http://dx.doi.org/10.14400/JDC.2015.13.9.129

Knowledge Creation Structure of Big Data Research Domain

Su-Hyeon Namn

Department of Global IT Business, Hannam University

빅데이터 연구영역의 지식창출 구조

남수현

한남대학교 글로벌IT경영전공

Abstract We investigate the underlying structure of big data research domain, which is diversified and complicated using bottom-up approach. For that purpose, we derive a set of articles by searching "big data" through the Korea Citation Index System provided by National Research Foundation of Korea. With some preprocessing on the author-provided keywords, we analyze bibliometric data such as author-provided keywords, publication year, author, and journal characteristics. From the analysis, we both identify major sub-domains of big data research area and discover the hidden issues which made big data complex. Major keywords identified include SOCIAL NETWORK ANALYSIS, HADOOP, MAPREDUCE, PERSONAL INFORMATION POLICY/ PROTECTION/PRIVATE INFORMATION, CLOUD COMPUTING, VISUALIZATION, and DATA MINING. We finally suggest missing research themes to make big data a sustainable management innovation and convergence medium.

Key Words : Big Data, Convergence, Bottom-up Approach, Keyword Analysis, Bibliometric Data, Research Domain

요 약 본 논문은 학제간 연구의 대표적인 사례인 빅데이터 연구가 어떤 주제로 구성되어 있는지를 상향식 접근법을 이용하여 분석한다. 분석을 위해서 연구재단에서 제공하는 학술지 인용색인시스템을 이용하였다. 영문 키워드 "big data"로 모든 등재지와 등재후보지를 대상으로 검색을 하여 이것을 원천 데이터로 하였다. 논문 저자가 직접 제공하는 키워드를 본 연구에서 사용하기 위해서 정제작업을 거친 후, 주요 키워드 분포, 참여 저널의 성격 분포, 참여 저자 수의 분포, 연도별 키워드 분포 등을 이용하여 빅데이터 연구주제의 구조를 설명하였다. 식별된 주요 키워드들은 사회네트워크 분석, 하둡, 맵리듀스, 개인정보/보호, 클라우드 컴퓨팅, 시각화, 데이터마이닝 등이다. 또한 빅데이터가 지속가능하고 융복합적인 경영혁신 도구로 사용되기 위해 향후 추가적으로 보완되어야 할 연구 키워드들을 제안한다.

주제어 : 빅데이터, 융복합, 상향식 접근, 키워드 분석, 서지정보, 연구 영역

1. Introduction

McAfee et al. [8] anticipated that big data analytics will become the core competency of an organization, not just augmenting trend reports.

* 본 논문은 2015년 한남대학교의 교비연구비에 의해 지원되었음.
Received 12 July 2015, Revised 15 August 2015
Accepted 20 September 2015
Corresponding Author: Su-Hyeon Namn
(Department of Global IT Business, Hannam University)
Email: namn@hnu.kr

ISSN: 1738-1916

Due to the Google Trend service we got familiarized to the application and power of big data analysis. We notice that in general 3Vs are used to describe big data: volume, velocity and variety. Sometimes the fourth dimension of "complexity" is added to them. These dimensions indicate that big data is a complicated domain. In this paper we are especially interested in the third and fourth dimensions, diversity and complexity; which are the major constituents of big data research domain.

To do that, we investigate the academic journals which are related with big data and analyze bibliographic data, especially keywords, authors, characteristics of contributing journals, and publication year. In this way we hope to understand the underlying structure of the complex domain of big data.

We also propose a new integrative framework where the sub domains of research on big data can be placed. In this way, we may recognize which part of the framework is actively researched or not.

We adopt a bottom-up approach for this research without any assumptions. The only assumption is that once we collect all the pieces of information provided by the bibliometrics, we may be able to have a whole picture of the big data research domain. For our empirical work, we used the Korean Citation Index (KCI) service to derive a set of articles by searching keyword "big data". Based on the biblometric data analysis, we both propose major sub-domains of research issues. and suggest the future research direction to fill the research gap.

