Visualization Analysis of NoSQL Research Field Based on SCI by CiteSpace Ⅴ

Ming He, Ying Zhang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: heming2018@foxmail.com

Jianning Zhang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: zhang1108545@gmail.com

Pixian Zhao
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: zhaopixian@bnuz.edu.cn

Yingxin She
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: syxingin@163.com

Yongjun Wu
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: jackripperwu@foxmail.com

Qike Jiang
School of Management
Beijing Normal University, Zhuhai
Zhuhai Guangdong, China
E-mail: jiangqike0504@foxmail.com

***Abstract*—NoSQL is one of the technical trends that rises in this context in the Web 2.0 Era. With the aim to explore the research status and development trends related to NoSQL technology, articles between 1998 and 2016 were collected from Thomson ISI's SCI. After the analysis by using CiteSpace Ⅴ, the pivotal documents related to NoSQL, as well as institutions, co-citation patterns, research hotspots and frontiers, etc., were visualized and identified.**

Keywords-NoSQL; Visual analysis; CiteSpace; SCI-E; Database

I. INTRODUCTION

With the continuous development of Internet technology, vast amounts of data have emerged in all areas of social life and scientific research. Faced with such a large-scale data, particularly in the SNS type of high concurrency scenarios, it has been a bit of powerless to store and query the users' dynamic data by using relational database [1]. There are a lot of advanced data management technology to alleviate this problem, NoSQL (originally referring to "non SQL", "non relational" or "not only SQL") is just one of the technical trends that rise in this context.

The databases like "NoSQL" have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century [2], triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com [3, 4].

Johan Oskarsson reintroduced the term NoSQL in 2009 when he organized an event to discuss "open source distributed, non-relational databases". The name attempted to label the emergence of an increasing number of non-relational, distributed data stores, including open source clones of Google's BigTable/MapReduce and Amazon's Dynamo.

The main objectives of this study are to identify the current development status, trends and frontiers, to find core researchers as well as their co-citation situation, and to detect the pivotal documents in the research area of NoSQL from 1998 to 2016.

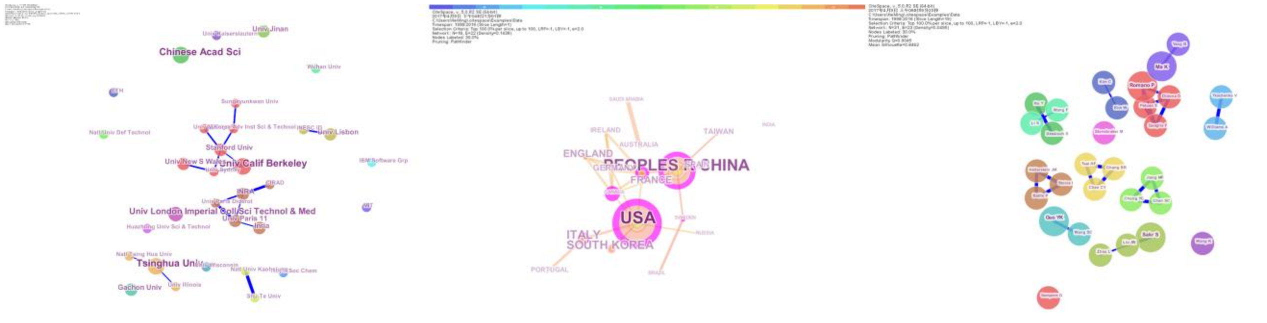
II. DATA SOURCES AND RESEARCH METHODS

A. Data Source

The data analyzed by this study is from Thomson ISI's SCI (Web of science in the Science Citation Index Expanded Edition). The time of collecting data is March 31, 2017. The author set the search mode to advanced search with the following formula: "TS = ((nosql) OR (non-relational database $) OR (nonrelational database $))". The timespan is set to 1998-2016 and the language is English. A total of 144 records include authors, titles, keywords, abstracts, and cited references.

