University of Malaya

September 29, 2017

A Bibliometric Approach to Tracking Big Data Research Trends

Ali Kalantari, *University of Malaya*
Amirrudin Kamsin, *Faculty of Computer Science and Information Technology, University of Malaya*
Halim Shukri Kamaruddin, *Department of Actuarial and Applied Statistics, Faculty of Business & Information Science, USCI University*
Nader Ale Ebrahim
Abdullah Gani, *University of Malaya*, et al.

A Bibliometric Approach to Tracking Big Data Research Trends

Ali Kalantari[1], Amirrudin Kamsin[2], Halim Shukri Kamaruddin[3], Nader Ale Ebrahim[4], Abdullah Gani[5], Ali Ebrahimi[6], Shahaboddin Shamshirband[7,8]

[1,2,5,6] Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
(Ali Kalantari email: alik.research@gmail.com)

[3] Department of Actuarial and Applied Statistics, Faculty of Business & Information Science, USCI University, 56000 Kuala Lumpur, Malaysia (e-mail: halimsukri@ucsiuniversity.edu.my)
[4] Research Support Unit, Centre for Research Services, Institute of Research Management and Monitoring (IPPP), University of Malaya (UM), Kuala Lumpur, Malaysia (e-mail: aleebrahim@um.edu.my)
[7] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[8] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Corresponding email: shahaboddin.shamshirband@tdt.edu.vn

Abstract-The explosive growing number of data from mobile devices, social media, Internet of Things and other applications has highlighted the emergence of big data. This paper aims to determine the worldwide research trends on the field of big data and its most relevant research areas. A bibliometric approach was performed to analyse a total of 6,572 papers including 28 highly cited papers and only papers that were published in the Web of Science™ Core Collection database from 1980 to 19 March 2015 were selected. The results were refined by all relevant Web of Science categories to computer science, and then the bibliometric information for all the papers was obtained. Microsoft Excel version 2013 was used for analyzing the general concentration, dispersion and movement of the pool of data from the papers. The *t*-test and ANOVA were used to prove the hypothesis statistically and characterize the relationship among the variables. A comprehensive analysis of the publication trends is provided by document type and language, year of publication, contribution of countries, analysis of journals, analysis of research areas, analysis of web of science categories, analysis of authors, analysis of author keyword and keyword plus. In addition, the novelty of this study is that it provides a formula from multi-regression analysis for citation analysis based on the number of authors, number of pages and number of references.

Index Terms-Big data; Research trends; Highly cited papers; Citation analysis.

1. INTRODUCTION

The age of big data has arrived, which has a significant role in the current Information Technology (IT) environment [1]. In 2015, there were over 3 billion Internet users around the world [2]. Accordingly, data have become more complex due to the increasing volume of structured and unstructured data with a growing number of various applications produced by the social media, Internet of Things (IoT), and multimedia, and etc. [3, 4]. Commonly scientists have introduced four V's for big data as: volume, velocity, variety and veracity. Meanwhile there is another study by [5] that presented three more V's for big data as: validity, volatility and the special V for value. Researchers have highlighted that the needs for big data are increasing, which will have a powerful impact on computer science, healthcare, society, educational systems, social media, government, economic systems and Islamic studies [6-10]. In [11], the state of the art of big data indexing using intelligent and non-intelligent approaches are investigated in order to show the strength of machine learning techniques in big data.

To identify the gaps on big data research trends in different fields, researchers have to investigate or review the comprehensive sources and databases about the papers published in the field. We found that Web of Science (WoS) is the most completed and well-known online scientific citation search that is provided by Thomson Reuters [12, 13]. Hence, it is a valuable reference for researchers to find and publish the latest technology, trends, enhancements, experimental, challenges and opportunities in research. In 1955, Garfield, E. wrote a paper entitled "Citation index for science: a new dimension in documentation through association of ideas" that introduced contemporary scientometrics. However, the Science Citation Index (SCI) has been used in indexing as a principal tool from 1964 [14-16].

Bibliometric is defined as the application of mathematical and statistical methods to papers, books and other means of communication that are used in the analysis of science publications [17]. To recognize the research trends, bibliometric methods are usually used to evaluate scientific manuscripts [18, 19]. Bibliometric methods have been used to measure scientific progress in many disciplines of science and engineering, and are a common research instrument for the systematic analysis of publications [20-24]. In this research, bibliometric analysis is employed in the field of "big data".

Highly cited papers have a greater chance of visibility, thus attracting greater attention among researchers [25]. Evaluating the top cited publications content is very useful for obtaining information about the trends of specific fields in the perspective of research progress [26]. It can reveal to researchers how they can find the best field or best journal to succeed in their publication. Although, the citation is not a scientific tool to assess the publication, it is a valuable metric that recognizes research parameters [27]. The citation index, as a type of bibliometric method, shows the number of times an article has been used by other papers [28]. Hence, citation analysis helps researchers to obtain a preliminary idea about the articles and research that has an impact in a particular field of interest and it deals with the examination of the documents cited by scholarly works [29, 30]. In addition, there are various

2

bibliometric studies has been evaluated based on different metrics and applications such as forecasting emerging technologies by using bibliometric and patent analysis [31], multiple regression analysis for Japanese patent case [32], medical innovation using medical subject headings [33], based on region or countries [34-36], number of authors [37, 38] and a bibliometric analysis based on number publications and cited references [39].

In this study, bibliometric tools have been selected to determine the major and essential research trends in the field of big data research, and the most relevant research areas upon which big data has a significant impact. The Thomson Reuters' Web of Science (WoS) database is used to extract the bibliometric information for "big data". The WoS is a structured database that indexes selected top publications that, covering the majority of significant scientific results [13]. A total of 6,572 papers were collected from WoS and the aim of this research is to provide a comprehensive analysis and evaluate the latest research trends followed by the Document Type and Language, Publication output, Contribution of Countries, Top WoS Categories and Journals, Top Authors, Top Research Areas and Analysis of Author Keywords and Keyword Plus related to the field of big data and its most relevant research areas. In addition, this paper analyzes the number of citation based on the impact of multiple factors of research paper as number of authors, number of pages and number of references. Therefore, our analysis makes an important contribution to researchers interested in the field of big data because we outline, research trends and identify the most relevant research areas to be taken into consideration when conducting future research on big data. We also provide a wide-ranging analysis on the relevant research areas that is mostly used in the field of big data with emerging research streams. Thus, this study will be useful for researchers to determine the relevant area of research in big data that has been broadly focused on along with the gaps that should be addressed. The rest of this paper is organized as follow: Section 2 discusses proposed methodology. Section 3 provides the results and following by discussions. Section 4 concludes this work.

2. METHODOLOGY

The methodology is based on bibliometric techniques which permit a robust analysis of "Big Data Research" publications at different levels. The proposed methodology depends on the quantity analysis of all publications in the field that were selected based on keywords search in the title of papers. In order to define initial keywords, 30 documents from various sources relevant to the topic of "big data" were reviewed. Based on the interview with experts in the related field, the keywords were modified. The comments and suggestions from the interviewees were used to finalize the keyword list.