In section 2, we provide literature review and background information regarding big data and bibliometric analysis. Preparation of data for the empirical work is given in Section 3. We analyze the data to understand the structure of the sub-domains of big data research in Section 4, followed by conclusion and limitation in Section 5.

2. Background

2.1 Perspectives of Big Data

Since big data can be viewed from diverse perspectives, it is not easy to define in concise terms (for a detailed survey see [2]).

For example, a dichotomous division from hardware platform for storage, server, and network to software tools for management and analysis [15]. Also it can be viewed from the process view, input (data) - process - output (solution). We can also approach it in terms of management process, planning, doing, and controlling.

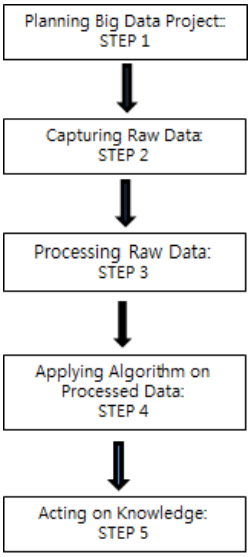
In addition, it can span from technical infrastructure to final application stage. It can be understood from different levels of algorithmic and visualization point of view for the data analysis. Moreover, by looking at the real applications of big data, we can indirectly understand the complexity of the domain. For example [14] provides a set of extensive cases across the industries. The cases were clarified in terms of application area, data employed, and analytic techniques utilized. Even though this approach can be practically useful, it is difficult to identify the structure and trend of big data research since we do not have specific target population to be considered.

In order to place each sub-domain of big data research we need a framework. In this way we can categorize and map the keywords identified in that framework. To achieve this objective, we propose a process model, which captures the transformation process from raw input data to knowledge customized to users as in [Fig. 1].

The process model starts from the planning stage in STEP 1 to define what we want to do with big data analysis. The next stage is to capture uncleaned raw data from the existing DB, sensors, etc. Once we capture the raw data, we need to preprocess to transform them into the usable form in STEP 3. In STEP 4, we can apply diverse analytical algorithms or

visualization techniques to the data, extracting desirable knowledge for the users. Finally in STEP 5 the users act on the knowledge.

Note that the value accrued becomes bigger at later stage of the model. We use the process model to find the distribution of keywords along the stages so that we know the intensity of research activity for each stage.



[Fig. 1] Value-Adding Process Model of Big Data

2.2 Bibliometric Analysis

Academic research articles are crucial for producing and sharing the core knowledge among the scholars and practitioners. Instead of analyzing the unstructured main body of an article, bibliographic data such as titles, abstracts, keywords, participating authors, and citations provide valuable information since they convey the essential aspects of the articles.

Bibliometric data can be used to identify the underlying structures of a research domain, especially when a research domain is interdisciplinary like Management Information Systems, Technology and Innovation, or Big Data.

Keywords are selected and provided by the authors to cover the most important concepts of the articles. That is why these days editorial board of academic journals requires the authors provide keywords. Even though the keywords are non-metric data, but they are more structured compared with main text. They have compressed meaning. Moreover, since keywords are easily accessed through the journal retrieval system, it is useful to analyze a large set of articles to look into the big picture of the research area. Thus keyword related analysis is an important branch of bibliographic study.

Keyword analysis has been extensively used to identify the trend of an academic discipline. For example, Choi et al. [3] used top five MIS journals to examine how the keywords were interrelated with each other over time. They identified keyword network measures such as frequency, degree, and betweenness as well as important keywords.