B. Research Tools

In 2004, CiteSpace was first developed by Chaomei Chen to facilitate the visual analysis of trends, patterns, and critical changes in a changeable information environment [5]. In CiteSpace, Timeline views and time-zone views display the publication time and peak time of articles and terms, Cluster views is node and link diagrams, where the nodes present author, institution,



The institution network related to NoSQL researches (1998-2016)
(a)

The country network related to NoSQL researches (1998-2016)
(b)

The author network related to NoSQL researches (1998-2016)
(c)

Figure 1 The network related to NoSQL researches (1998-2016)

Country, term, keyword, cited reference, cited journal, and so on [6, 7]. The node size represents the overall citation frequency. Link represents co-citation or co-occurrence, the line's thickness represents the strength proportion of co-citation or co-occurrence. Each color corresponds to a time slice following the legend bar above the visualization area. However, if a node has a purple ring, which means that the node has a high betweenness centrality and tends to be strategically important in terms of the macroscopic structure of a new work. Those nods with high betweenness centrality are called pivotal points or turning points; if a node has a red ring; it means the node has burst in one of its attributes, notably citations [8].

Therefore, researchers can easily analyze the trends, patterns, and critical changes by studying the size, color of nodes, and links of colorful network. The version of the CiteSpace 5.0. R2 SE was the main research tool used in this paper.

III. DISTRIBUTION OF NOSQL RESEARCH FIELDS ANALYSIS

Distribution on research field analyses, including institution co-occurrence, country co-occurrence, author co-occurrence, are used to reveal the development status of NoSQL from different dimensions.

A. Institution Co-occurrence Network Analysis

Node type on the interface of CiteSpace was selected as the network node for the analysis of institutions. Because the total of institutions was fewer, time slicing was set to 19, which is the maximum value. Through running CiteSpace, then we can get holding the Fig. 1 (a) with 32 institutions and 17 links. Each country issued a relatively average number of articles. Among these institutions, Chinese Acad Sci, Tsinghua Univ, and Univ Calif Berkeley issued the largest number of documents.

B. Country Co-occurrence Network Analysis

The author set the node type to Country and time slicing to 1, and then run CiteSpace. A network consisting of nodes represented collaborating countries is presented in Fig. 1 (b).

In the NoSQL field, the United States has the greatest advantage, living in the world's first, and cooperation with other countries or regions more closely. China ranked second, significantly more than other countries and regions, but relatively less in terms of cooperation, followed by Canada, France, etc.

C. Author Co-occurrence Network Analysis

A total number of 31 authors and 23 links between the authors were shown in Fig. 1 (c). Because authors belong to the organization, so the cooperation between the authors is similar to that of the institution.

Romano P (Paolo Romano), Sakr S (Sakr Sherif), Guo YK (Yike Guo), Ma K (Ma Kun). The four authors have published the most documents and have a higher frequency of collaboration with other authors.

D. Paolo Romano

Dr. Paolo Romano is a senior researcher at the division systems group at INESC-ID, his main research directions are Distributed Data Management, Dependability, Cloud Computing, and Autonomic Computing.

E. Sakr Sherif

Sakr Sherif is currently a Professor of Computer Science at King Saud bin Abdulaziz University for Health Sciences. His research interest is data and information management in general, particularly in big data processing systems, big data analytics, data science and big data management in cloud computing platforms. And Dr. Sakr has published more than 100 refereed research publications in international journals and conferences.

F. Yike Guo

Yike Guo is an Imperial College London Parallel Computing Center Technical Director, Lifetime Professor in London E-Science Research Center, and the Chairman and CEO of the board of directors of InforSense Ltd. His main research direction is large-scale data mining and parallel computing.

G. Ma Kun

Ma Kun, Professor of Key Laboratory of Intelligent Computing Technology for network environment in Shandong Province, member of IEEE, member of China Computer Federation, the main research directions are Big Data Management for Multi-tenant Applications in the Cloud, and Data Intensive Computing.

IV. CO-CITATION NETWORK ANALYSIS OF NOSQL

A. Author Co-citation Network Analysis

The higher the frequency of the authors, the stronger the academic authority, so the author have analyzed for authors cited by the data above this paper. Fig. 2 is the Author co-citation map with 180 authors and 549 links generated by CiteSpace and the node named anonymous deleted. Form Fig.2, three core researchers are presented.

