Discovering interactions of big data research: A learning-based bibliometric study

Yi Zhang[1], Ying Huang[2], Alan L. Porter[3], Guangquan Zhang[1], Jie Lu[1]

[1]Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[2]School of Management and Economics, Beijing Institute of Technology, Beijing, P. R. China

[3]Technology Policy and Assessment Centre, Georgia Institute of Technology, Atlanta, USA

Abstract: As one of the most representative emerging technologies, big data analytics and related applications rapidly lead the development of information technologies and widely influence the thinking model and management behavior of today's interconnected world. Exploring technological evolution of big data research can be considered as an effective way to enhance the ability of technology management and create value for research and development (R&D) strategies in related government and industry sectors. This paper applies a learning-based bibliometric study to discover the interactions of big data research by detecting and visualizing its evolutionary pathways. Concentrating on a set of 5,840 publications derived from Web of Science, covering the period from 2000 to 2015, text mining and bibliometric techniques are used to profile big data-related hotspots and core scholars, and a learning-based process is used to identify the interactive relationships between topics in sequential time slices, indicating valuable information on technological evolution, fusion, and death. The outputs include a landscape of such interactions in big data research from 2000 to 2015 with detailed routes of specific sub technologies and empirical insights for related studies in science policy, innovation management, and entrepreneurship.

Keywords: Bibliometrics; text mining; scientific evolutionary pathways; big data.

I. INTRODUCTION

It has been several years since the big data boom led the revolution in both management behavior and thinking model in all sectors of modern society [1]. Big data analytics can be defined as "the means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets[1]," and its importance has been highlighted by both academic and business communities, with a broad range of applications in business intelligence [2]. Currently, the capability to explore insights from big data seems to become the only way to succeed for organizations and individuals, and big impact has been raised as one highlighted accomplishment gained from big data [2, 3]. However, things are always easier said than done, and one investigation conducted by Mckinsey Global Institute indicates that despite those successful examples in Amazon and Google the success of big data analytics in legacy companies is still limited [4]. At this stage, it is significant to trace the evolution of big data research in the past years, discover the interactions between related techniques of big data analytics, and identify crucial technological connections that hold the capability to create and enlarge such "big impact."

Oriented to the 5840 publications derived from Web of Science, a learning-based bibliometric study is conducted to address the above concern. On one hand, a model of research and development (R&D) profile is used to review the landscape of big data research by 1) profiling the statistical dynamics and geographic distribution of related scientific publications and 2) identifying core players (i.e., leading journals, organizations, and countries) and their competitive and collaborative relationships in this area. On the

[1] The definition is given in "Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)," Program Solicitation NSF 12-499. More information can be addressed on the website: https://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

other hand, we apply a model of scientific evolutionary pathways (SEP) to detect and visualize the technological change of big data research from 2000 to 2015, in which a learning process is used to obtain the dynamics of semantic concepts in sequential time slices. In addition, aiming to retain the benefit of the objective analysis and at the same time enhance the reliability of our recommendations, expert knowledge is engaged.

The rest of this paper is organized as follows: we describe the data and propose the analytic framework of our study in Section II. Section III presents the results of the R&D profile and the SEP, and provides recommendations on science policy and entrepreneurship, with the help of expert knowledge. We further conclude our research, limitations, and future study in Section IV.

II. DATA AND METHODOLOGY

A. Data and pre-processing

As discussed in [5], scientific publications can be a good resource for exploring information on the research frontier of a given technological area, and Web of Science (WoS) is approved to be a quality-guaranteed database under this circumstance. We collected 5840 publications from WoS by using an updated version (given in Table I) of the search strategy proposed in [6].

TABLE I. SEARCH STRATEGY

NO	Strategy
#1	TS= ("Big Data" or Bigdata or "Map Reduce" or MapReduce or Hadoop or Hbase or Nosql or Newsql)
#2	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")
#3	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")
#4	#1 OR (#2 AND #3)

We emphasized the combined field of titles and abstracts and used a function of natural language processing (NLP) in VantagePoint[2] to retrieve terms. A term clumping process [7] was applied to remove noise and consolidate technological synonyms. The stepwise results of the term clumping process are given in Table II, and the 10,921 terms are considered as the core technological terms in the field of big data research.