Chen et al. [1] searched academic journals during the period of 2000-2011. They used three keywords such as “business intelligence”, “business analytics”, and “big data”. They mentioned the limitation of bibliometric analysis since different keywords such as “business analytics” and “data mining” can be used interchangeably by some authors. The total number of articles containing these terms were 3,602, divided into 3,146, 213, and 243 for “business intelligence”, “business analytics”, and “big data”. It is notable in the western world, “business intelligence” is the most popular term and “big data” is in low popularity. However, they did not analyzed the interrelationship among the keywords.

Lee et al. [6] attempted to find out the knowledge production structure in Law. They analyzed 10 year period to derive important issues in Law, using descriptive statistics, keywork analysis, and network analysis techniques. They found that “privacy” is one of the most important keywords in terms of degree centrality.

In technology innovation sector, [10] performed co-author network analysis to find out the cooperative research network in innovation studies of Korea. [13]

also investigated the coauthorship network to analyze the sub fields of information systems research. Especially they considered the coauthorship network as knowledge sharing and social capital to make the research deeper. [11] analyzed author-provided keywords appeared in technology innovation journals of Korea. They used the cluster analysis to test the characteristics of keywords used across the three representative journals in the domain.

3. Data

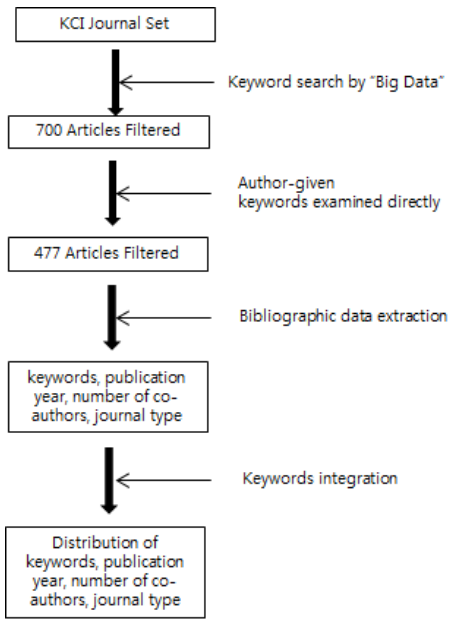
We used the KCI system for our empirical work.. We only used the keyword "big data" to search the KCI database for two reasons. First, big data conveys an extensive coverage and meaning. Second, we waned to make sure the samples to be consistent and relevant. The single term was just used to link other keywords so that we encompass related subjects under "big data" research domain. We used English term for the searching categorizing the keywords because the origin of the most of terminologies in big data is in English and there is no consensus for translating them into Korean terms.

Using the search term, we identified 700 papers, which we call it the original set. (KCI system accessed in August, 2015) If the term is included in the article title or in the author-provided keywords, then the article is counted in the original data set. We found many articles which contain keywords in Korean, but not in English. In that case, We translated the Korean keywords into English keywords.

From the original set, we first filtered out 223 articles 1) if the article does not provide keywords in neither Korean nor English, or 2) even if the article included keywords, but if keyword "big data" is not found in the keyword list. After the screening process, we were left with 477 articles, or 68% of the original set. This reduced set is to be considered in our analysis.

From the reduced set of articles, we sorted out 2,150 keywords. The average number of keywords per article is 4.42. After duplicated keywords are exluded, we are left with 1,296 unique keywords. Note that since we searched the KCI DB based on the keyword of "big data", all the articles considered in this paper include big data in the keyword. Thus we assume that the set of 1,296 keywords describes the intellectual ecosystem of big data research area.

In section 4, we analyze the bibliometric data included in the reduced set of 477 articles.



[Fig. 2] Steps of data processing

The bibliometric data include the author provided keywords, publication year, characteristics of journals, and the number of authors. The data preparation process is described in [Fig. 2].

4. Analysis

In this section, we analyze the bibliometric data from

the 477 papers to delineate the diverse nature of big data research domain.

4.1 Distribution of articles and authors

We first look into the distribution of articles and participating authors. Considering the diversity and complexity dimensions of big data characteristics, noted in Section 1, we build the following conjectures:

Conjecture 1: We expect the number of articles will increase as time goes by, since opportunities from big data are expected big [8].

