Research Trends

Volume 1 | Issue 30 Article 2

9-1-2012

The evolution of big data as a research and scientific topic: Overview of the literature

Gali Halevi
Elsevier

Henk F. Moed Dr

Section 1: The Evolution of Big Data as a Research and Scientific Topic

Overview of the Literature.

Gali Halevi, MLS, PhD
Dr. Henk Moed

The term Big Data is used almost anywhere these days; from news articles to professional magazines, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005 (1), 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. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behavior and market behavior. It can also be seen in the life sciences where big sets of data such as genome sequencing, clinical data and patient data are analyzed and used to advance breakthroughs in science in research. Other areas of research where Big Data is of central importance are astronomy, oceanography, and engineering among many others. The leap in computational and storage power enables the collection, storage and analysis of these Big Data sets and companies introducing innovative technological solutions to Big Data analytics are flourishing.

In this article, we explore the term Big Data as it emerged from the peer reviewed literature. As opposed to news items and social media articles, peer reviewed articles offer a glimpse into Big Data as a topic of study and the scientific problems methodologies and solutions that researchers are focusing on in relation to it. The purpose of this article, therefore, is to sketch the emergence of Big Data as a research topic from several points: (1) timeline, (2) geographic output, (3) disciplinary output, (4) types of published papers, and (5) thematic and conceptual development. To accomplish this overview we used Scopus™.

Method

The term Big Data was searched on Scopus using the index and author keywords fields. No variations of the term were used in order to capture only this specific phrase. It should be noted that there are other phrases such as "large datasets" or "big size data" that appear throughout the literature and might refer to the same concept as Big Data. However, the focus of this article was to capture the prevalent Big Data phrase itself and examine the ways in which the research community adapted and embedded it in the mainstream research literature.

The search results were further examined manually in order to determine the complete match between the articles' content and the phrase Big Data. Special attention was given to articles from the 1960s and 1970s which were retrieved using the above fields. After close evaluation of the results set, only 4 older articles were removed from the final results set which left 306 core articles. These core articles were then analyzed using the Scopus analytics tool which enables different aggregated views of the results set based on year, source title, author, affiliation, country, document type and subject area. In addition, a content analysis of the titles and abstracts was performed in order to extract a timeline of themes and concepts within the results set.

Results

The growth of research articles about Big Data from 2008 to the present can be easily explained as the topic gained much attention over the last few years (see Figure 1). It is, however, interesting to take a closer look at older instances where the term was used. For example, the first appearance of term Big Data appears in a 1970 article on atmospheric and oceanic soundings (according to data available in Scopus; see study limitations). The 1970 article discusses the Barbados Oceanographic and Meteorological Experiment (BOMEX) which was conducted in 1969 (2). This was a joint project of seven US departments and agencies with the cooperation of Barbados. A look at the BOMEX site features a photo of a large computer probably used at the time to process the large amounts of data generated by this project (3). Other early occurrences of the term are usually related to computer modeling and software/ hardware development for large data sets in areas such as linguistics, geography and engineering.

When segmenting the timeline and examining the subject areas covered in different timeframes, one can see that the early papers (i.e. until 2000) are led by engineering especially in the areas of computer engineering (neural networks, artificial intelligence, computer simulation, data management, mining and storage) but also in areas such as building materials, electric generators, electrical engineering, telecommunication equipment, cellular telephone systems and electronics. From 2000 onwards, the field is led by computer science followed by engineering and mathematics.

Another interesting finding in terms of document types is that conference papers are most frequent followed by articles (see Figures 2 and 3). As we see in the thematic analysis, these conference papers become visible through the abstracts and titles analysis.

The top subject area in this research field is, not surprisingly, computer science; but one can notice other disciplines that investigate the topic such as engineering, mathematics, business and also social and decision sciences (see Figure 4). Other subject areas that are evident in the results sets but not yet showing significant growth are chemistry, energy, arts and humanities and environmental sciences. In the arts and humanities for example, there is a growing interest in the development of infrastructure for e-science for humanities digital ecosystems (for instance, text mining), or in using census data to improve the allocation of funds from public resources.