TABLE II. STEPWISE RESULTS OF THE TERM CLUMPING PROCESSING

Step	Description	#Term
0	Raw terms retrieved by the NLP technique;	120,427
1	Removing terms starting with non-alphabetic characters, e.g., 1.5%;	115,381
2	Removing meaningless and common terms, e.g., pronouns, prepositions, and conjunctions;	110,362
3	Removing common terms in scientific publications, e.g., “method” and “introduction;”	109,137
4	Consolidating meaningless terms in the field of computer science, e.g., “classification” and “classification analysis;”	108,344
5	Consolidating terms with the same stem, e.g., the singular and plural of a noun;	91,918
6	Removing single words [a], e.g., “internet” and “information;”	84,949
7	Removing terms appearing in only one publication;	12,229
8	Consolidating technological synonyms with expert knowledge [b], e.g., “time series” was used to represent terms such as “time series forecasting,” “time series	10,921

[2] VantagePoint is commercial software in text mining and in particular in Science, Technology, and Innovation (ST&I) text analysis. More detail can be addressed on the website: https://www.thevantagepoint.com/

	mining" and "time series economics"	

[a.] Since we would consolidate terms with their related single words (e.g., the case of "classification") in Step 5, these single words that have ever consolidated with terms will not be removed.

[b.] Yi Zhang and Ying Huang reviewed the terms derived by Step 7, and based on the list of big data techniques and technologies outlined in [8], related technological synonyms were consolidated.

B. Methodology

The methodology of this study includes a model of performance analysis and a model of scientific evolutionary pathways, and the analytic framework is given in Fig. 1.

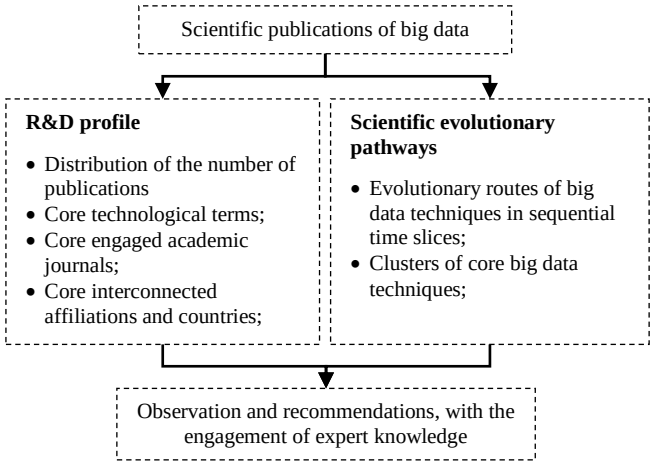

Fig. 1. Analytic framework

The model of R&D profile focuses on the basic statistical information of the collected scientific publications, including the number of publications, term frequency, and correlated journals, affiliations, and countries. In addition, based on the core terms derived by the term clumping process, a series of association analysis to investigate the research similarity between affiliations and between counties will be conducted.

The model of scientific evolutionary pathways (SEP) is mainly based on the algorithm proposed in [9], and considering the preference of fewer number of topics, we use the K-means algorithm refined in [10] to take the place of the hierarchical agglomerative clustering approach, and the configuration of related parameters also depends on the actual situation of our empirical dataset. The main concept of the SEP approach can be described as follows:

Definition 1: a topic is a collection of publications and can be mathematically represented by its centroid, which is identified as the publication sharing the highest similarity with all other publications in the topic.

Definition 2: a topic can be geometrically represented as a circle, and its boundary is the largest Euclidean distance between its centroid and all publications.

The stepwise algorithm of the SEP approach is given below:

Step 1: To simulate the dataset as a data stream, consisting of certain sequential time slices, and the K-means algorithm is used to group the publications in Time Slice 0 into several initial topics;

Step 2: To process the data stream in an iterative flow, i.e., one time slice will be treated as one iteration and publications in a time slice will be read one by one;

Step 3: To measure the similarity between a forthcoming publication and the centroids of all existing topics by using Salton's cosine [11], and assign the publication to the most similar topic;

Step 4: To calculate the Euclidean distance between the publication and the centroid of its assigned topic. If the distance is within the boundary, we set the publication as "normal;" if it is nearby the boundary within a given interval, we set the publication as "evolution;" if it is much larger than the boundary, we set it as "novelty/noise" and its assignment with the existing topic will be deleted;

Step 5: At the end of each iteration, 1) to group publications labeled with "evolution" and "novelty/noise" respectively by using the K-means algorithm. New-born topics consisting of publications labeled with "evolution" will be set as the generation of their assigned topic in Step 3, while new-born topics consisting of publications labeled with "novelty/noise" will not have any predecessor; 2) to detect the accumulation of the number of publications in each topic, and set a topic as "death" if the accumulation is 0 in certain sequential time slices; 3) to measure the similarity

between all new-born and all topics (including both existing and dead ones). If a new-born topic shares the highest similarity with a topic which is not its predecessor, we will combine the new-born topic to the old one, and the linkage with its predecessor will be removed. If the old topic is dead, it will be resurged.

