Bibliometric Analysis of Big Data Research

Ramiz Aliguliyev[1], Nigar Ismayilova[2]
[1,2] ANAS Institute of Information Technology, Baku, Azerbaijan
[1]*r.aliguliyev@gmail.com*, [2]*nigar@iit.ab.az*

***Abstract* — The paper analyzes research performance on an area of big data during 2001-2015 years, using data from Scopus. Also was demonstrated geographical distribution of big data research, and predicted future of this area by means of received information.**

Keywords — big data; bibliometric analysis; research performance

I. INTRODUCTION

Big Data is related to technologies for collecting, processing, analyzing and extracting useful knowledge from very large volumes of structured and unstructured data generated by different sources at high speed [1]. The term Big Data is used almost anywhere in modern times; from news articles to professional journals, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005 [2], refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity [3]. This paper is aimed to provide a summary of research activity on big data and characterize its most important aspects.

II. RESULTS OF BIBLIOMETRIC ANALYSIS

Scopus is the largest bibliometric database covering peer-reviewed literature: scientific journals, books and conference proceedings [4]. Scopus was used as a main source in our analysis, the results of which are presented below.

A. Publication output

Big data research has begun in 2001with one published document in Scopus, and 2014 was the most productive year with 3472 documents (45 % of all publications from 2001 to 2015) (figure 1).

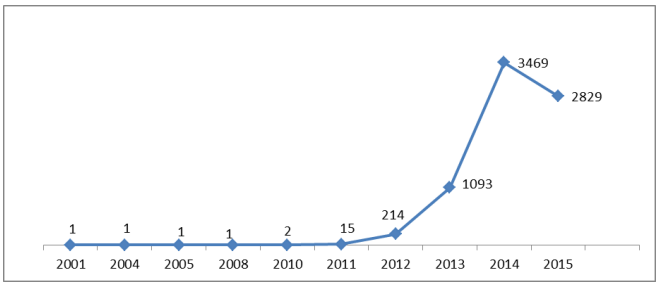

Figure 1. Number of publications in big data.

Exponential and polynomial dependence of publications' number on publication year are C=0.0567exp(1.1011Y) and C= 83.76Y2 -582.95Y+744.02, where Y – is the year and C- is the cumulative number of published documents. According to these equations we can predict that, the number of publications in big data research will be approximately 4467 in 2016.

B. Type of document

The queries result consists of 7633 documents; most of them are conference papers - 4887 (64 %) and journal articles - 2190 documents (28,7 %) . Reviews (204 publications – 2.7%), notes (15 publications – 0.2%), editorial materials (32 publications – 0,4 %) and book chapters (33 documents - 0,5 %) showed much-lesser significance than articles and conference materials. Totally, 7685 articles were published in 98 sources listed in the Scopus. Most of the articles published in journals edited by IEEE, Springer, Elsevier and reported in International Conferences as IEEE Big Data, Data Engineering, Advanced Cloud and Big Data, Symposium on Computational Intelligence in Big Data, Big data Congress, etc. The most productive authors in big data are from Australia and China (Table 1).

TABLE I. TOP TEN AUTHORS ON BIG DATA

N	*Authors*	*Affiliations*	*h-index*	*Number of publications in big data*
1	Ranjan Rajiv	Commonwealth Scientific and Industrial Research Organization, Melbourne, Australia	14	22
2	Wang Lizhe	Chinese Academy of Sciences, Institute of Remote Sensing and Digital Earth, Beijing, China	19	15
3	Cuzzocrea Alfredo	Universita della Calabria, Cosenza, Italy	18	13
4	Liu Chang	Tongji University, Department of Chemistry, Shanghai, China	51	14
5	Chen Jinjun	University of Technology Sydney, Faculty of Engineering and IT, Sydney, Australia	20	14
6	Wang Wei	Xi'an Jiaotong University, School of Electronic and Information Engineering, Xi'an, China	85	15
7	Zhang Xuyun	NICTA, Machine Learning Research Group, Canberra, Australia	6	15
8	Rong Chunming	Universitetet i Stavanger, Centre of Innovation Technology, Stavanger, Norway	10	13

N	*Authors*	*Affiliations*	*h-index*	*Number of publications in big data*
9	Puliafito Antonio	Universita degli Studi di Messina, Department of Civil Engineering, Messina, Italy	21	12
10	Xhafa Fatos	Universitat Politecnica de Catalunya, Department of Languages and Informatics Systems, Barcelona, Spain	21	12

C. Distribution in subject areas

All the documents were published in 26 scientific areas. Most of them (76, 7 %) belongs to Computer Sciences, 23.1 % Engineering and 14.8 % Mathematics. The publications was in Social sciences, Business, Medicine, Physics and other disciplines showed lesser significance (Figure 2).