To achieve a complete set of keywords, an online questionnaire design and was distributed to about 400 people through posting in "big data" community groups on different social media platforms such as Facebook and LinkedIn. In addition, several questionnaires were sent by email to those authors whose

papers were reviewed. The purpose of the survey was to obtain the participants' comments and then analyze the collected data to illustrate the percentage of correct keywords that had been chosen and other relevant keywords that might be used for this study. The details of the data collection process are illustrated in figures 1 and 2, showing the results obtained from the survey analysis of which a total of 142 responses were received. After comparing with the results of the survey, the final relevant research keywords to the field of big data are shown in Table 1.
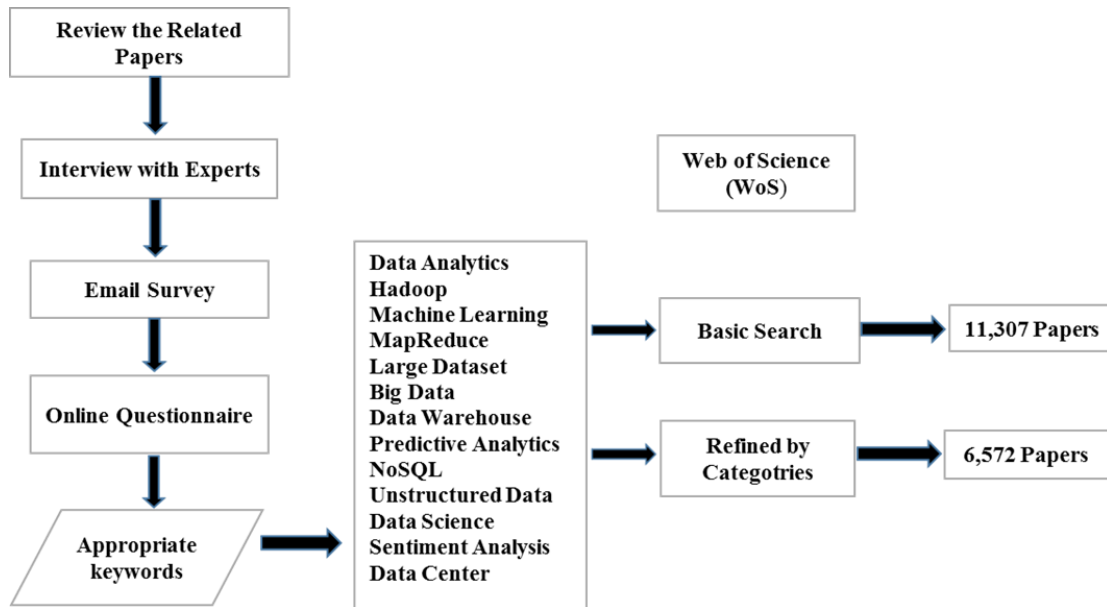

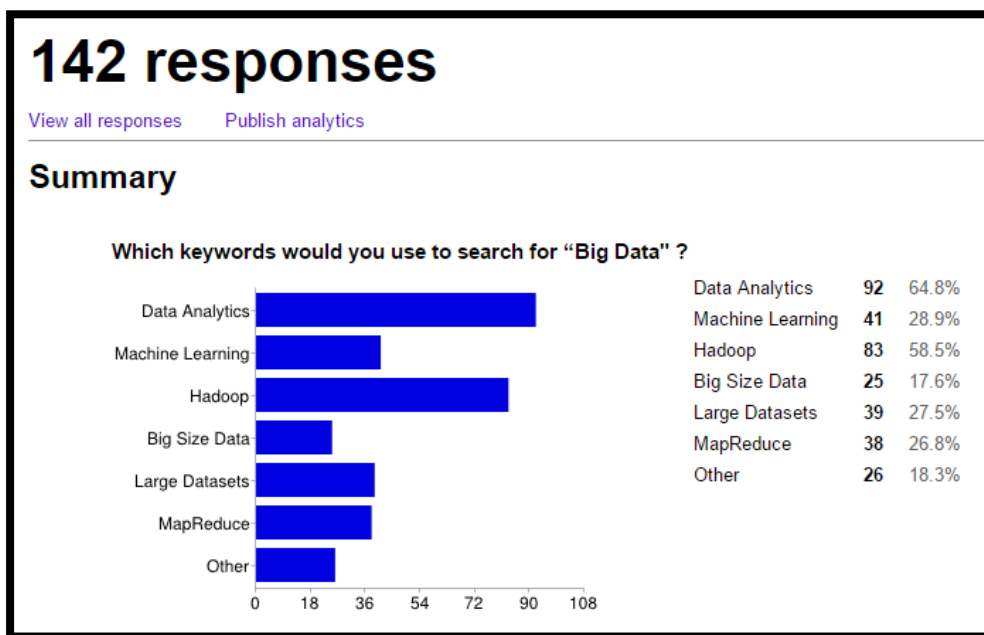
Figure 1: Data Collection Processes

Figure 2: Survey Results (Number of Responses)

4

Table 1: Most Relevant Research Keywords for "Big Data"

Selected Keywords
Data Analytics
Hadoop
Machine Learning
MapReduce
Large Dataset
Big Data
Data Warehouse
Predictive Analytics
NoSQL
Unstructured Data
Data Science
Sentiment Analysis
Data Center

The data for this paper were derived from the online version of the Web of Science^{TM} (WoS) Core Collection database, which consists of Science Citation Index Expanded (SCI), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI) from 1980 to 19 March 2015, and Conference Proceeding Citation Index-Science (COCI-S), and Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH) from 2004 to 19 March 2015. To ensure that the article was relevant to the research topic, the title of the published papers in WoS were scrutinized for the list of keywords in Table 2. Wildcard (*) and Boolean operator (OR) with a combination of keywords was used to purify the results. The preliminary results included 11,307 papers. The results were refined by all relevant WoS categories to computer science and then the bibliometric information for all the papers was obtained (Supplementary file: Table I). The final result consists of data for 6,572 papers, which were downloaded into a Microsoft Excel spreadsheet. There are different ways of calculating author-level impact: number of article citations, number of publications, or combine the publication and citation counts to create a "hybrid indicator" [13]. In this study, the citation counts were selected for evaluation. Following which, 28 highly cited papers were selected according to the Essential Science Indicators^{SM} (ESI) provided by Thomson Reuters [40]. Since the citation rates vary by field and older papers are cited more than recent papers, the selection of highly cited papers is an important issue [41]. The procedure for selection is summarized in (Supplementary file: Table II), and the illustration for the citation report of the highly cited and all papers is shown in table 3. In summary, highly cited papers are the ones that ranked within the top 1% over the past ten years [42, 43].

Table 2: Keywords Phrase and Search Results

Keyword in Title	Search term	Example	Results
Data Analytics	"Data* Analytic*"	Data Analytic/s, Data-analytical, DATABASE ANALYTICS,	264
Hadoop	"Hadoop*"	Hadoop, Hadoop-based, HadoopToSQL, HadoopRDF, HadoopM	312
Machine Learning	"Machine* Learn*"	Machine/s learning, MACHINERY LEARNING, machine learners, Machinery Learners, Machine Learned, machine-learned,	4,466
MapReduce	"MapReduce*" OR "Map$Reduce*"	MapReduce, Map-Reduce,	752
large datasets	"Large$ Dataset*"	large dataset/s, larger datasets	309
Big Data	"Big Data*"	Big Data, Big Datasets, Big Databases	1,310
Data Warehouse	"Data Warehouse*"	Data Warehouse/s	1,200
Predictive Analytics	"Predictive Analytic*"	Predictive Analytic/s	60
No SQL	"No SQL" OR "NoSQL" OR "NoSQL Database"	No-SQL, NoSQL, No SQL	72
Unstructured Data	"Unstructured Data"	Unstructured Data	82
Data Science	"Data Science*"	Data Science/s	46
Sentiment Analysis	"Sentiment Analy*"	Sentiment Analysis, Sentiment Analyzing, Sentiment Analyzer	303
Data Centers	"Data Cent*"	Data Center/s, Data Centre/s, data centric,	2,384
All Above			11,307

Table 3: Citation Report of Highly Cited and All Papers

Item	Highly Cited	All papers
Results founds	28	6,572
Sum of the Times Cited	3,549	32,683
Sum of Times Cited without self-citations	3,540	28,617
Average Citations per Item	126.75	4.97
h-index	19	64

Besides the highly cited papers, which were reported by ESI, the citations per year were calculated as a division of the total citations by life year of the article. The citations per year are more accurate and more scientific than the total citations to identify the top cited papers [44, 45]. Citation statistics produced for a period of less than three years may not be sufficiently stable [46, 47]. Therefore, we only select the papers published up to 19 March 2015 for citation analysis. The rest of the analyses were based on the whole dataset. The emphasis of the research was to describe trends in physical activity and ageing research from the following five aspects:

- Trend of publications during 1980-2015
- Analysis of distribution of author keywords
- Analysis of distribution of KeyWords Plus

- Comparison of papers citation based on author keywords with the KeyWords Plus
- Citation analysis of the research output

The research will provide a guideline based on the publication trend and impact for future research.