Step 6: To return to Step 2 until the stream ends.

We visualize the topics and their relationships identified by the SEP approach via Gephi [12], and a landscape of core technological clusters in big data research and their detailed evolutionary routes can be addressed in a clear way. Expert knowledge is engaged to provide in-depth recommendations.

III. Results: How Big Data Interconnects the World

The results of our study include the R&D profile and the scientific evolutionary pathways of big data research, and we also provide recommendations on related studies in science policy and entrepreneurship.

A. R&D profile

The distribution of the number of publications in big data research is given in Fig. 2. Despite common sense that the big data boom started in the late 2000s when a number of world-leading IT companies developed architectures to handle large-scale data (e.g., MapReduce by Google in 2004 [13]), "big data" is still a new term to the public, and so to academia. This circumstance can be endorsed by the few and relatively unchanged number of publications from 2000 to 2010 in Fig. 2. The dramatic increase in the number of scientific publications after 2012 can be credited to the "*Big Data Research and Development Initiative*[3]," which was announced by the Obama administration and formally raised the significance of big data research to the stage of national strategy. Funding provided by the governments of the United States, the European Union, China, and many other countries effectively stimulated big data research.

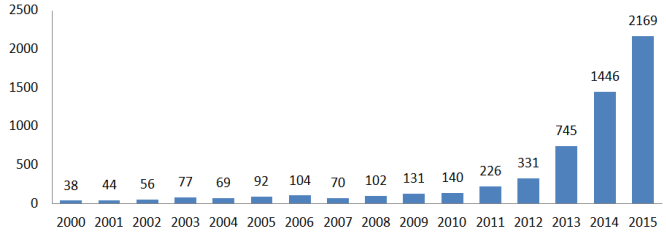

Fig. 2. Distribution of the number of publications in big data research

The main components of big data [8] summarized by Mckinsey Global Institute in 2011 (Mckinsey List) were widely recognized by both industry and academia. It has been more than five years since it was first released, and it becomes interesting to explore the answer of the questions such as "What happened in the past five years?" and "How the importance of different big data techniques is now?" At this stage, we selected the top 70 core technological terms identified by Step 8 of Table II and visualized them in Fig. 3 via a function of Word Cloud in VantagePoint, in which term frequency is used to decide the size of these terms.

Fig. 3. Core technological terms in big data research

Despite the fact that "MapReduce" and "Hadoop" are still two hot terms, the importance of "machine learning," "cloud computing," and "optimization" is highlighted in Fig. 3, and other techniques such as "prediction model," "internet of things," and "classification" follow the trend. Generally, there is no "unexpected" term, and all terms in Fig. 3 or their synonyms can be tracked

[3] Information of the big data initiative can be addressed on the website: https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

back to the Mckinsey List. However, it is clear that the importance and the internal content of some techniques have changed. Three examples are given as follows:

1) The importance is weakened - terms relating to "A/B testing" cannot be found in our list. Apparently, A/B testing is a basic approach of statistical hypothesis testing, and scientific publications might not hold strong interest to such a mature and basic technique, compared with those novel and intelligent algorithms;

2) The internal content is extended - the term "artificial intelligence" has a relatively low frequency, which is out of our imagination, but it is clear that in our list there are a large number of terms closely relate to artificial intelligence, e.g., machine learning, neural network, and natural language processing. In other words, many sub-domains of artificial intelligence have evolved into relatively-mature research areas, which in some sense negatively influence the frequency of "artificial intelligence." A similar case occurs among the terms "natural language processing," "sentiment analysis," and "emotion recognition."

3) The internal content is changed – actually the Mckinsey List did not consider "internet of things (IoT)" as one big data technique, but, undoubtedly, the interaction between big data and IoT is broader and deeper than that of we imagined years ago. Technically, before the 2010s, IoT closely related to radio-frequency identification (RFID) techniques, but now, involved new techniques include sensor networks, Wi-Fi techniques, and a wide range of smart and mobile devices. From the perspective of national R&D strategy, China identified IoT as one of its top 5 emerging industries in 2009 [14]. Comparably, in 2014 the Obama administration highlighted IoT in a big data report "*Big Data: Seizing Opportunities Preserving Values*[4]," and the concern on data privacy in big data age was first raised officially.