Conjecture 2: The average number of co-authors per article will increase as time goes by, since the research on big data would pursue more complex problems and thus more co-authors will be involved in the research.

To test the conjectures, we use <Table 1>.

〈Table 1〉 Distribution of Articles and Authors

Categories	2015[(1)]	2014	2013	2012	2011	Total
Number of articles (A)	111	196	121	47	2	477
Number of Participating Authors (B)	230	442	292	122	2	1,088
Average number of author (B/A)	2.07	2.25	2.41	2.60	2	2.28

(1) Includes the first half of 2015.

It shows that “big data” keyword appears in 2011 for the first time. From 2012 to 2013 and from 2013 to 2014 we see a rapid increase in the number of articles published. Since the number of articles in 2015 reflects only the first half of the year, the increasing trend seems to hold, even though the rate is slower than those in the previous years.

The distribution indicates that the Conjecture 1 holds at least for the 5 year-period.

In terms of the number of co-authors, the average number of authors in fact decreases as time goes by. From 2012 to 2015, the average number of co-authors decreased from 2.6 to 2.07, resulting in 20% decrease. The result is just the opposite of Conjecture 2.

We did not statistically test the significance of the yearly difference in terms of the average number of authors per article, but the difference looks quite significant.

The reasons might be the followings:

1) In Korea, most of the universities, the number of authors is important when research performance is measured. The contribution of an author diminishes as the number of co-authors increases. In recent years, researchers might be pursuing high evaluation marks, rather than doing deep research.

2) The big data issues became more structured and clear in terms of what to do as time goes by.

This means that big data research is in maturity stage. But it is not true because even rapid processing of high volume and diverse data types is not realized technically or algorithmically yet. Moreover big data is not embedded in organizations as a management innovation media yet [8]. In this sense, we might explain the unexpected phenomenon as in the following: shallow big data research has been prevalent, even though superficial big data is accepted by many organizations and researchers.

4.2 Distribution of journal characteristics

To recognize the popularity of big data research from diverse disciplines, we summarize the characteristics of academic journals contributing to big data domain research in <Table 2>.

As expected, “Computer Science” related journals contributed the most, followed by “Management” and “Law”. The contribution by “Law” is remarkable. Based on the keyword analysis, issues such as personal information protection, privacy, the right to be forgotten, and criminal investigation triggered active research in “Law” discipline. The domain of “Design”

with 12 articles is an emerging discipline, dealing with visualization in 3D as well as 2D renderings. We expect Design discipline becomes more important since visualization during decision making processes is crucial for creative design of alternative solutions. Mintzberg & Westly [9] coined "seeing first", compared with Simon's rational "thinking first" approach. We also notice from <Table 2> that other disciplines publishes relatively similar number of articles except "Computer Science".

〈Table 2〉 Major Discipline Areas of Journals

Discipline	Articles[1]	Discipline	Articles
Computer Science	73	Management	36
Law	36	Electronics / Telecom Engineering	35
Extra Science & Technology	29	Statistics	24
Interdisciplinary	23	Telecom Theory / Applications	21
Medical	21	Extra Engineering	18
General Social Science	15	Industrial Engineering	13
Electrical Engineering	12	Design	12

[1] Number of articles published on specific discipline during 2011-2015.

4.3 Distribution of Major Keywords

Note that the keywords are given by the authors of the article. Therefore, there is no objective way of categorizing them. Based on personal judgment, keywords with similar meanings were grouped together and we compiled the major keywords from 477 articles as in <Table 3>.

The most significant keyword is the group of SNS, SOCIAL NETWORK ANALYSIS, and SOCIAL NETWORK SERVICE, used in 44 articles. Along with this keyword, SOCIAL DATA/MEDIA (23), TWITTER (14), OPINION MINING (8) convey similar contexts in terms of massive and unstructured data processing and applications. New big data processing technologies such as HADOOP (38) and MAPREDUCE (29) are also popular keywords.