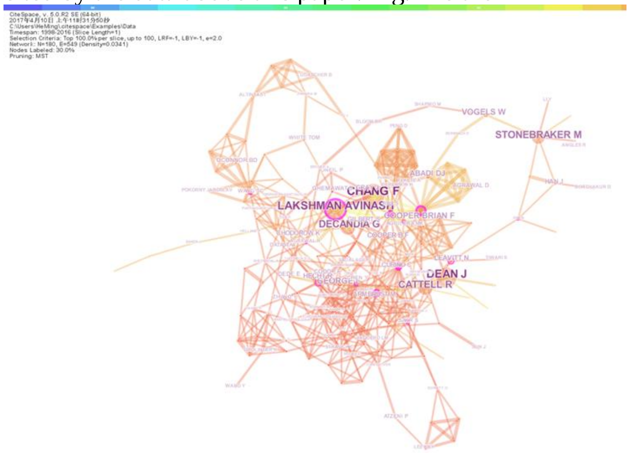

CiteSpace, v. 5.0.R2 SE (64-bit)
2017年4月10日 上午11时31分50秒
C:\Users\HeMing\.citespace\Examples\Data
Timespan: 1998-2016 (Slice Length=1)
Selection Criteria: Top 100.0% per slice, up to 100, LRF=-1, LBY=-1, e=2.0
Network: N=180, E=549 (Density=0.0341)
Nodes Labeled: 30.0%
Pruning: MST
VOGELS W
STONEBRAKER M
WHITE TOM
OCONNOR BD
ABADI DJ
AGRAWAL D
HAN J
CHANG F
LAKSHMAN AVINASH
DECANDIA G
COOPER BRIAN F
COOPER B F
CHODOROW K
LEAVITT N
DEAN J
CATTELL R
GEORGE L
HECHT R
WANG Y
ATZENI P

Figure 2 The author co-citation map related to NoSQL researches(1998-2016)

TABLE I. is the key cited authors in the author co-citation map and the cited number more than 10.

B. Fay Chang

The largest node is CHANG F (Fay Chang). He is now at Google. In his research, he worked on developing a general, automatic approach to I/O prefetching based on speculative execution. Prior to work at Google, he worked on the Network-attached Secure Disks (NASD) project.

Fay Chang worked with Dean Jeffrey as the first author to complete *Bigtable: a distributed storage system for structured data*.

C. Jeffrey Dean

The second largest node is DEAN J (Jeffrey Dean). Jeff Dean is a Google Fellow in the Systems Infrastructure Group. A summa cum laude graduate of the University of Minnesota with a M.S. degree in Computer Science, he obtained a Ph.D. degree in Computer Science from the University of Washington.

Research areas include large–scale distributed systems, performance monitoring, compression techniques, information retrieval, microprocessor architecture, compiler optimizations. Products Jeff has developed for Google include AdSense, MapReduce, BigTable, and Google Translate.

His 6 papers are included in Web of Science Web of Science Core Collection such as *MapReduce: Simplified data processing on large clusters*.

D. Lakshman Avinash

Lakshman Avinash is the third largest node with a purple ring, which means that he has a high betweenness centrality and tends to be strategically important in terms of the macroscopic structure of a new work. The reason he was known to most of the people is that he co-invented Amazon Dynamo and invented Apache Cassandra. His main papers include Cassandra: a decentralized structured storage system (cited 297 in Web of Science Core Collection), Dynamo: amazon's highly available key-value store.

Currently Avinash is the CEO and co-founder of Hedvig founded in 2012. Hedvig is positioned as a pure SDS (Software Defined Storage) company to help companies become more responsive to the data demands of today's digital businesses, which received a $ 21.5 million C round of financing on March 4, 2017.

TABLE I. THE KEY CITED AUTHORS IN THE AUTHOR CO-CITATION MAP(THE CITED NUMBER≥10)

No.	Frequency	Centrality	Year	Author
1	34	0.03	2010	CHANG F
2	31	0.08	2013	DEAN J
3	28	0.21	2012	LAKSHMAN AVINASH
4	22	0.01	2012	DECANDIA G
5	21	0.04	2013	CATTELL R
7	20	0.07	2014	STONEBRAKER M
8	10	0.04	2014	GEORGE L
9	10	0.05	2014	VOGELS W

E. Document Co-citation Analysis

Fig. 3 shows a document co-citation map with 97 documents and 242 co-citation links. Each node in the graph represents a document in which the thickness of the circle is proportional to the number of citations in the corresponding year.