After classifying and extracting the data, the process of analyzing was started and the observations include statistical analysis, statistical descriptions, statistical tests, ANOVA, and regression analysis of some of the factors chosen. The ANOVA table was constructed using Microsoft Excel version 2013 [48], which can be used to show the statistical relationship between two variables. In addition, a free version of StatPlanet Plus [49] software was used in this study for creating the interactive world map to visualize the distribution of all papers among different countries.

The four relationships among two variables were observed in this study. The first observation was the relationship between the number of publications of each country in all papers and the impact in a highly cited paper. The second observation was the relationship between the number of the author's publications in all the papers and the impact in highly cited papers. The third observation was the relationship between the number of publication of journals in all papers and the impact in highly cited papers. Microsoft Excel 2013 was used in constructing the ANOVA table and to conclude this two-factor relationship. The fourth observation was the relationship between the author keyword and keyword plus in all and the impact in highly cited papers.

The last part of the analysis focused on the multiple regressions of three factors concerning the number of citations. The three independent variables in this regression were number of authors, number of pages and number of references, form each of the 6,572 papers, and the dependent variable was the number of citations of each paper (out of 6,572). By using Microsoft Excel 2013, the t-test was constructed to analyze how strongly these three factors contribute to the number of citations. At the end of this analysis, this study produced a multiple regression equation, which could be used to forecast the number of citations giving the number of pages, authors and references.

3. RESULTS AND DISCUSSION

As discussed above, the total of 6,572 papers was refined by all WoS categories relevant to computer science (Supplementary file: Table III). Hence, various bibliometric tools to evaluate the different metrics were selected to determine the research trends based on the two groups that extracted from WoS: All Papers (6,572) and Highly Cited Papers (28) groups. In addition, we have not discounted double counting in papers co-authored from multi countries; therefore, the results shown below or in other tables might be counted more than the real data as 28 papers for highly cited papers and 6,572 papers for all papers

Table 4 shows the highly cited papers sorted by publication year, which was used to analyze and compare with the total number of publications (all papers group).

Table 4: List of Highly Cited Papers

Tittle	Authors	Year	NR	TC (Rank)	Refs.
Trends in Big Data Analytics	Kambatla, K; Kollias, G; Kumar, V; Grama,A	2014	75	6 (27)	[50]
Big Data: A Survey	Chen, M; Mao,SW; Liu, YH	2014	155	7 (26)	[6]
A Comparison of Parallel Large-Scale Knowledge Acquisition Using Rough Set Theory on Different MapReduce Runtime Systems	Zhang, JB; Wong, JS; Li, TR; Pan, Y	2014	46	9 (24)	[51]
A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud	Zhang, XY; Yang, LT; Liu, C; Chen, JJ	2014	31	6 (27)	[52]
Data Mining with Big Data	Wu, XD; Zhu, XQ; Wu, GQ; Ding, W	2014	56	12 (23)	[1]
Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis	Balahur, A; Turchi, M	2014	39	9 (24)	[53]
Techniques and Applications for Sentiment Analysis	Feldman, R	2013	39	19 (20)	[54]
New Avenues in Opinion Mining and Sentiment Analysis	Cambria, E; Schuller, B; Xia, YQ; Havasi, C	2013	33	41 (18)	[55]
Review of Performance Metrics for Green Data Centers: A Taxonomy Study	Wang, LZ; Khan, SU	2013	43	18 (21)	[56]
G-Hadoop: MapReduce Across Distributed Data Centers for Data-Intensive Computing	Wang, LZ; Tao, J; Ranjan, R; Marten, H; Streit, A; Chen, JY; Chen, D	2013	39	27 (19)	[57]
Data Center Network Virtualization: A Survey	Bari, MF; Boutaba, R; Esteves, R; Granville, LZ; Podlesny, M; Rabbani, MG; Zhang, Q; Zhani, MF	2013	67	17 (22)	[58]
Business Intelligence and Analytics: From Big Data To Big Impact	Chen, HC; Chiang, RHL; Storey, VC	2012	68	53 (15)	[59]
Energy-Aware Resource Allocation Heuristics for Efficient Management of Data Centers for Cloud Computing	Beloglazov, A; Abawajy, J; Buyya, R	2012	39	88 (12)	[60]
A Survey on Optical Interconnects for Data Centers	Kachris, C; Tomkos, I	2012	64	49 (16)	[61]
Scikit-Learn: Machine Learning in Python	Pedregosa, F; Varoquaux, G; Gramfort, A; Michel, V; Thirion, B; Grisel, O; Blondel, M; Prettenhofer, P; Weiss, R; Dubourg, V; Vanderplas, J; Passos, A; Cournapeau, D;	2011	16	299 (2)	[62]

8

	Brucher, M; Perrot, M; Duchesnay, E				
Lexicon-Based Methods for Sentiment Analysis	Taboada, M; Brooke, J; Tofiloski, M; Voll, K; Stede, M	2011	120	64 (14)	[63]
MapReduce: A Flexible Data Processing Tool	Dean, J; Ghemawat, S	2010	14	110 (11)	[64]
Faster and Better: A Machine Learning Approach to Corner Detection	Rosten, E; Porter, R; Drummond, T	2010	102	156 (7)	[65]
VL2: A Scalable and Flexible Data Center Network	Greenberg, A; Hamilton, JR; Jain, N; Kandula, S; Kim, C; Lahiri, P; Maltz, DA; Patel, P; Sengupta, S	2009	23	121 (10)	[66]
A study of Statistical Techniques and Performance Measures for Genetics-Based Machine Learning: Accuracy and Interpretability	Garcia, S; Fernandez, A; Luengo, J; Herrera, F	2009	46	160 (5)	[67]
Improving the Performance of Predictive Process Modeling for Large datasets	Finley, AO; Sang, HY; Banerjee, S; Gelfand, AE	2009	17	47 (17)	[68]
CloudBurst: Highly Sensitive Read Mapping with MapReduce	Schatz, MC	2009	20	146 (9)	[69]
A Scalable, Commodity Data Center Network Architecture	Al-Fares, M; Loukissas, A; Vahdat, A	2008	33	148 (8)	[70]
MapReduce: Simplified Data Processing on Large Clusters	Dean, J; Ghemawat, S	2008	15	1249 (1)	[71]
Analysis of Interpretability-Accuracy Tradeoff of Fuzzy Systems by Multiobjective Fuzzy Genetics-Based Machine Learning	Ishibuchi, H; Nojima, Y	2007	33	158 (6)	[72]
A Machine Learning Information Retrieval Approach to Protein Fold Recognition	Cheng, JL; Baldi, P	2006	83	86 (13)	[73]
Machine Learning for High-Speed Corner Detection	Rosten, E; Drummond, T	2006	35	251 (3)	[74]
Predicting Subcellular Localization of Proteins Using Machine-Learned Classifiers	Lu, Z; Szafron, D; Greiner, R; Lu, P; Wishart, DS; Poulin, B; Anvik, J; Macdonell, C; Eisner, R	2004	21	193 (4)	[75]

NR: Cited Reference Count, TC: Web of Science Core Collection Times Cited Count, Refs: References

A. *Document Type and Language*

Table 5 illustrates the main distribution of document types in both groups (all and highly cited papers). In the all papers group, proceeding papers (62.73 %) and articles (38.61%) are the main contributors, whereas in the highly cited papers, articles (89.28%) is only the highest contributor. We found that English was the dominant language with 6,549 records (99.65%). Other less significant outputs for all papers were editorial material, review, meeting abstract, news item, book review, letter, correction, software review, book chapter, item about an individual, note, and reprint; for highly cited papers it was the same with lower significant percentages for review and proceeding paper (Table 5).