Finally, we took a look at the geographical distribution of papers. The USA has published the highest number of papers on Big Data by far, followed by China in second place (see Figure 5). In both countries the research on Big Data is concentrated in the areas of computer science and engineering. However, while in the USA these two areas are followed by biochemistry, genetics and molecular biology, in China computer science and engineering are followed by mathematics, material sciences and physics. This observation coincides with other research findings such as the report on International Comparative Performance of the UK Research Base: 2011 (4) which indicated that the USA is strong in research areas such as medical, health and brain research while China is strong in areas such as computer science, engineering and mathematics.

In addition to the overall characteristics of the publications on Big Data, we also conducted a thematic contextual analysis of the titles and abstracts in order to understand how and in what ways the topics within this field have evolved. In order to accomplish this, the abstracts and titles in each article were collected in two batches; one file containing abstracts and titles of articles from 1999-2005 and the second file from 2006-2012. The analysis concentrated on these years rather than the entire set, as there were multiple publications per year during this period. The texts were then entered into the freely available visualization software Many Eyes (www.manyeyes.net).

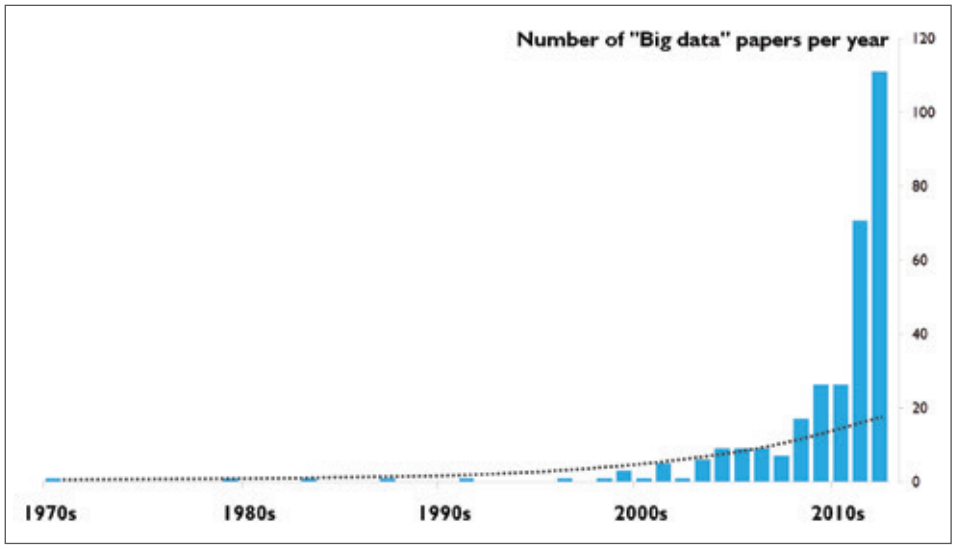

Figure 1: Time line of Big Data as topic of research. The dotted line represents the exponential growth curve best fitting the data represented by the blue bars. This shows the number of Big Data articles increasing faster than the best exponential fit.