The core documents shown in Fig. 3 constitutes the most important knowledge foundation in the NoSQL domain. There are 8 core documents cited more than 5 times shown in TABLE II.

Bigtable: A distributed storage system for structured data

The most cited core document is *Bigtable: A distributed storage system for structured data* published in 2008 by Chang Fay as the first author and Jeffrey Dean, and the citation frequency is 18 times. This paper laid the foundation of HBase which was published in 2006 and cited up to 510 times in Web of Science Core Collection. Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers [9]. This article describes the simple data model, dynamic control data layout and format for the client provided by Bigtable, and describes the design and implementation of Bigtable.

F. MapReduce: Simplified data processing on large clusters

The second place is MapReduce: Simplified data processing on large clusters published in 2008 by Dean Jean. MapReduce is a programming model and an associated implementation for processing and generating

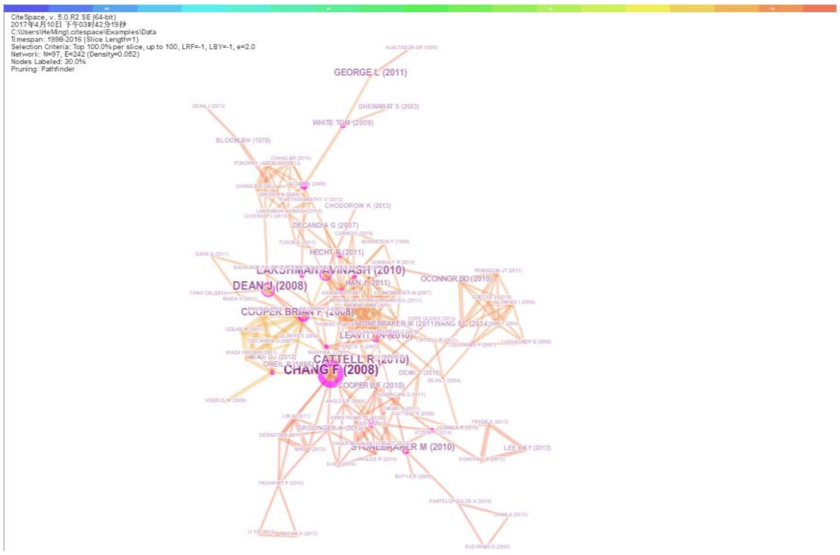


Figure 3 The document co-citation map related to NoSQL researches(1998-2016)

TABLE II. THE KEY CITED AUTHORS IN THE AUTHOR CO-CITATION MAP(THE CITED NUMBER≥5)

No.	Frequency	Centrality	Author	Year	Document
1	18(510)	0.7	CHANG F	2008	Bigtable: A distributed storage system for structured data
2	12(3438)	0.18	DEAN J	2008	Mapreduce: Simplified data processing on large clusters
3	12(168)	0.03	CATTELL R	2010	Scalable SQL and NoSQL Data Stores
4	11(297)	0.17	LAKSHMAN AVINASH	2010	Cassandra: a decentralized structured storage system
5	9(160)	0.32	COOPER BRIAN F	2008	Pnuts: Yahoo!'s hosted data serving platform
6	8(143)	0.04	GEORGE L	2011	Hbase: The Definitive Guide: Random Access to Your Planet-Size Data
7	8(87)	0.12	STONEBRAKER M	2010	SQL Databases v. NoSQL Databases
8	7(90)	0.16	LEAVITT N	2010	Will NoSQL Databases Live Up to Their Promise?

The number in brackets is the cited times in Web of Science Core Collection

large data sets. And many real world tasks are expressible in this model, as shown in this paper [10]. The citation of the article in the Web of Science Core Collection is as high as 3438 times.

V. RESEARCH HOTSPOTS AND FRONTIERS

A. Research Hotspot Analysis

Keyword is the core in obtaining the information of an article. Only by accurately grasp the distribution of key words can we better analyze and study the hotspots.

Keywords were set as the network nodes for analysis. By selecting all the documents that appear in each time period, the keywords knowledge mapping was constructed with some irrelevant keywords (such as: lung cancer) deleted. We can get the Fig. 4 with 38 keywords and 58 links.

The largest node in the mapping except NoSQL is nosql database, other larger nodes are big data, cloud computing, mapreduce, mongodb, and so on. Based on Fig. 4, we can find the main areas of the current development states of NoSQL and several important branches of NoSQL domain.