Fig. 3 provides an overview to help answer the question of what happed in big data research, and the following part in this section will focus on the core players of big data research, offering insights on the questions such as "Which journals are holding interest to the frontier of big data research?" "Which countries are leading the global competition in big data?" and "Which organizations are leading the world, and how they interact with each other?"

Based on the 5840 publications we collected from WoS, 1759 journals were retrieved, and the top 20 journals with the largest number of publications are listed in Table III. It is interesting that two bioinformatics-related journals are in the list, which might reflect the increasing interest of analyzing large-scale dataset in biological areas and particularly genomic data. Two multidisciplinary journals also attract our eyes. On one hand, the special issue of *Nature* entitled *Big Data* in 2008[5] and the special issue of *Science* entitled *Dealing with Data* in 2011[6] can be considered as the milestone of big data research, indicating the start of the big data boom in academia. Under this circumstance, the appearance of *Nature* in Table III demonstrates the continuous interest of world-leading research communities to big data research. On the other hand, the fact that big data research was published in journals in the field of multidisciplinary study might be able to endorse the argument that big data belongs to the scope of emerging technologies and it can be the common interest of researchers in the areas of both natural science and social science.

TABLE III. TOP 20 JOURNALS IN BIG DATA RESEARCH

# P. [a]	**Journal**	**# P.**	**Journal**
84	Futur. Gener. Comp. Syst.	46	Computer
79	PLoS One	44	Bioinformatics
69	Concurr. Comput. Pract. Exp.	41	Inf. Sci.
62	Big Data	36	J. Parallel Distrib. Comput.
57	IEEE Trans. Parallel Distrib. Syst.	36	Nature
56	BMC Bioinformatics	36	Neurocomputing
56	IEEE Trans. Knowl. Data Eng.	35	ACM Sigplan Not.
51	J. Supercomput.	34	Expert Syst. Appl.
49	Int. J. Distrib. Sens. Netw.	34	IBM J. Res. Dev.
48	Cluster Comput.	33	Commun. ACM

a. The number of publications

[4] https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[5] http://www.nature.com/news/specials/bigdata/index.html