Even though, proceeding paper and article were the most used document types, we also considered all the document types, as shown in table 5. We believe that each document type has its own intrinsic value and would provide insight on the research trend.

Table 5: Distribution of Document Type

Document Types in General	Percentage (out of 6572)	Document Types for highly Cited Papers	Percentage (out of 28)
Proceedings Paper	62.73	Article	89.28
Article	38.61	Review	10.71
Editorial Material	3.97	Proceedings Paper	10.71
Review	1.01	-	
Meeting Abstract	0.27	-	
News Item	0.27	-	
Book Review	0.24	-	
Letter	0.16	-	
Correction	0.12	-	
Software Review	0.07	-	
Book Chapter	0.06	-	
Item About An Individual	0.06	-	
Note	0.03	-	
Reprint	0.01	-	

B. *Publication Trends: annually, regions/countries, Contribution of Countries*

1) **Publication Output**
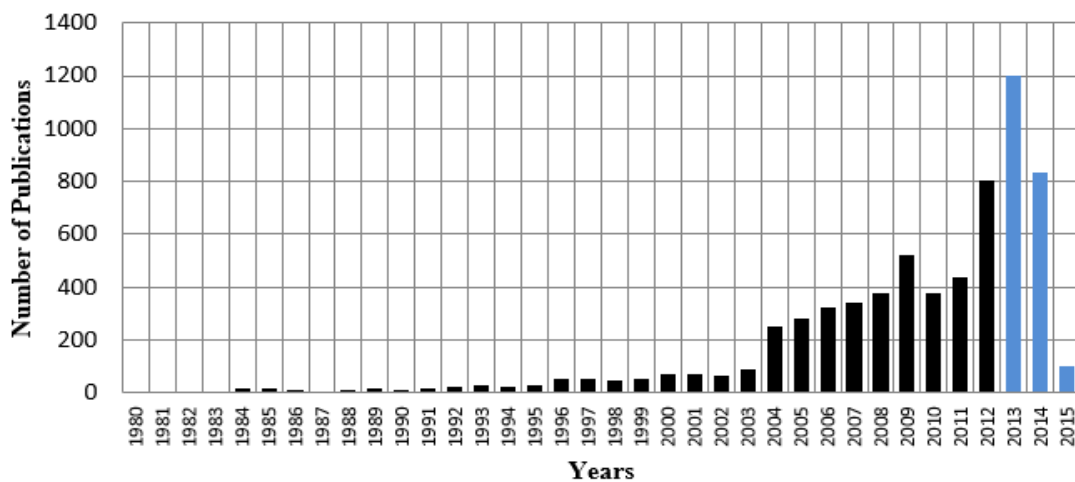
Figure 3: Number of published items from 1980 to 19 March 2015 (citation statistics produced for a period of less than three years may not be sufficiently stable, as indicated in blue color bars)

The recent research concentration is reflected in its publication output [76, 77]. Figure 3 shows the number of published items spanning 36 years from 1980 to 19 March 2015. In general, the number of publications increased over the period studies. A huge positive jump from one year to another can be

10

seen from 2012 to 2013, with a difference of 395 publications. If the last three recent years is excluded from the observation, the highest drop can be observed between 2009 and 2010, the published items dropped by 146 items. In the first 16 years in Figure 3, the count movement of publications steadily increased but with a very small increment. It can be clearly seen from the trend of the graph that there is a sudden jump in the initial period after a long steady movement, which was between 1993 and 1994. On average, the number of papers published every year was 182.56 with a standard deviation of 280.03. It can be concluded that, the shape of the graph is skewed to the right, which consistently agrees with the huge value for standard deviation generated from Excel.

In general, the number of publications is increasing over the years considered. But, in 2009, 2012 and 2013 we observed sudden spikes in the number of publications. Further analysis we found, there were more research activities reporting on proceeding papers in those years compared to their previous year. For instance, there were 556 proceeding papers in 2012 and 260 proceeding papers in 2011.

2) *Contribution of Regions/Countries*
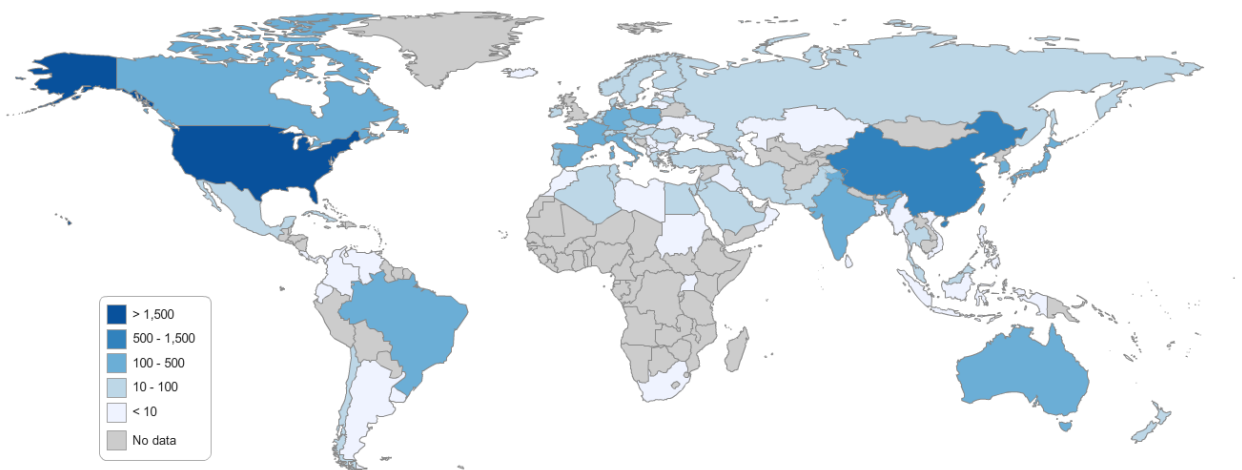
Figure 4: Distribution of All Papers in the World

It has been reported that, "Each author of an article has made an independent contribution to the manuscript and therefore the institution and country the author affiliated could be considered the important contributors for the evaluation of research" [78, 79]. Consequently, the number of publication counts for each country were used to evaluate the research contribution of any region/country in the related field.

Figure 4 shows the geographical distribution of the published papers in the world relevant to the field of big data. As shown in figure 4, the USA (1852) was the most productive country with the largest number of publications regardless of the participation of international collaborators, followed by China

(1059), Germany (303), England (285), Spain (282), Canada (255), India (253), France (226), Italy (198) and Australia (193).

The geographical world map shows that there has been a gradual increase in the number of publications in North and South America, and that it has a higher impact in the world. We found that among the 196 countries in the world, 96 countries such as South Africa and a few countries in the Middle East have no publications. The present results show that big data is a growing area of research in most countries.

C. Analysis of Countries between All and Highly Cited Papers

The comparison of the top ten countries with the highest publications for all and highly cited papers are shown in (Supplementary file: Table III). The top three countries for both groups were the USA, China and Germany with the combined total numbering over 3,000 publications produced. Over 36 years, the average for the USA was 51.44 meaning that there were more than 51 publications. In other words, this was over 22 publications (22.02) higher than China for each year. Our analysis shows that there was huge positive correlation between the number of publications by country in all and highly cited papers. The coefficient of determination between these two groups was 0.90. The ANOVA table, which was extracted from Excel, is shown in (Supplementary file: Table IV). The F-value retrieved from the ANOVA table was 10.54; this was higher than its F-crit. (3.88), which was set at the 5% level of significance. Based on the result obtained from the ANOVA test, the null hypothesis was rejected and there was strong evidence to show that the number of publications for each country for all papers had an impact on the number of publications of the country in highly cited papers.