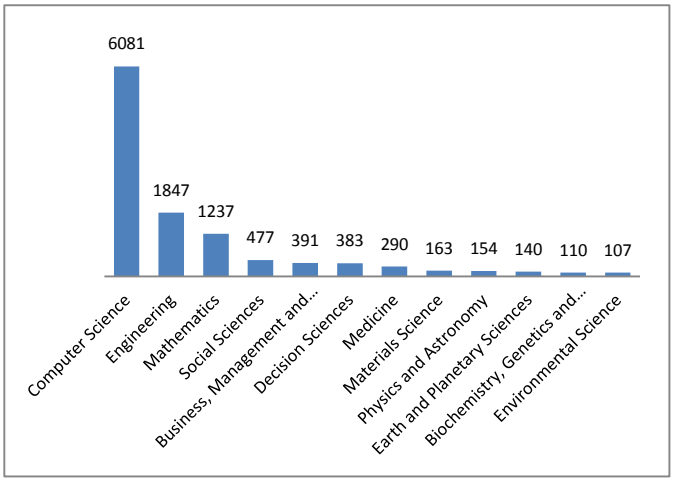

Figure 2. Distribution of publications by subject areas

D. Geographical distribution of publications

The analyzed documents were from 100 countries; researchers from USA have published 2250 documents (29 %), researchers from China 1881 documents (24,4 %). Asian countries as India and South Korea are in top ten of the most productive countries in big data research. The publications are in 13 languages, the documents published in English is 99,7 %, Chinese is 3,3, % and 1% of all publications are written in German, Spanish, French, Japanese, Turkish, Russian, Korean, Poland, Hungarian, Italian and Serbian. The notion of big data was first used by German researcher in 2001, and afterwards by the US and Chinese authors in 2004 and 2005 accordingly. In 2010 were fixed 2 documents from UAE and South Korea. During 2011-2012 years USA was the leader in 'big data' research, in 2013-2014 China have gone after the USA with the insignificant difference and in 2015 have passed it. They are followed by Italy in 2011, UK in 2012 and 2013, Germany in 2014 and India in 2015. In 2015 the Asian countries become more active than the developed European countries (Figure 3, 4, 5, 6).

