Exploring the networks, hot-topics and clusters of big data: evidence from Internet

Wenjie Zhou，Yu Bai
Business School, Northwest Normal University，China
*wj_lp@sina.com

Keywords：big data; networks; hot-topics; clusters; Internet

Abstract. Present study shed light on the networks, hot topics and clusters of big data via Social Network Analysis based on the retrieval of Internet. The findings of this study include: The term ‘big data’ has strong ties with service, platform, analysis, technology, application and enterprise. The technologies in relate to big data is a hottest topic on Internet nowadays. Internet is both the closest word and the most influential word to big data in the network. 8 clusters were found in the big data network and each cluster showed the specific aspect of big data network. The findings of this study may helpful for both big data related research and business.

1 Background

Big data become a hot discussed topic in recent few years. Ever so many people focus on this topic; however, some questions, such as what is the big data and what kinds of factors are in related to so call big data, remind vague. Aiming to explore the landscape of the meaning of big data, this paper shed light on the networks, hot discussed topics and clusters of big data through a social network analysis procedure based on the data collected from Internet.

2 Research Design

2.1 Data collection

We retrieved the keyword ‘big data’ through Baidu, which is the most popular search engine in Chinese, on 02/04, 2016. The scan of the retrieval period limited to the recent week of the day we retrieval. Finally, we got 1000 papers, articles, and web news from Internet. Then, a word-frequency analysis procedure was conducted, thus, a co-words metric is available to be analyzed further. Based on the co-words metric, we explored the key words network of big data and found hot discussed topics and clusters among big-data related papers and news on Internet through Social Network Analysis procedure.

2.2 Instruments

A word frequency analysis program developed by Waseda University, named AntConc, is chosen to find the exact word frequency and their rank of corpus. Then, we use Ucinet which developed by Borgatti, S.P., Everett, M.G. and Freeman, L.C. in 2002 is selected to find the networks, hot-topics and clusters of big data.

3 Results

3.1 Networks

Through Uninet program, networks under different level of strength, ties equals to 1, 50, 100, 150, 200, 300, of high frequency words metric, are available as Fig. 1.

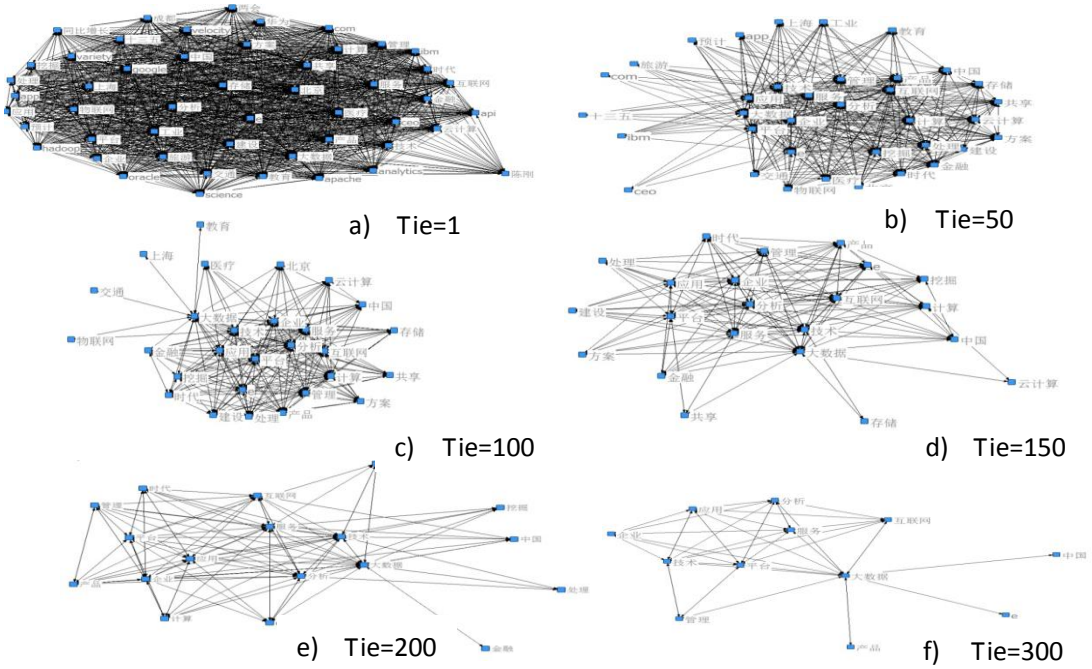

Fig 1 networks of big data at different ties

Fig 1-e) shows that the term 'big data' has strong ties with service, platform, analysis, technology, application and enterprise. This means that from the perspective of Internet, people tend to view big data as a tool which could be used as a platform to conduct some tasks as service and analysis. From fig 1-f) to 1-a), we show the words link between big data and other key words from different ties. As these figs showing, more items become connected to the big data. Such as the terns mining, management, production and so on. We will make a detailed analysis on these terms in following part of this paper.