D. Analysis of Web of Science Categories and Journals between All and Highly Cited Papers

The comparison the top ten WoS categories based on the total number of publications in all and highly cited papers are shown in (supplementary file: Table V). In total, there were 96 WoS categories for which the Computer Science Theory Methods (2624) had the highest number of publications in the all papers group, while, in the highly cited group, the first category of WoS was Computer Science Artificial Intelligence (10). It is evident that computer science categories and related subjects are the leading fields in big data research.

According to Garfield, E. [80], "Journal impact factors generally involve relatively large populations of articles and citations". Among the most key striking metric that can evaluate the contribution of a journal is the Journal Impact Factor [81].

Table 6 depicts the top 10 Journals between all and highly cited papers by their impact factor for 2012 and 2013. Commonly, researchers believe that the high journal impact factor shows the high value of

the journals. But, we found that there were some journals with a lower impact factor that contained the highly cited papers.

Overall, there were 2,866 source titles including Journals and Conferences. The average of every source title was 2.29, showing that there were more than 2 papers published for each source title. The coefficient of determination (r-squared) of these two groups was 0.21 and these two groups (all and highly cited papers) had a weak linear relationship. The constructed ANOVA table was observed and analyzed to show how these two arrays of data are related. Table 7 highlights the results of ANOVA, which was set at the 5% significant level. It is obvious that the null hypothesis is rejected as the F value equals 790.75, which is far greater than F-crit. (3.84). The result shows that there is strong evidence to show that the number of publications of source tittles in all papers has an impact on the number of publications in highly cited papers.

Table 6: Comparison of Top 10 Journals in All and Highly Cited Papers (*the top 2 journals in all papers group is not applicable for year 2012, 2013)

Top 10 Journals in Highly Cited Papers	Impact Factor		Top 10 Journals in All Papers	Impact Factor	
	2012	2013		2012	2013
Bioinformatics	5.323	4.621	Lecture Notes in Computer* Science	N/A	N/A
Communications of The Acm	2.511	2.863	Lecture Notes in Artificial Intelligence*	N/A	N/A
ACM Sigcomm Computer Communication Review	N/A	1.102	Expert Systems With Applications	1.854	1.965
Future Generation Computer Systems The International Journal of Grid Computing and eScience	1.864	2.639	Bioinformatics	5.323	4.621
IEEE Communications Surveys and Tutorials	4.818	6.490	Journal of The American Medical Information Association	3.571	3.932
International Journal of Approximate Reasoning	1.729	1.977	Decision Support Systems	2.201	2.036
Computational Linguistics	0.940	1.468	Communications of The ACM	2.511	2.863
Computational Statistics & Data Analysis	1.304	1.151	ACM Sigcomm Computer Communication Review	N/A	1.102
Computer Speech and Language	1.463	1.812	Neurocomputing	1.634	2.005
IEEE Intelligent Systems	1.930	1.920	Machine Learning	1.454	1.689

Table 7: The Result of ANOVA test for the effect of the number of Publications in All and Highly Cited Source titles

Source of Variation	Sum of Square	Degree of Freedom	Mean Square	F value	Pr(>F)
Between Groups	13988.33	1	13988.33	106.98	7.39E-25
Within Groups	751787.8	5750	130.74		
Total	765776.2	5751			

(F-crit=3.84)

E. Analysis of Authors between All and Highly Cited Papers

The comparison of the top ten authors with the highest number of publications from the all and highly cited papers groups are shown in (supplementary file: Table VI). In total, there were 14,949 authors in all papers and the average was 2.27, meaning that there were more than 2 authors for each paper (1.45 papers for each author with a standard deviation of 1.45). There were five authors with two publications in the highly cited group and 105 authors with one publication. The coefficient of determination (r-squared) for these two groups was 0.02, implying that these two data have a close to zero linear relationship. For further analysis of this case, the ANOVA table was constructed to observe the relationship between the all and highly cited papers. The results of ANOVA, which was set at the 5% significant level is shown in (supplementary file: Table VII). It is obvious that the null hypothesis is rejected, as the F value is equal to 14909.21, which is far greater than F-crit. (3.84). From the observation of ANOVA (supplementary file: Table VII), the null hypothesis was rejected and there was strong evidence to show that the number of publications of authors for the all papers had an impact on the number of publications of authors in highly cited papers.

F. Analysis of Research Areas between All and Highly Cited Papers

This section provides the research areas between all and highly cited papers, which were sorted by the number of records based on the total publications. Overall, there were 54 research areas for the all papers group and 11 research areas for the highly cited group. As shown in (supplementary file: Table VIII), the top two research areas for both groups were Computer Science and Engineering, with the highest number of records and the respective number of publications in other research areas showing that there are slightly different research areas compared to each group. The result shows that the most relevant research areas were in Computer Science with the highest number of publications. However, there were other important research areas with a lower number of records. For instance, in the research areas for the all papers group, this was followed by "Telecommunications" (527), "Operations Research Management Science" (194), "Medical Informatics" (184) and for the highly cited group it was followed

by "Mathematics" (4), "Biochemistry Molecular Biology" (3), and "Biotechnology Applied Microbiology" (3).

G. *Analysis of Author Keywords and KeyWords Plus*

Author keyword is one of the essential types of information about the research trends from the view of researchers and has been proven to be important for monitoring the development of science [82-84].

In this study, 10,002 author keywords were used for the analysis from 1980 to 19 March 2015. Table 8 depicts the top 20 frequency for the author keywords used in all papers. The author keywords were compared by the total number of records for three different periods. The distribution of each keyword would assist the researchers to identify the importance of each author keyword used in different years or decades. Accordingly, the result shows that among the top 20 frequency for author keywords, only a few keywords were used between 1980 and 1999. These included "machine learning", "data warehouse", "data mining", "classification", "neural networks" and "clustering". However, over the past 15 years the number of keywords increased. For instance, the frequency of use of "machine learning" was only 6 % between 1980 and 1999 and from 2000 to 2009 it was 40 %, but, 53 % of total publications were used between 2010 and 2015. This means that more focus was given for each keyword from 2000 to 2015. In addition, machine learning was used as the highest frequency of author keywords which has added a big value to big data. The main objective of machine learning is to learn from data in order to make a suitable decision. The mean of data in term of "big data" refers to complex data which are not easy to process in a single machine learning platform. Therefore, the need of a platform such as Hadoop to run machine learning for big data is essential [85].

In another example, the keyword of "big data" itself was used from 2010 to 2015. On the other hand, table 8 reveals that there was a pivotal role for the top frequent author keywords from 2000 to 2015. Thus, researchers are able to evaluate the latest trends of the research based on the top frequent author keywords relevant to the field in any particular decade to see the efficiency of any keyword that might be used to extend the research study. In addition, "mapreduce", "data warehouse", "big data", "hadoop" and "cloud computing" had high growth in the ranking of author keyword frequency.

**Table 8:** Top 20 Highest Frequency of Author Keywords Used (From 1980-2015)

Author Keyword	1980-2015 TP	1980-1999 TP (%)	2000-2009 TP (%)	2010-2015 TP (%)
Machine Learning	757	48(0.06)	304(0.40)	405(0.53)
MapReduce	514	N/A	24(0.04)	490(0.95)
Data Warehouse(s)/Warehousing	353	11(0.03)	215(0.60)	127(0.35)
Big Data	292	N/A	N/A	292(1)
Hadoop	236	N/A	5(0.02)	231(0.97)
Cloud Computing	232	N/A	4(0.01)	228(0.98)
Data Center(s)	232	N/A	40(0.17)	192(0.82)
Data Mining	181	4(0.02)	80(0.44)	97(0.53)
Support Vector Machine(s)	180	N/A	64(0.35)	116(0.64)
Sentiment Analysis	147	N/A	6(0.04)	141(0.95)
Classification(s)/Classifier(s)	112	4(0.03)	53(0.47)	55(0.49)
Neural Network(s)	85	9(0.10)	41(0.48)	35(0.41)
Performance	84	N/A	14(0.16)	70(0.83)
Energy Efficiency	84	N/A	4(0.04)	80(0.95)
Online Analytic(al) Processing (OLAP)	77	N/A	47(0.61)	30(0.38)
Virtualization	64	N/A	14(0.21)	50(0.78)
Feature Selection	57	N/A	28(0.49)	29(0.50)
Cluster/Clustering	54	2(0.03)	16(0.29)	36(0.66)
Opinion Mining	59	N/A	5(0.10)	44(0.89)
Scheduling	47	N/A	5(0.10)	42(0.89)

TP: total number of publications, %: the percentage of keywords over total number of records

Furthermore, another metric was used to evaluate the publications based on the title which is known as KeyWords Plus. It has been proven that [86, 87], "KeyWords plus, which provides search terms extracted from the titles of papers cited in each new article in the ISI database, is an independent supplement for title-words and author keywords".