[6] http://science.sciencemag.org/content/331/6018

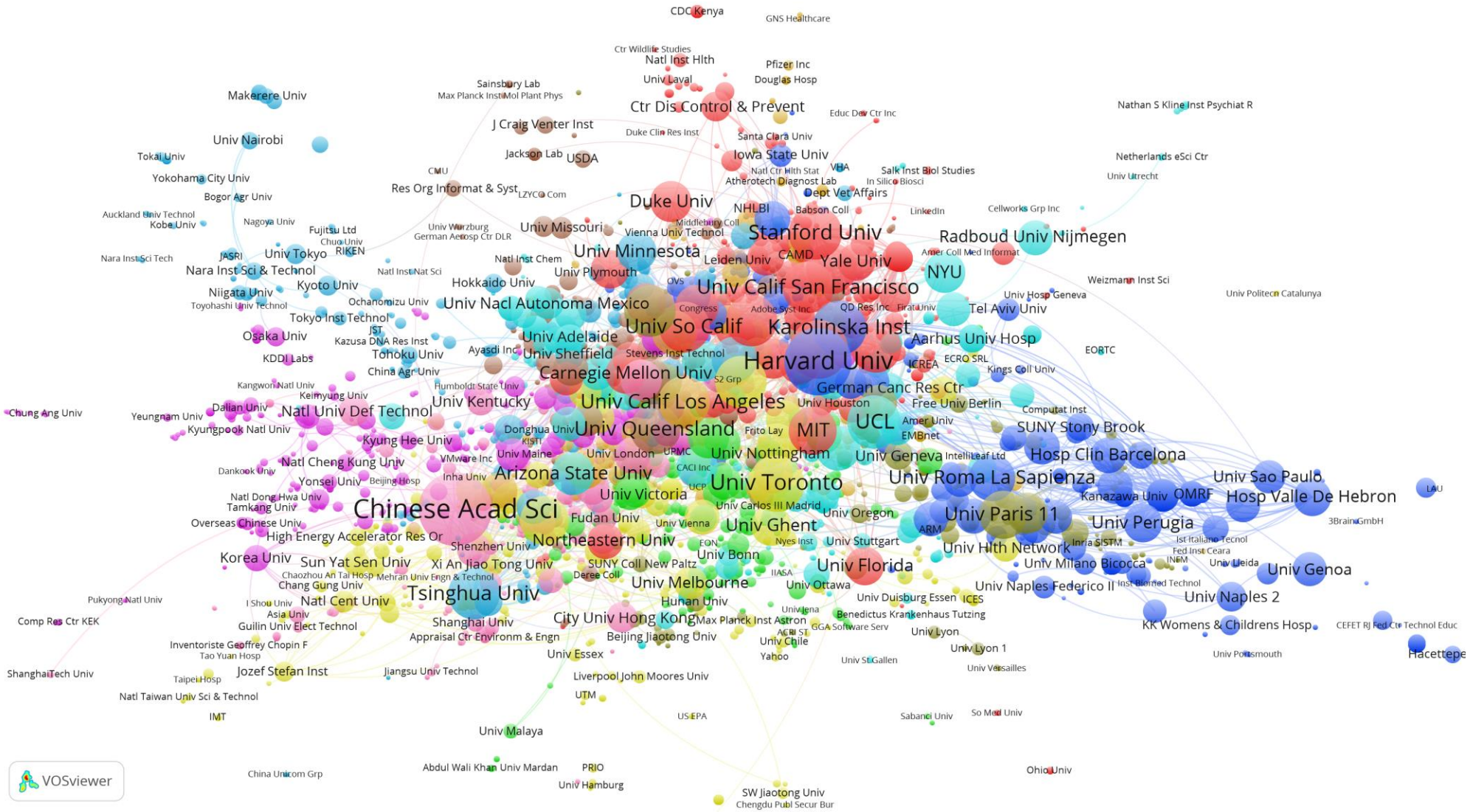

Fig. 4. Co-authorship map between affiliations in big data research

B. Scientific evolutionary pathways

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as PICMET, INFORMS, IEEE, etc. do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