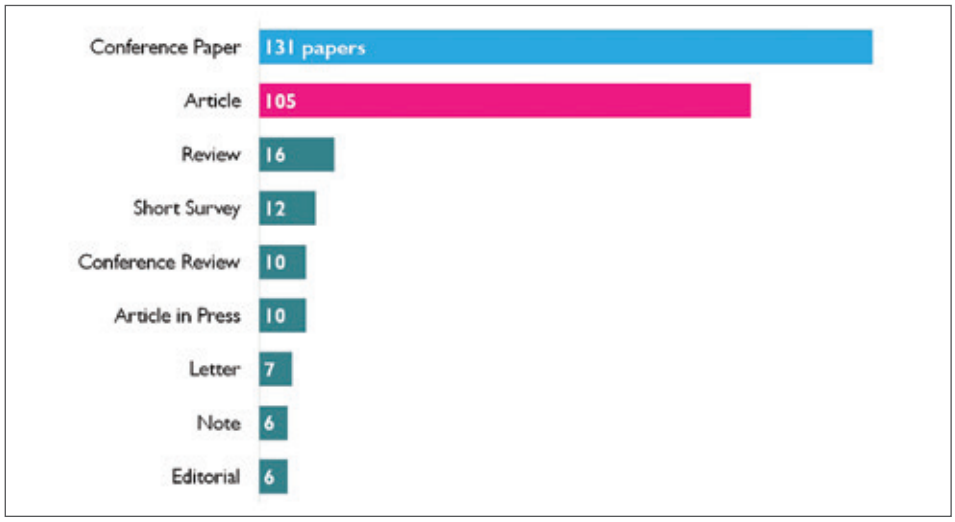

Figure 2: Document types of Big Data papers.

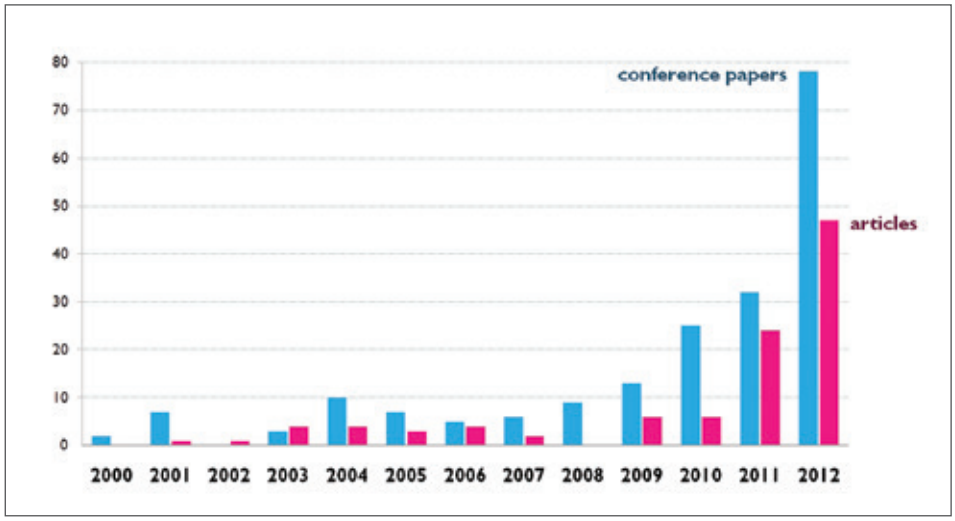

Figure 3: Conference papers and Articles growth over time.

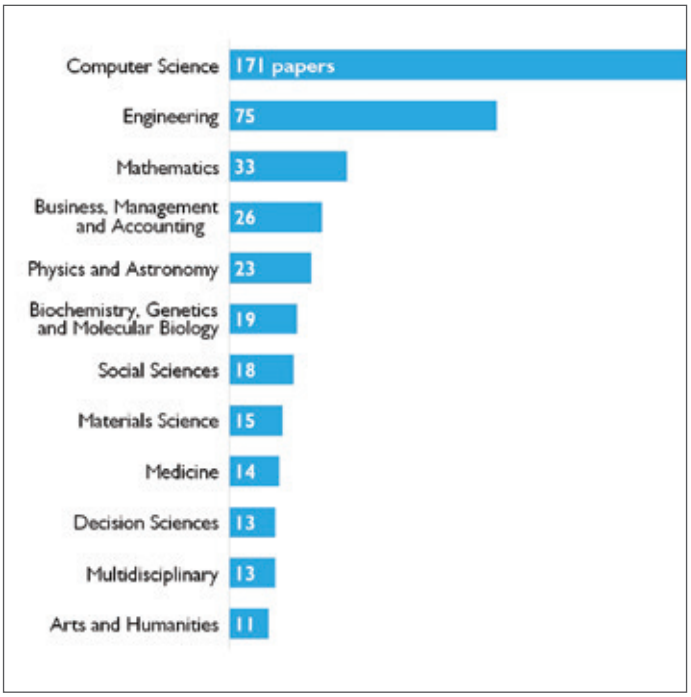

Figure 4: Subject areas researching Big Data.