Figure 5 provides the top 10 most frequently used terms from KeyWords Plus used over the total number of 3,750 keywords. Based on table 8 and figure 5, the most similar keywords among the most frequently used terms from author keyword and KeyWord Plus were "classification(s)/classifier(s)", "neural

network(s)", "support vector machine(s)", "Performance" and "mapreduce", which played a pivotal role over all keywords. In addition, there were also other keywords with a significant growth in percentage, which were not in the list of both author keyword and KeyWord Plus. This means that several types of analysis were used in big data.
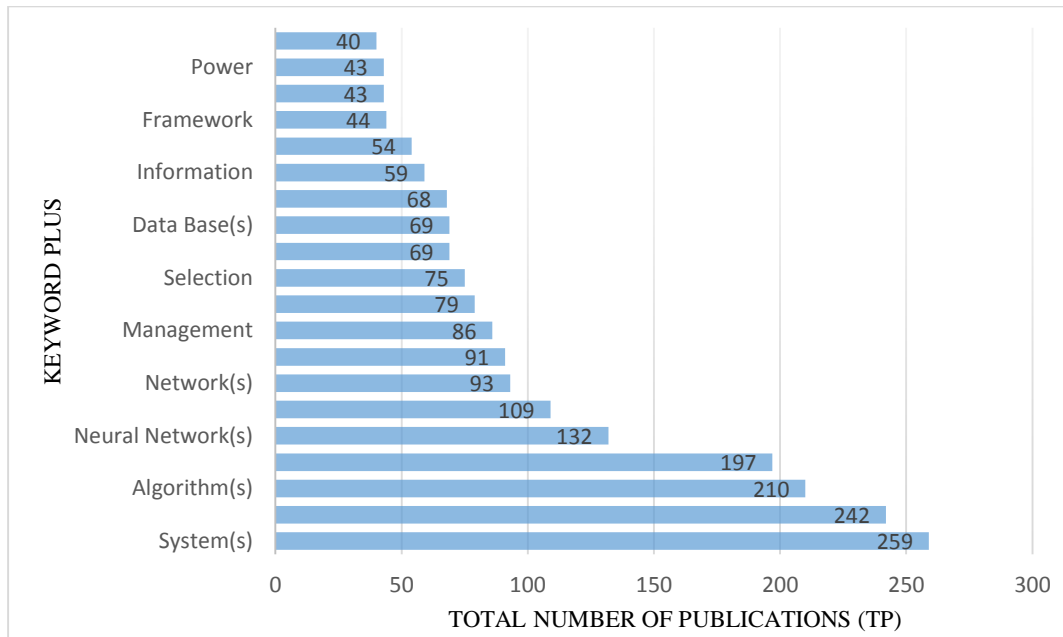
Figure 5: Top 10 Most Frequently used terms of KeyWords Plus (From 1980-2015)

Moreover, the ANOVA test was used to provide the analysis between the author keywords and KeyWords Plus. As mentioned above, there were 10,002 author keywords and 3,750 KeyWords Plus in all papers (6,572). From the total of author keywords, the average of a single author keyword was 1.84 (with a standard deviation of 8.35), meaning that there was more than one author keyword for each paper. The high value of standard deviation contributed from the wide range of frequency of data, or, in other words the frequency distribution was widely spread. In addition, from the total of KeyWords Plus, the average for a single KeyWords Plus was 2.16 (with a standard deviation of 5.93), showing that there were more than two KeyWords Plus for each paper. The reason for the standard deviation of KeyWords Plus being lower compared to the author keywords' standard deviation is that the distribution of KeyWords Plus was more condensed and concentrated. The coefficient of determination (r-squared) for these two groups, as retrieved from Excel was 0.012, which indicates that the two sets of data had a very low linear relationship. Table 9 shows the ANOVA table of KeyWords Plus and author keywords data set with 5% significant level. From the outcome of the ANOVA table, the null hypothesis is rejected as F value equals 121.97, which is greater than F-crit. (3.84). As a summary, there is strong evidence to shows that the frequency of author keywords had an impact on the KeyWords Plus in the total number of publications.

17

Table 9: Results of ANOVA test for effect of Author Keyword and KeyWords Plus

Source of Variation	Sum of Square	Degree of Freedom	Mean Square	F value	Pr(>F)
Between Groups	3956.402	1	3956.40	121.97	2.69E-28
Within Groups	852197.5	26274	32.43		
Total	856153.9	26275			

(F-crit=3.84)

H. Multi-Regression Analysis

Three factors were used to observe the multiple regression for each paper that is, the number of pages (NP), the number of references (NR) and the number of authors (NA). Microsoft Excel version 2013 was used for the regression analysis to analyze the effect of these factors on the Number of Citation (NC) over the total number of papers. The result obtained from the analysis is shown in table 10, to determine how each single factor contributes to the value of the dependent variable. The coefficient column in Table 10 shows the value of the percentage of effect of each factor on the number of citations. Below is a multiple regression equation produced from Excel, which correlates to the number of citation formula:

$$NC = -2.47204 - 0.36744\,NA + 0.38623\,NP + 0.226613\,NR$$

If we are given the number of authors, number of pages and number of references, then the number of citations can be predicted using this multiple regression. The t-test for each of the coefficient is displayed in Table 10. For the number of pages and number of references the p-values are very small and acceptable for factors in the multiple regression equation. For the number of authors the coefficient at the 10% significant level was rejected (but it was not rejected at the 5% significant level). Consequently, for this given multiple regression, all the mentioned factors were acceptable to determine the number of citations with the above given weightage. Therefore, all of the factors had a very small P-value except the number of pages factor which needed a smaller significant level to not be rejected.

Table 10: Result of t-test for the effect of the number of publications in all and highly cited papers

	Coefficients	Standard Error	t Stat	P-value
Intercept	-2.47	0.98	-2.49	0.012
Number of Authors (NA)	-0.36	0.21	-1.74	0.081
Number of Pages (NP)	0.38	0.07	5.12	3.09E-07
Number of References (NR)	0.22	0.02	8.40	5.03E-17

4. CONCLUSION

The bibliometric analysis of the big data revealed that the worldwide research trends and performance in the subject areas. So far, there are significant gaps in current research about the bibliometric analysis of the big data. In this study, selected keywords were used to extract the most relevant papers from the Web of Science^{TM} (WoS) Core Collection database, which consists of Science Citation Index Expanded (SCI), Social Science Citation Index (SSCI), and Arts & Humanities Citation Index (A&HCI) from 1980 to 19 March 2015, and Conference Proceeding Citation Index-Science (COCI-S), and Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH) from 2004 to 19 March 2015. In total, 6,572 papers (including 28 highly cited papers reported by ESI) were refined by all relevant WoS categories to computer science and then the bibliometric information of all the papers was obtained. In total, 2,866 source tittles including journals and conferences were listed in 96 web of science categories with the 14 different document types. English was the dominant language with 6549 records (99.65%) and the five most popular trends areas were Computer Science, Engineering, Telecommunications, Operations Research Management Science, and Medical Informatics. The findings showed that the USA, China and Germany were the most productive countries that played a predominant role in this study with the highest number of published papers in the world. However, other top countries as mentioned above, made a significant contribution to the field. In contrast, there is an essential lack of research to the field of big data in 96 countries with no publications.