B. Research Trend and Frontier Analysis

Research frontier analysis can provide researchers the latest information, and then quickly provide valuable information or references in their potential research area [8]. The development trends and research frontiers can be analyzed according to the keyword frequency changed in the trend. Therefore, the author changed the keyword

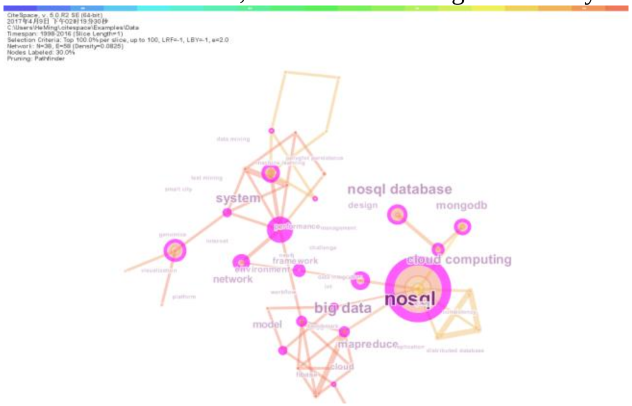

CiteSpace, v. 5.0.R2 SE (64-bit)
2017年4月9日 下午02时19分30秒
C:\Users\HeMing\.citespace\Examples\Data
Timespan: 1998-2016 (Slice Length=1)
Selection Criteria: Top 100.0% per slice, up to 100, LRF=-1, LBY=-1, e=2.0
Network: N=38, E=58 (Density=0.0825)
Nodes Labeled: 30.0%
Pruning: Pathfinder
data mining
text mining
system
nosql database
design
mongodb
performance
management
framework
environment
network
cloud computing
nosql
big data
model
mapreduce
cloud
distributed database

Figure 4 The keyword network related to NoSQL researches (1998-2016)

network to time zone view which provided by CiteSpace, and got Fig. 5.

Research Trends

C. The origin of NoSQL

Although the "data warehouse" in 1998 is the first hotspot in Fig. 7, NoSQL is not originated from the "data warehouse" according to the reference situation of view. It is obvious that NoSQL has evolved from "system", "model", "network", "database" between 2006 and 2007.

1) Development period

In 2009-2010, the concept of opening distributed non-relational database was proposed, but the development of this technology is still at an unexpected stage.

Until 2011, with the rise of "cloud computing" and "MapReduce" technology, NoSQL field has also been unprecedented developing.

In the 2012-2013, NoSQL technology continues to develop forward, and puts some theoretical knowledge into practice, mainly reflected in the nodes named as “NoSQL database”, “Web”, “cloud storage”, and so on.

2) differentiation trend

Since 2014, NoSQL researches gradually split into various fields, mainly focusing on fields such as "bioinformatics", "text mining".

From 2015 to 2016, the differentiation trend is more pronounced. Some main research topics are “mongodb”, “framework”, “cloud”, “hbase”, “data integration”, etc.

Therefore, it can be seen that NoSQL has experienced from the concept to the continuous improvement, and achieved the change from technology theory to practice.

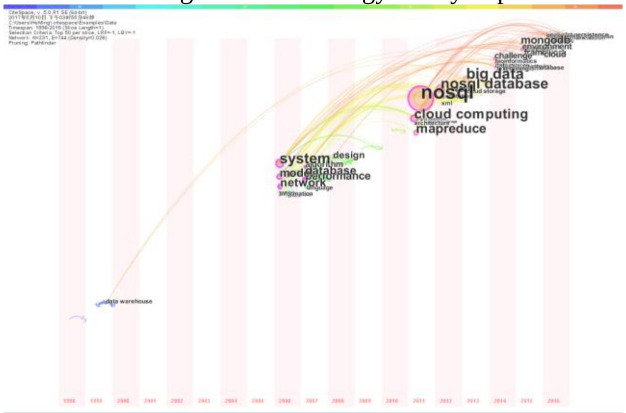


Figure 5 The keyword time zone network related to NoSQL researches (1998-2016)

TABLE III. THE KEYWORDS RELATED TO NOSQL RESEARCHES FROM 2014 TO 2016 (THE FREQUENCY>1)