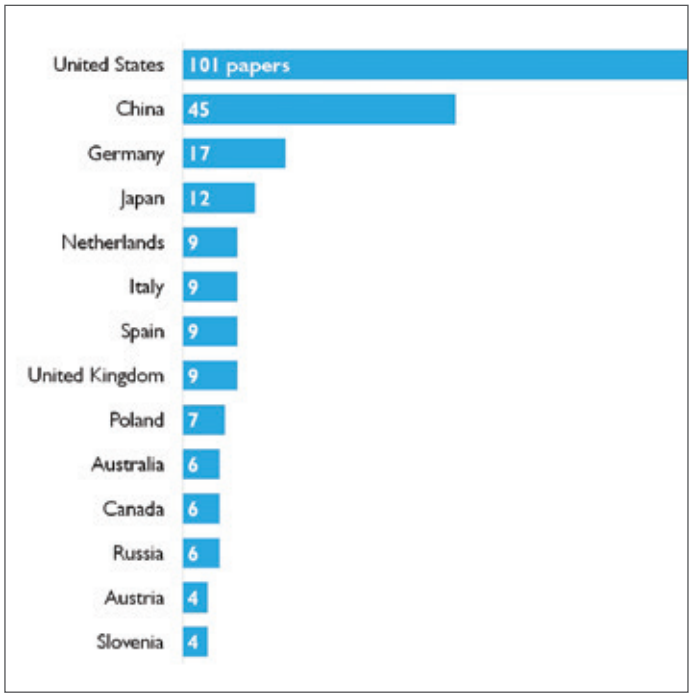

Figure 5: Geographical Distribution of Big Data papers.

This was used to create phrase-maps using the top 50 occurring keywords in these texts. These visualizations were produced by ignoring common words and connecting words such as 'and', 'the', 'of' etc. and used one place space between terms to determine the connections between the terms (see Figures 6 and 7).

These maps visualize two main characteristics of the text: (1) connections between terms are depicted by the gray lines, where a thicker line notes a stronger relationship between the terms; and (2) the centrality of the terms which are depicted by their font size (the bigger the font, the more frequently a term appears in the text). Clusters of connections may appear when a connection is found between single words but not to other clusters.

The first two striking observations when looking at these two maps are the complexity of Figure 6 compared to Figure 7 and the close connectivity of the themes in Figure 7 compared to the scattered nature of their appearance in Figure 6.

The thematic characteristics of the 1999-2005 abstracts and titles text show several scattered connections between two words, seen on the right and left sides of the map. For example, neural networks analysis, on the right side of the map, is a common concept in the field of artificial intelligence computing. This map is conceptually quite simple, with most concepts concentrated around computer related terms, such as 'data mining ',' data sets', XML and applications. When compared to Figure 7 it can be easily seen how the term 'big', although strongly connected to 'data' is not as noticeable as it is in later dates.

The map in Figure 7 represents a tighter network of terms all closely related to one another and to the Big Data concept. Figure 7 also represents a much richer picture of the research on Big Data. There's a clear evolution from basic data mining to specific issues of interest such as data storage and management which lead to cloud and distributed computing. It could be said that the first period demonstrates a naïve picture, in which solutions and topics revolve around a more 'traditional' view of the topic using known concepts of XML and data mining while the second period shows a more complex view of the topic while demonstrating innovative solutions such as cloud computing with emphasis on networks. This also holds for terms such as 'model', 'framework', and 'analytics', that appear in Figure 7, which indicate development and growth in research directions.

A comparison of these two maps also reveals the appearance of diversity in the topics surrounding Big Data such as 'social data' , 'user data' and even specific solutions such as 'MapReduce', a model for processing large datasets implemented by Google (http://mapreduce.meetup.com/), and 'hadoop', an open source software framework that supports data-intensive distributed applications (www.hadoop.apache.org).

As mentioned in the section above analyzing document types, conference papers are central to research in this area. As can be seen in Figure 7, the ACM or IEEE conferences in 2010-2012 play an important role in this area which can be seen by the clear appearance of these terms and their connection to the topic.