〈Table 3〉 Major Keyword Distribution

Keyword	Articles[1]	STEP[2]
SOCIAL NETWORK SERVICE / ANALYSIS	44	2, 3, 4, 5
HADOOP	38	2, 3
MAPREDUCE	29	2, 3
PERSONAL INFORMATION POLICY/ PROTECTION/PRIVATE INFORMATION	28	1, 2, 5
CLOUD / COMPUTING	25	2, 3, 4
VISUALIZATION	24	4
DATA MINING	23	4
SOCIAL DATA/ MEDIA	23	2, 3, 4, 5
DISTRIBUTED SYSTEM / PROCESSING/COMPUTING	21	2, 3
TEXT MINING / ANALYTICS	21	3, 4
PUBLIC DATA / SECTOR	18	2, 3
SMART CITY / DEVICE/SOCIETY	18	1, 2
SPATIAL DATA / ANALYSIS	16	3, 4
CRM / CUSTOMIZATION	15	5
HEALTH / HOSPITAL	15	5
TWITTER	14	2, 5
SENTIMENT ANALYSIS / DICTIONARY	13	2, 3, 4, 5
INFOGRAPHIC	11	4
CLUSTERING	10	4
RISK FACTOR / MANAGEMENT / PREDICTION	10	5
WEB ANALYTICS / CONTENT	10	4
DE-IDENTIFICATION / RIGHT TO BE FORGOTTEN	9	5
SECURITY MANAGEMENT / POLICY	9	5
CRIME / CRIMINAL LAW	8	5
INTERNET OF THINGS / IOT	8	2
KNOWLEDGE SEARCH / SERVICE / MANAGEMENT	8	5
OPINION MINING	8	2, 3, 4
SPORTS INDUSTRY / MARKETING / ANALYSIS	8	5
TOTAL	497	51

[1] Number of articles where the keyword appears.
[2] Placement of keyword follows from the STEP defined in the Process Model of [Fig. 1]. A keyword may be assigned to more than one STEP. The placement is based on the author's judgment.

In <Table 3> we also provide the distribution of STEPs where each of the keywords can be placed according to the framework of [Fig. 1]. For example, since HADOOP is used for both capturing raw data and preprocessing before specific algorithms are applied, the keyword is placed in the STEP 2 and 3. SOCIAL NETWORK ANALYSIS research may span all the stages from the preprocessing unstructured raw data through utilizing or applying the derived

knowledge in the real business practice, thus being placed in STEPs 2, 3, 4, and 5.

〈Table 4〉 Distribution of STEPS

STEP	Frequency
STEP 1	2
STEP 2	13
STEP 3	11
STEP 4	12
STEP 5	13
Total	51

In <Table 4>, the frequency of each STEP appeared in the major keyword list of <Table 3> is tabulated.

<Table 4> shows that research efforts on big data are almost evenly distributed across the STEP 2 through STEP 5. It is remarkable that very few keywords are in STEP 1, or planning stage.

By scanning the list, however, we failed to find the future oriented directional keywords such as "education", "creativity" [14], etc. This might indicate that big data is just another wave of temporary IT upheaval. Not just as an extension of previous data analysis methods such as OLAP and data mining, we need to incorporate big data into sustainable managerial innovation process. To do that, we wish to see in the future more keywords such as creativity and decision making [5], which are not based on predefined top-down, but based on bottom-up [4,7] and prototyping approach. As a reference, both bottom-up and prototyping approaches to big data planning were discussed in [9,12]. These are quite different from the traditional top-down. This is in the same line of Stanford's d school empathy or design thinking paradigm.