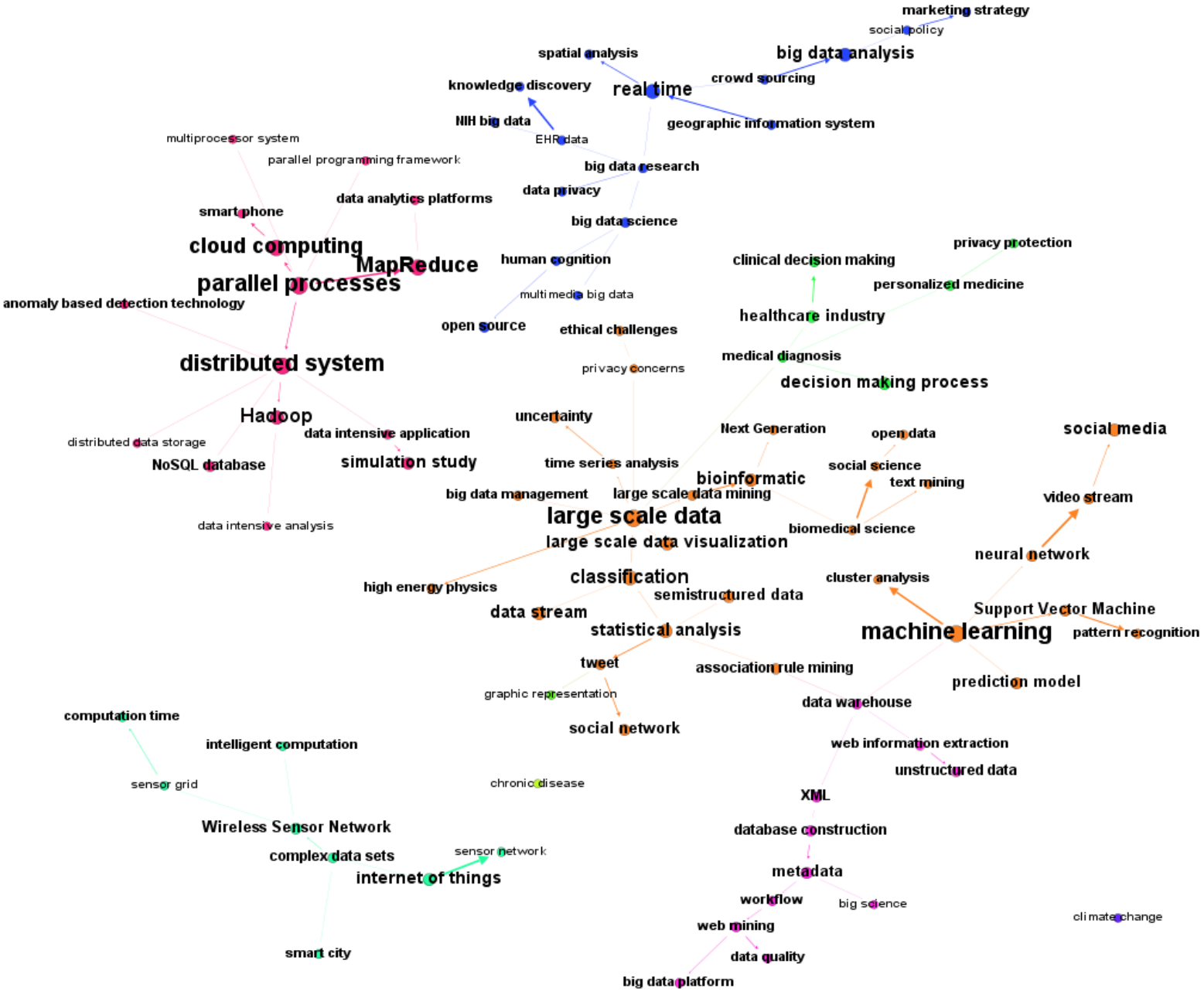

Fig. 5. Scientific evolutionary pathways of big data research

C. Recommendations

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as PICMET, INFORMS, IEEE, etc. do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

IV. DISCUSSIONS AND CONCLUSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the PICMET submission ID for the name of your paper, but also keep the original template as reference for the styles to be used for various parts of your document. Copy the title of your paper in the text file and paste on the new file with the PICMET submission ID. Highlight it. Use the pulled down window on the MS Word Styles toolbar (See figure 1 below). Select the *paper title* style (See figure 1 below).

Copy the authors' names and affiliations in the text file and paste on the new file. Highlight the authors' names. Select the *Author* style from the MS Word Styles menu. Then highlight the affiliations and select the *Affiliation* style from the MS Word Styles menu.

Continue with copying the rest of the contents in the text file and pasting on the template, highlighting part by part, and selecting the appropriate style from the MS Word Styles menu.

A. Key findings

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

B. Limitations and future study

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

This work is partially supported by the Australian Research Council under Discovery Grant DP150101645.

REFERENCES

[1] V. Mayer-Schönberger and K. Cukier, *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt, 2013.

[2] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Quarterly*, vol. 36, pp. 1165-1188, 2012.

[3] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234-246, 2015.

[4] D. Court, "Getting big impact from big data," *McKinsey Quarterly*, January, 2015.

[5] Y. Zhang, H. Chen, and D. Zhu, "Semi-automatic technology roadmapping composing method for multiple science, technology, and innovation data incorporation. ," in *Anticipating Future Innovation Pathways through Large Data Analytics*, T. Daim, A. L. Porter, and D. Chiavetta, Eds., ed New York: Springer, 2015.

[6] Y. Huang, J. Schuehle, A. L. Porter, and J. Youtie, "A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for 'Big Data'," *Scientometrics*, vol. 105, pp. 2005-2022, 2015.

[7] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, ""Term clumping" for technical intelligence: A case study on dye-sensitized solar cells," *Technological Forecasting and Social Change*, vol. 85, pp. 26-39, 2014.

[8] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, *et al.*, "Big Data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute2011.

[9] Y. Zhang, G. Zhang, D. Zhu, and J. Lu, "Science evolutionary pathways: Identifying and visualizing relationships for scientific topics," *The Journal of the Association for Information Science and Technology,* accepted, 2016.
[10] Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, and J. Lu, "Topic analysis and forecasting for science, technology and innovation: Methodology and a case study focusing on big data research," *Technological Forecasting and Social Change,* vol. 105, pp. 179-191, 2016.
[11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management,* vol. 24, pp. 513-523, 1988.
[12] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," *ICWSM,* vol. 8, pp. 361-362, 2009.
[13] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM,* vol. 51, pp. 107-113, 2008.
[14] J. Wen, "Let Science and Technology Lead China's Sustainable Development," 2009.