The analysis of authors among the all and highly cited papers showed that there were 14,949 authors in all papers, and the average was 2.27, meaning that there were more than two authors for each paper. In addition, there were five authors with two publications and 105 authors with one publication in the highly cited group. In addition, the analysis of the author keywords showed that among the top 20 most frequent keywords there were fewer records from 1980-1999. However, the most keywords were used from 2000 to 2015. The results obtained from the analysis of KeyWords Plus revealed that similar keywords were "classification(s)/classifier(s)", "neural network(s)", "support vector machine(s)", "performance", and "mapreduce" which played a pivotal role over all keywords. The analysis of top 20 highest frequency of author keywords used from 1980-2015, shows that all to 20 fields are increasing and none of them declining during the mentioned period. Moreover, the multi-regression analysis of the number of pages, number of references and number of authors compared to the number of citations were analyzed and the correlation formula is provided in this study.

The findings of this study provide relevant researchers with a panorama of worldwide big data research and the established direction for further study to the field and most relevant research areas.

REFERENCES

1.   Wu, X., et al., *Data mining with big data.* Knowledge and Data Engineering, IEEE Transactions on, 2014. **26**(1): p. 97-107.

2.   Banks, R. *There are now 3 billion Internet users worldwide in 2015*. Mobile Industry Review 2015; Available from: http://www.mobileindustryreview.com/2015/01/3-billion-internet-users-2015.html.

3.   Hashem, I.A.T., et al., *The rise of "big data" on cloud computing: Review and open research issues.* Information Systems, 2015. **47**: p. 98-115.

4.   Diaz, M., et al. *Big Data on the Internet of Things*. in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. 2012.

5.   Khan, M., M.F. Uddin, and N. Gupta. *Seven V's of Big Data understanding Big Data to extract value*. in *American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the*. 2014. IEEE.

6.   Chen, M., S. Mao, and Y. Liu, *Big data: A survey.* Mobile Networks and Applications, 2014. **19**(2): p. 171-209.

7.   Menacer, M., A. Menacer, and A. Arbaoui, *Islamic Resources Big Data mining, Extraction and Archiving.* Enhanced Research in Management & Computer Applications, 2014. **3**(12): p. 20-25.

8.   Murdoch, T.B. and A.S. Detsky, *The inevitable application of big data to health care.* Jama, 2013. **309**(13): p. 1351-1352.

9.   Michael, K. and K.W. Miller, *Big Data: New Opportunities and New Challenges [Guest editors' introduction].* Computer, 2013. **46**(6): p. 22-24.

10.   Xiang, Z., et al., *What can big data and text analytics tell us about hotel guest experience and satisfaction?* International Journal of Hospitality Management, 2015. **44**: p. 120-130.

11.   Gani, A., et al., *A survey on indexing techniques for big data: taxonomy and performance evaluation.* Knowledge and Information Systems, 2016. **46**(2): p. 241-284.

12.   Drake, M., *Encyclopedia of library and information science*. Vol. 1. 2003: CRC Press.

13.   Wildgaard, L., *A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar.* Scientometrics, 2015: p. 1-34.

14.   Garfield, E., *Citation indexes for science: A new dimension in documentation through association of ideas.* Science, 1955. **122**(3159): p. 108-111.

15.   Ho, Y.-S., *The top-cited research works in the Science Citation Index Expanded.* Scientometrics, 2013. **94**(3): p. 1297-1312.

16.   Garfield, E., *Science Citation Index-A new dimension in indexing.* Science, 1964. **144**(3619): p. 649-654.

17.   Repanovici, A. *Measuring the visibility of the University's scientific production using GoogleScholar," Publish or Perish" software and Scientometrics*. in *world library and information congress: 76th ifla general conference and assembly (10-15August 2010, Gothenburg, Sweden)*. 2010.

18.   Zitt, M., S. Ramanana-Rahary, and E. Bassecoulard, *Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation.* Scientometrics, 2005. **63**(2): p. 373-401.

19.   Li, L.L., et al., *Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006.* Scientometrics, 2009. **80**(1): p. 39-58.

20.   Ale Ebrahim, N., et al., *Visibility and Citation Impact.* International Education Studies, 2014. **7**(4): p. 120-125.

21.   Budd, J.M., *A Bibliometric Analysis of Higher-Education Literature.* Research in Higher Education, 1988. **28**(2): p. 180-190.

22. Canas-Guerrero, I., et al., *Bibliometric analysis in the international context of the "Construction & Building Technology" category from the Web of Science database.* Construction and Building Materials, 2014. **53**: p. 13-25.

23. Canas-Guerrero, I., et al., *Bibliometric analysis of research activity in the "Agronomy" category from the Web of Science, 1997-2011.* European Journal of Agronomy, 2013. **50**: p. 19-28.

24. Ingwersen, P., *The international visibility and citation impact of Scandinavian research articles in selected Social Science fields: The decay of a myth.* Scientometrics, 2000. **49**(1): p. 39-61.

25. Wohlin, C., *An analysis of the most cited articles in software engineering journals—1999.* Information and Software Technology, 2005. **47**(15): p. 957-964.

26. Fardi, A., et al., *Top-cited Articles in Endodontic Journals.* Journal of Endodontics, 2011. **37**(9): p. 1183-1190.

27. Shadgan, B., et al., *Top-cited articles in rehabilitation.* Archives of physical medicine and rehabilitation, 2010. **91**(5): p. 806-815.

28. Fooladi, M., et al., *Do Criticisms Overcome the Praises of Journal Impact Factor?* Asian Social Science, 2013. **9**(5): p. 176-182.

29. Ale Ebrahim, N., et al., *Equality of Google Scholar with Web of Science Citations: Case of Malaysian Engineering Highly Cited Papers.* Modern Applied Science, 2014. **8**(5): p. 63-69.

30. Gomez-Jauregui, V., et al., *Information management and improvement of citation indices.* International Journal of Information Management, 2014. **34**(2): p. 257-271.

31. Daim, T.U., et al., *Forecasting emerging technologies: Use of bibliometrics and patent analysis.* Technological Forecasting and Social Change, 2006. **73**(8): p. 981-1012.

32. Yoshikane, F., *Multiple regression analysis of a patent's citation frequency and quantitative characteristics: The case of Japanese patents.* Scientometrics, 2013. **96**(1): p. 365-379.

33. Leydesdorff, L., D. Rotolo, and I. Rafols, *Bibliometric perspectives on medical innovation using the Medical Subject Headings of PubMed.* Journal of the Association for Information Science and Technology, 2012. **63**(11): p. 2239-2253.

34. Bornmann, L., C. Wagner, and L. Leydesdorff, *BRICS countries and scientific excellence: A bibliometric analysis of most frequently cited papers.* Journal of the Association for Information Science and Technology, 2015. **66**(7): p. 1507-1513.

35. Kozak, M., L. Bornmann, and L. Leydesdorff, *How have the Eastern European countries of the former Warsaw Pact developed since 1990? A bibliometric study.* Scientometrics, 2015. **102**(2): p. 1101-1117.

36. Zhou, P. and L. Leydesdorff, *Chemistry in China–A bibliometric view.* Chimica Oggi-Chemistry Today, 2009. **27**(6): p. 19-22.

37. Abramo, G. and C.A. D'Angelo, *The relationship between the number of authors of a publication, its citations and the impact factor of the publishing journal: Evidence from Italy.* Journal of Informetrics, 2015. **9**(4): p. 746-761.

38. Fox, C.W., C.E. Paine, and B. Sauterey, *Citations increase with manuscript length, author number, and references cited in ecology journals.* Ecology and Evolution, 2016: p. n/a-n/a.

