiling (c.f., Guo et al. 2010), social network analysis (c.f., Wang et al. 2014), semantic network analysis (c.f., Hsu et al. 2013; Danowski and Park 2014), or to generate competitive technical intelligence (c.f., Guo et al. 2015). Precision/recall sensitivities and critical coverage attributes (e.g., key players or emergent topics) will vary, and search strategies need tuning accordingly.

This paper was motivated by our experiences in weighing what would make for a "best" search for Big Data research activity. We felt that a systematic approach to set out search strategy elements and processes to select a viable term set was lacking. Figure 1 is our response that we offer as a model for systematic, empirical search strategy development. The Big Data case presents particular challenges as an emerging technology with legacy technologies and "multi-use terminology." We suggest that this provides a model approach to be adapted to help think through generation and assessment of other emerging technology searches.

**Acknowledgments** We acknowledge support from the US National Science Foundation (Award #1527370—"Forecasting Innovation Pathways of Big Data & Analytics"). Besides, we are grateful for the scholarship provided by the China Scholarship Council (CSC Student ID 201406030005). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation and China Scholarship Council.

## References

Arora, S. K., Porter, A. L., Youtie, J., & Shapira, P. (2013). Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs. *Scientometrics, 95*(1), 351–370.

Campbell, P. (2008). Editorial on special issue on big data: Community cleverness required. *Nature, 455*(7209), 1.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Danowski, J. A., & Park, H. W. (2014). Arab spring effects on meanings for Islamist web terms and on web hyperlink networks among Muslim-majority nations: A naturalistic field experiment. *Journal of Contemporary Eastern Asia, 13*(2), 15–39.

Garfield, E., Paris, S., & Stock, W. G. (2006). HistCiteTM: A software tool for informetric analysis of citation linkage. *Information Wissenschaft und Praxis, 57*(8), 391–400.

Gorjiara, T., & Baldock, C. (2014). Nanoscience and nanotechnology research publications: A comparison between Australia and the rest of the world. *Scientometrics, 100*(1), 121–148.

Guo, Y., Huang, L., & Porter, A. L. (2010). The research profiling method applied to nano-enhanced, thin-film solar cells. *R&d Management, 40*(2), 195–208.

Guo, Y., Zhou, X., Porter, A. L., & Robinson, D. K. R. (2015). Tech mining to generate indicators of future national technological competitiveness: Nano-enhanced Drug Delivery (NEDD) in the US and China. *Technological Forecasting and Social Change, 97*, 168–180.

Halevi, G., & Moed, H. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends, 30*(1), 3–6.

Hsu, C. L., Park, S. J., & Park, H. W. (2013). Political discourse among key Twitter users: The case of Sejong city in South Korea. *Journal of Contemporary Eastern Asia, 12*(1), 65–79.

Huang, C., Notten, A., & Rasters, N. (2011). Nanoscience and technology publications and patents: A review of social science studies and search strategies. *The Journal of Technology Transfer, 36*(2), 145–172.

Kable, A. K., Pich, J., & Maslin-Prothero, S. E. (2012). A structured approach to documenting a search strategy for publication: A 12 step guideline for authors. *Nurse Education Today, 32*(8), 878–886.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5*(12), 2032–2033.

Leydesdorff, L., & Zhou, P. (2007). Nanotechnology as a field of science: Its delineation in terms of journals and patents. *Scientometrics, 70*(3), 693–713.

Manyika, J., Chiu, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*, 60–67.

Miller, H. E. (2013). Big-data in cloud computing: A taxonomy of risks. *Information Research*, *18(1)*. http://InformationR.net/ir/18-1/paper571.html

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy, 36*(6), 893–903.

Park, H. W., & Leydesdorff, L. (2013). Decomposing social and semantic networks in emerging "big data" research. *Journal of Informetrics, 7*(3), 756–765.

Porter, A. L., & Cunningham, S. W. (2005). *Tech mining: Exploiting new technologies for competitive advantage*. New York: Wiley. **[Chinese edition, Tsinghua University Press, 2012]**.

Porter, A. L., Huang, Y., Schuehle, J., & Youtie, J. (2015). *MetaData: BigData research evolving across disciplines, players, and topics*. New York (July): IEEE BigData Congress.

Porter, A. L., Youtie, J., Shapira, P., & Schoeneck, D. J. (2008). Refining search terms for nanotechnology. *Journal of Nanoparticle Research, 10*(5), 715–728.

Robinson, D. K., Huang, L., Guo, Y., & Porter, A. L. (2013). Forecasting Innovation Pathways (FIP) for new and emerging science and technologies. *Technological Forecasting and Social Change, 80*(2), 267–285.

Rousseau, R. (2012). A view on big data and its relation to informetrics. *Chinese Journal of Library and Information Science, 5*(3), 12–26.

Thomas, D. G., Pappu, R. V., & Baker, N. A. (2011). NanoParticle Ontology for cancer nanotechnology research. *Journal of Biomedical Informatics, 44*(1), 59–74.

Wang, X., Li, R., Ren, S., Zhu, D., Huang, M., & Qiu, P. (2014). Collaboration network and pattern analysis: Case study of dye-sensitized solar cells. *Scientometrics, 98*(3), 1745–1762.

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Mmanagement, 42*(6), 1513–1531.

Zitt, M., Lelu, A., & Bassecoulard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields? *Journal of the American Society for Information Science and Technology, 62*(1), 19–39.

Zucker, L. G., Darby, M. R., Furner, J., Liu, R. C., & Ma, H. (2007). Minerva unbound: Knowledge stocks, knowledge flows and new knowledge production. *Research Policy, 36*(6), 850–863.