Conclusions

Research on Big Data emerged in the 1970s but has seen an explosion of publications since 2008. Although the term is commonly associated with computer science, the data shows that it is applied to many different disciplines including earth, health, engineering, arts and humanities and environmental sciences. Conferences, especially those sponsored by IEEE and/or ACM, are the leaders in the progression of publications in this area followed by journal articles. Geographically, research is led by the USA followed by China and some European countries.

A closer look at the concepts and themes within the abstracts and titles over time show how this area, which began as a computer and technology focus area with some satellite applications, developed into a close and tight-knit discipline featuring applications, methodologies and innovative solutions ranging from could to distributed computing and focusing on user experience. In May 2012, Elsevier sponsored a 2-day conference in Canberra, Australia dedicated to the topics of Big Data, E-science and Science policy (see videos and links to the presentations here: http://www.youtube.com/playlist?list=PL61DD522B24108837). The topic was treated from a variety of viewpoints including the analytics of Big Data sets in publishing, digital scholarship, research assessment and science policy. The multi-dimensional characteristic of this topic is seen in the literature as well as in the social media and online publications. The concept of Big Data as a research topic seems to be growing and it is probable that by the end of 2012 the number of publications will double, if not more, and its analytics and applications will be seen in various disciplines.

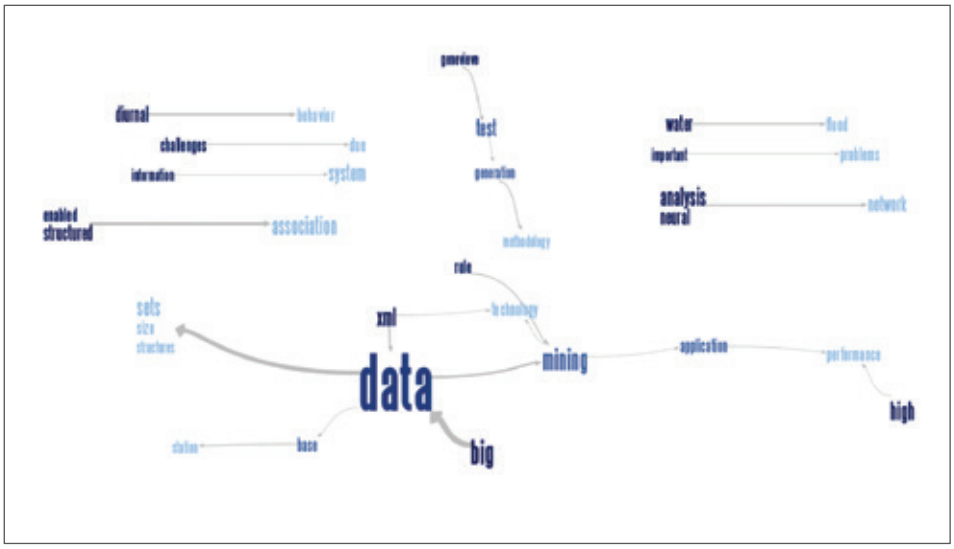

Figure 6: Phrase map of highly occurring keywords 1999-2005.

Figure 7: Phrase map of highly occurring keywords 2006-2012.

Limitations

This study was conducted using Scopus.com in August 2012 and the numbers and percentages presented in this article reflect the indexed publications at the time. These are bound to change as Scopus.com is updated daily with new publications, covering articles in press.

In addition, the dates and document types presented in this study are direct derivatives of Scopus coverage as far as sources and dates. A similar search on other databases might result in slightly different findings and may vary according to the database coverage.

Useful Links:

1. http://strata.oreilly.com/2010/01/roger-magoulas-on-big-data.html
2. http://www.eol.ucar.edu/projects/bomex/
3. http://www.eol.ucar.edu/projects/bomex/images/DataAcquisitionSystem.jpg
4. http://www.bis.gov.uk/assets/biscore/science/docs/i/11-p123-international-comparative-performance-uk-research-base-2011.pdf