5. Conclusion and Limitations

From the bibliometric data of KCI, we found the research on big data just starated in 2011, very young discipline. Based on the filtered bibliometric data using the keyword search by "big data", we compiled the distributions of the number of articles, the number of co-authors, the theme of contributing journals, major keywords, and the placement of STEPs defined in big data value added processing mode of [Fig. 1]. These distributions of 5 years along with the conjectures provided us windows through which we look into the structure of the big data research domain.

Based on the current sub domain of big data research, we also suggested the normative direction of big data research in the future.

In this research we identified important keywords, but we did not investigate the interrelationship among the keywords using network analysis tools, which might be the direct extension of the current research. However, to have meaningful results from the network analysis, we first need to come up with consistent and reliable methods to handle ill-structuredness of keywords.

ACKNOWLEDGMENTS

This research was supported by the 2015 Hannam University Research Fund.

REFERENCES

[1] H. Chen, R. Chiang, and V. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact", MIS Quarterly, Vol. 36, No. 4, pp. 1165-1188, 2012.

[2] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey", Mobile Networks and Applications, Vol. 19, No. 2, pp. 179-20, 2014.

[3] J. Choi, S. Yi, and K. Lee, "Analysis of keyword networks in MIS research and implications for predicting knowledge evolution", Information & Management, Vol. 48, pp. 371-381, 2011.

[4] A. Hayashi, "Thriving in a Big Data World", Sloan

Management Review, Vol. 55, No. 2, pp. 35-39, 2014.

[5] D. Kiron, P. Prentica, and R. Ferguson, "Raising the Bar With Analytics", Sloan Management Review, Vol. 55, No. 2, pp. 29-33. 2014.

[6] J. Lee, S. Han, and K. Kwon, "A Study on Korean legal research trend during the last 10 years based on keyword network analysis", Ajou Law Studies, Vol. 8, No. 4, pp. 519-539, 2015.

[7] V. Mayer-Schonberger and K. Cukier, "Big Data: A Revolution That Will Transform How We Live, Work, and Think", John Murray, 2013.

[8] A. McAfee and E. Brynjolfsson, "Bid Data: The Management Revolution", Harvard Business Review, Vol. 90, No. 10, pp 61-68, 2012.

[9] H. Mintzberg and F. Westley, "Decision making: It's not what you think", Sloan Management Review, Vol. 42, No. 3, pp 89-93, 2001.

[10] S. Namn and S, Seol, "Coauthorship Analysis of Innovation Studies in Korea: A Social Network Perspective", Korea Technology Innovation Studies, Vol. 10, No. 4, pp 605-628, 2007.

[11] S. Namn, J. Park, S. Seol, "Quantitative Analysis of Knowledge Flow - Technology Innovation Reserach in Korea", Korea Technology Innovation Studies, Vol. 8, No. Special Issue, pp 337-359, 2005.

[12] S. Namn and K. Noh, "A Study on the Effective Approaches to Big Data Planning", Journal of Digital Convergence, Vol. 13, No. 1, pp 227-235, 2015.

[13] W. Oh, J. Choi, and K. Kim, "Coauthorship Dynamics and Knowledge Capital: The Patterns of Cross-Disciplinary Collaboration in Information Systems Research", Journal of Management Information Systems, Vol. 22, No. 3, pp 265-292, 2006.

[14] Society of Digital Policy & Management,"Big Data Analytics for Business", WOW Pass, 2015.

[15] Society of Digital Policy & Management, "Big Data Analytics for Business", Kwangmoonkag, 2015.

남 수 현(Namn, Su Hyeon)

- 1982년 2월 : 고려대학교 통계학과 (경제학사)
- 1988년 8월 : Texas Tech Univ 경영대학원 (경영정보학 석사)
- 1996년 5월 : Rutgers Univ 경영대학원 (경영정보학 박사)
- 1996년 9월 ~ 현재 : 한남대학교 글로벌IT경영전공 교수
- 관심분야 : 빅데이터분석, 네트워크이론 활용, 지식관리
- E-Mail : namn@hnu.kr