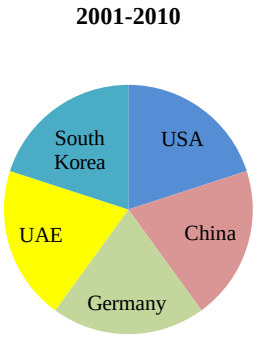

Figure 3. Distribution of publications by countries in 2001-2010 years

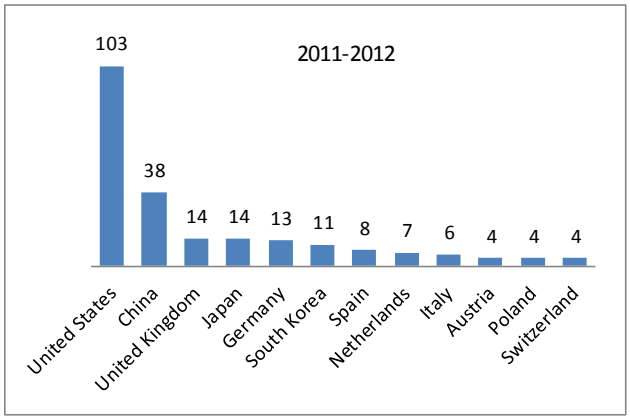

Figure 4. Distribution of publications by countries in 2011- 2012

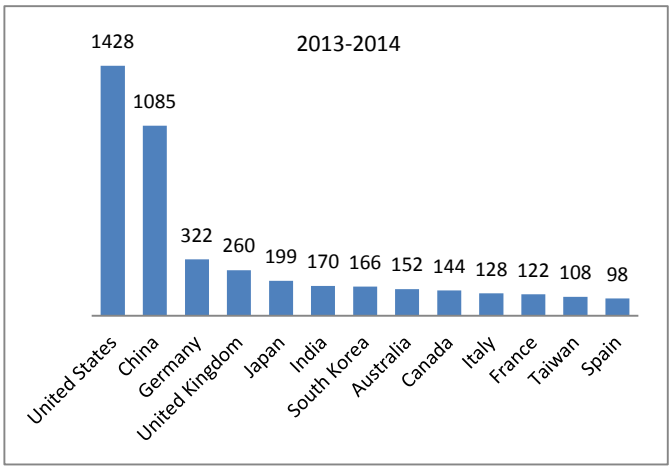

Figure 5. Distribution of publications by countries in 2013-2014

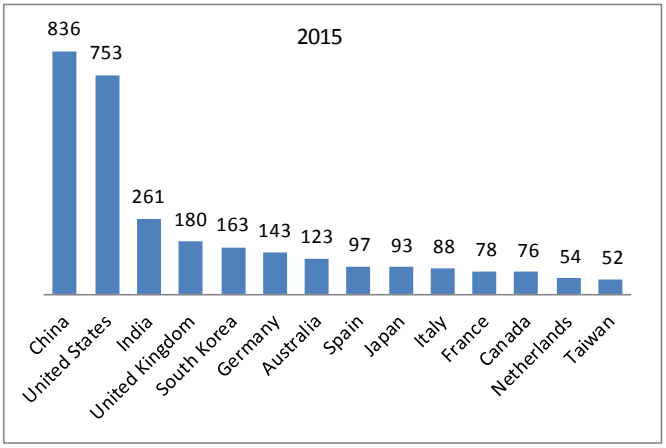

Figure 6. Distribution of publications by countries in 2015

We can conclude that, big data research gets its beginning from Germany, has grown in developed countries as the USA and UK and is prospering in Asian countries as China and India.

E. Topic longevity and the most cited article

Using the citation data we can calculate topic longevity of big data [5]:

$$a = i\sqrt{\frac{k}{k+l}}, \quad (1)$$

where n depends on three parameters, i - the number of years (in this case 6), k - the number of cited papers in topic t published i or more years before the current year, and l - the number of cited papers in topic t published less than i years previously. These parameters determine the aging rate a. The estimated half-life is then

$$\text{Longevity(t)} = -\frac{\log 2}{\log a}. \quad (2)$$

The cumulative number of cited documents during 2008-2014 and 2010 – 2014 years is equal to 1089, according to equation 2 topic longevity is approximately 0.28 [6]. Small longevity score of big data topic indicates that, publications in this topic are relatively recent. The most cited paper by Chen H. et al. was published in 2012 in the journal MIS Quarterly: Management Information Systems named as "Business intelligence and analytics: From big data to big impact" was cited 256 times during 2012-2015 years (Figure 7).

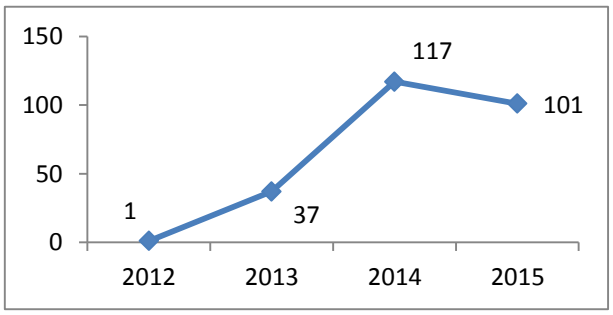

Figure 7. The history of the most cited article

CONCLUSION

Bibliometric analysis carried out in this paper, gives reason for following discourses: 'big data' topic now is progressing and it is directed from Europe and the USA toward Asian countries. Asian scientists have begun actively research and provide their contributions to big data research.

REFERENCES

[1] R. Alguliyev, Y. Imamverdiyev "Big Data: Big Promises For Information Security 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT)", 2014, pp. 1-4.

[2] http://radar.oreilly.com/2010/01/roger-magoulas-on-big-data.html

[3] G. Halevi, H. Moed "The Evolution of Big Data as a Research and Scientific Topic", Research Trends Issue, 2012. pp. 03-04

[4] https://www.elsevier.com/solutions/scopus

[5] R. Əliquliyev, N. İsmayılova "Bibliometriya: müasir vəziyyəti, problemləri və inkişaf perspektivləri," Bakı, 2015, 78 səh.

[6] S. Mann, D. Mimno, A. McCallum " Bibliometric Impact Measures Leveraging Topic Analysis", Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 2006, pp. 1-10.