Frequency	Centrality	Year	Keyword	Frequency	Centrality	Year	Keyword
4	0.13	2014	challenge	2	0.05	2015	association
2	0.11	2014	data mining	2	0.02	2015	workflow
2	0.1	2014	bioinformatics	4	0.05	2016	cloud
2	0.06	2014	availability	3	0.01	2016	hbase
2	0.04	2014	internet	2	0.07	2016	data integration
2	0.04	2014	smart city	2	0.04	2016	neo4j
2	0.04	2014	text mining	2	0.02	2016	benchmark
2	0	2014	consistency	2	0.02	2016	platform
2	0	2014	distributed database	2	0	2016	genomics
2	0	2014	machine learning	2	0	2016	iot
2	0	2014	replication	2	0	2016	management
7	0.01	2015	mongodb	2	0	2016	polyglot persistence
4	0.15	2015	framework	2	0	2016	query
4	0.11	2015	environment	2	0	2016	visualization

And now it is attempting to perfect each branch and utilized by more fields.

Research Frontiers

By the analysis of NoSQL research trend in recent years, the possible frontiers in the future can be found out, such as “data mining”, “data integration”, “and iot”.

The detailed information of keywords which occurrence frequencies higher than 1 from 2014 to 2016 are listed in TABLE III.

VI. CONCLUSION

In the paper, using mapping knowledge domains Software CiteSpace Ⅴ, visualization analysis on 144 documents in NoSQL field were studied to analyze research distribution, co-citation situation as well as the hotspots and frontiers.

A. It Revealed the Distribution Situation of NoSQL Research between Countries, Institutions, and Authors

- The Chinese academy of sciences, Tsinghua University, University of California Berkeley, have published the largest number of documents, leading the

- way in NoSQL researches.

- The United States ranks first in the world in the research field of NoSQL, followed by China, Canada, and France.
- In these 144 documents, Paolo Romano, Sakr Sherif, Yike Guo, Ma Kun have published the most documents and have a higher frequency of collaboration with other authors.

B. It Defined the Key Researchers and Documents

- Fay Chang, Jeffrey Dean, and Lakshman Avinash are three core researchers in which Lakshman Avinash may bring greater influence to NoSQL research in the future because he has a high betweenness centrality.
- Bigtable: A distributed storage system for structured data and MapReduce: Simplified data processing on large clusters laid the foundation for NoSQL researches.
- It showed the hotspots and frontier related to NoSQL ersearches
- The research hotspots are shown out, such as big data, NoSQL database, system cloud, computing, MapReduce, mongodb, etc.
- The technical frontiers of NoSQL like data mining, data integration are discovered.

In the end, this study is summarized: NoSQL research started relatively late, but the development is very rapid. Only in about ten years, it has experienced from reintroduce of concept to continuous improvement, and achieved the change from technology theory to practice with a lot of excellent researchers springing out. There is still huge development space to research and explore in the future.

References

[1] Kai Fan. "NoSQL database overview." programmer 6(2010):76-78.

[2] Leavitt, Neal. "Will NoSQL databases live up to their promise?." Computer 43.2 (2010).

[3] Mohan, C. "History repeats itself: sensible and NonsenSQL aspects of the NoSQL hoopla." Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013.

[4] "Amazon Goes Back to the Future With 'NoSQL' Database."unpublished.

[5] Chen, Chaomei. "Searching for intellectual turning points: Progressive knowledge domain visualization." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5303-5310.

[6] Chen, Chaomei. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." Journal of the American Society for information Science and Technology 57.3 (2006): 359-377.

[7] Chen, Chaomei, Fidelia Ibekwe ‐ SanJuan, and Jianhua Hou. "The structure and dynamics of cocitation clusters: A multiple ‐ perspective cocitation analysis." Journal of the American Society for Information Science and Technology 61.7 (2010): 1386-1409.

[8] Liu, Hailong, et al. "Visualization Analysis of Subject, Region, Author, and Citation on Crop Growth Model by CiteSpace II Software." Knowledge Engineering and Management. Springer Berlin Heidelberg, 2014. 243-252.

[9] Chang, Fay, et al. "Bigtable: A distributed storage system for structured data." ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.

[10] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." Communications of the ACM 51.1 (2008): 107-113.