39. Bornmann, L. and R. Mutz, *Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references.* Journal of the Association for Information Science and Technology, 2015. **66**(11): p. 2215-2222.

40. Huang, M.-H., H.-W. Chang, and D.-Z. Chen, *Research evaluation of research-oriented universities in Taiwan from 1993 to 2003.* Scientometrics, 2006. **67**(3): p. 419-435.

41. *ESI.* 2015 [cited 2015 2 June]; Web Of Science Core Collection Help Essential Science Indicators Highly Cited Papers]. Available from: http://images.webofknowledge.com/WOKRS517B4/help/WOS/hs_citation_applications.html#dsy7851-TRS_highly_cited_papers.

42. Bornmann, L., et al., *A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication?* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2011. **174**(4): p. 857-879.

43.   Fu, H.-Z., et al., *Characteristics of research in China assessed with Essential Science Indicators.* Scientometrics, 2011. **88**(3): p. 841-862.

44.   Chuang, K.Y., M.H. Wang, and Y.S. Ho, *High-impact papers presented in the subject category of water resources in the essential science indicators database of the institute for scientific information.* Scientometrics, 2011. **87**(3): p. 551-562.

45.   Ho, J.C., et al., *Technological barriers and research trends in fuel cell technologies: A citation network analysis.* Technological Forecasting and Social Change, 2014. **82**: p. 66-79.

46.   Adams, J., *Early citation counts correlate with accumulated impact.* Scientometrics, 2005. **63**(3): p. 567-581.

47.   UZUN, A., *Statistical relationship of some basic bibliometric indicators in scientometrics research*, in *International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*. 2006: Nancy (France). p. 5.

48.   *Microsoft Excel* 2013   [cited 2015 April]; Available from: https://products.office.com/en-us/excel.

49.   *StatPlanet Plus*.   [cited 2015 April]; Available from: http://www.statsilk.com/software/statplanet.

50.   Kambatla, K., et al., *Trends in big data analytics.* Journal of Parallel and Distributed Computing, 2014. **74**(7): p. 2561-2573.

51.   Zhang, J., et al., *A comparison of parallel large-scale knowledge acquisition using rough set theory on different MapReduce runtime systems.* International Journal of Approximate Reasoning, 2014. **55**(3): p. 896-907.

52.   Zhang, X., et al., *A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud.* Parallel and Distributed Systems, IEEE Transactions on, 2014. **25**(2): p. 363-373.

53.   Balahur, A. and M. Turchi, *Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis.* Computer Speech & Language, 2014. **28**(1): p. 56-75.

54.   Feldman, R., *Techniques and applications for sentiment analysis.* Communications of the ACM, 2013. **56**(4): p. 82-89.

55.   Cambria, E., et al., *New avenues in opinion mining and sentiment analysis.* IEEE Intelligent Systems, 2013. **28**(2): p. 15-21.

56.   Wang, L. and S.U. Khan, *Review of performance metrics for green data centers: a taxonomy study.* The Journal of Supercomputing, 2013. **63**(3): p. 639-656.

57.   Wang, L., et al., *G-Hadoop: MapReduce across distributed data centers for data-intensive computing.* Future Generation Computer Systems, 2013. **29**(3): p. 739-750.

58.   Bari, M.F., et al., *Data center network virtualization: A survey.* Communications Surveys & Tutorials, IEEE, 2013. **15**(2): p. 909-928.

59.   Chen, H., R.H. Chiang, and V.C. Storey, *Business Intelligence and Analytics: From Big Data to Big Impact.* MIS quarterly, 2012. **36**(4): p. 1165-1188.

60.   Beloglazov, A., J. Abawajy, and R. Buyya, *Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing.* Future generation computer systems, 2012. **28**(5): p. 755-768.

61.   Kachris, C. and I. Tomkos, *A survey on optical interconnects for data centers.* Communications Surveys & Tutorials, IEEE, 2012. **14**(4): p. 1021-1036.

62.   Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* The Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.

63.   Taboada, M., et al., *Lexicon-based methods for sentiment analysis.* Computational linguistics, 2011. **37**(2): p. 267-307.

64.   Dean, J. and S. Ghemawat, *MapReduce: a flexible data processing tool.* Communications of the ACM, 2010. **53**(1): p. 72-77.

65.   Rosten, E., R. Porter, and T. Drummond, *Faster and better: A machine learning approach to corner detection.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010. **32**(1): p. 105-119.

66.   Greenberg, A., et al. *VL2: a scalable and flexible data center network*. in *ACM SIGCOMM computer communication review*. 2009. ACM.

67.   García, S., et al., *A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability.* Soft Computing, 2009. **13**(10): p. 959-977.

68.   Finley, A.O., et al., *Improving the performance of predictive process modeling for large datasets.* Computational statistics & data analysis, 2009. **53**(8): p. 2873-2884.

69.   Schatz, M.C., *CloudBurst: highly sensitive read mapping with MapReduce.* Bioinformatics, 2009. **25**(11): p. 1363-1369.

70.   Al-Fares, M., A. Loukissas, and A. Vahdat, *A scalable, commodity data center network architecture.* ACM SIGCOMM Computer Communication Review, 2008. **38**(4): p. 63-74.

71.   Dean, J. and S. Ghemawat, *MapReduce: simplified data processing on large clusters.* Communications of the ACM, 2008. **51**(1): p. 107-113.

72.   Ishibuchi, H. and Y. Nojima, *Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning.* International Journal of Approximate Reasoning, 2007. **44**(1): p. 4-31.

73.   Cheng, J. and P. Baldi, *A machine learning information retrieval approach to protein fold recognition.* Bioinformatics, 2006. **22**(12): p. 1456-1463.

74.   Rosten, E. and T. Drummond, *Machine learning for high-speed corner detection*, in *Computer Vision–ECCV 2006*. 2006, Springer. p. 430-443.

75.   Lu, Z., et al., *Predicting subcellular localization of proteins using machine-learned classifiers.* Bioinformatics, 2004. **20**(4): p. 547-556.

76.   Garfield, E., *Citation indexing for studying science.* Nature, 1970.

77.   Qian, F., et al., *A bibliometric analysis of global research progress on pharmaceutical wastewater treatment during 1994–2013.* Environmental Earth Sciences, 2015. **73**(9): p. 4995-5005.

78.   Coats, A.J., *Ethical authorship and publishing.* International journal of cardiology, 2009. **131**(2): p. 149-150.

79.   Sun, Y., H.-Z. Fu, and Y.-S. Ho, *A bibliometric analysis of global research on genome sequencing from 1991 to 2010.* African Journal of Biotechnology, 2013. **12**(51): p. 7043-7053.

80.   Garfield, E., *The history and meaning of the journal impact factor.* Jama, 2006. **295**(1): p. 90-93.

81.   Eshraghi, A., et al., *100 top-cited scientific papers in limb prosthetics.* Biomed Eng Online, 2013. **12**(1): p. 1-12.

82.   Li, L.-L., et al., *Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006.* Scientometrics, 2009. **80**(1): p. 39-58.

83.   Chiu, W.-T. and Y.-S. Ho, *Bibliometric analysis of tsunami research.* Scientometrics, 2007. **73**(1): p. 3-17.

84.   Liao, J. and Y. Huang, *Global trend in aquatic ecosystem research from 1992 to 2011.* Scientometrics, 2014. **98**(2): p. 1203-1219.

85.   Landset, S., et al., *A survey of open source tools for machine learning with big data in the Hadoop ecosystem.* Journal of Big Data, 2015. **2**(1): p. 24.

86.   Garfield, E., *KEYWORDS PLUS-ISI'S BREAKTHROUGH RETRIEVAL METHOD. 1. EXPANDING YOUR SEARCHING POWER ON CURRENT-CONTENTS ON DISKETTE.* Current Contents, 1990. **32**: p. 5-9.

87.   Dong, B., et al., *A bibliometric analysis of solar power research from 1991 to 2010.* Scientometrics, 2012. **93**(3): p. 1101-1117.