3.2 Hot topics

3.2.1 Centrality degree

Centrality degree is a very useful tool to explore the hot discussed topics among a specific field (Freeman, 1979; Snijders, 1981a, 1981b). As table 1 shows, the technologies in relate to big data is a hottest topic on Internet nowadays. This means that from the perspective of Internet, big data is considered as technological issue first. And then service, analysis, application and management are located at top 10 hot discussed issues which close in related to big data. This shows the main ways that big data serve the society as big data service, big data analysis, big data application or big data management.

Table 1 Top 10 high centrality degree words of the big data network

R	word	Degree	R	word	Degree
1	Big data	35.341	6	application	27.020
2	technology	29.293	7	platform	26.983
3	service	28.808	8	internet	24.009
4	analysis	27.899	9	management	22.536
5	enterprise	27.767	10	production	20.964

Table 2 Top 20 high closeness degree words of the big data network

R	word	R	word	R	word	R	word
1	Big data	6	Beijing	11	technology	16	Compute
2	Internet	7	Time	12	enterprise	17	service
3	Production	8	Travel	13	e	18	sharing
4	Medical	9	financial	14	management	19	industrial
5	Cloud compute	10	analysis	15	platform	20	storage

3.3.2 Closeness degree

According to Hakimi (1965) and Sabidussi (1966), closeness degree shows the minimum steps from one actor to anther in a network. In another word, the higher closeness degree a word has the higher relationship it will have with the central word, big data. Table 2 shows top 20 high closeness degree words of Fig1-a). We listed those words which the farness of these words equal to 49 and closeness equal to 100.

As Table 2 shows, Internet is closest word to big data in the network. This means that big data is mainly based on the Internet. As we could find in table 2, some professionals are close related to big data, such as medical, travel, financial and industrials. And some items are closely in related to the outcomes of big data, such as production, cloud compute, sharing and storage.

Table 2 Top 20 high closeness degree words of the big data network

R	word	R	word	R	word	R	word
1	Big data	6	Beijing	11	technology	16	Compute
2	Internet	7	Time	12	enterprise	17	service
3	Production	8	Travel	13	e	18	sharing
4	Medical	9	financial	14	management	19	industrial
5	Cloud compute	10	analysis	15	platform	20	storage

3.2.3 Proximal Betweenness degree

In the social network, if the interactions between several actors were controlled by a specific actor potentially, the controller tends to have a higher between then others. In another words, if an actor has a high betweenness degree, this actor will affect others a lot (Freeman, 1979; Friedkin, 1991).

Table 3 shows that Internet is the most influential word in the big data network. This is reasonable because as an information source, internet providing the most information of data for big data analysis. And cloud compute also has a high betweenness degree. This shows that cloud compute may control other big data analysis behavior potentially. Besides, Beijing, as the captain city of China is very influential for big data due to there are many academic institutions located at Beijing and the city becomes a core area for both research and business of big data. Other words have high betweenness degree include financial, analysis, medical, service and so on. These words have a similar influence among the big data network.

Table 3 Top 20 proximal Betweenness degree words of the big data network

Rank	word	Rank	word	Rank	word	Rank	word
1	Big data	6	analysis	11	ceo	16	construction
2	Internet	7	medical	12	technology	17	industrial
3	Cloud compute	8	service	13	enterprise	18	time
4	Beijing	9	e	14	platform	19	travel
5	Financial	10	storage	15	production	20	management

3.3 Clusters

Aiming to explore the cluster among the big data network, we conducted a hierarchical cluster analysis procedure via Ucinet and 8 clusters were formulated as Fig 2.

Fig 2 shows that following key words is located at the first cluster: education, transportation, medical etc. Obviously, the first cluster mainly focuses on the professional use of big data. Besides, some tools of big data such as app, cloud computing, mining and technology also involved in this cluster. This shows that the professional use of big data is closely connect to the tools. The second cluster formulated by 3 keywords: Shanghai, Chengdu and yearly growth. This shows that these 2 cities have an advantage on the development of big data business. Cluster 3-8 also show the close related keywords and the structure of big data network. Through these clusters, we could confirm what had happened to the big data research and practice via a perspective of Internet.

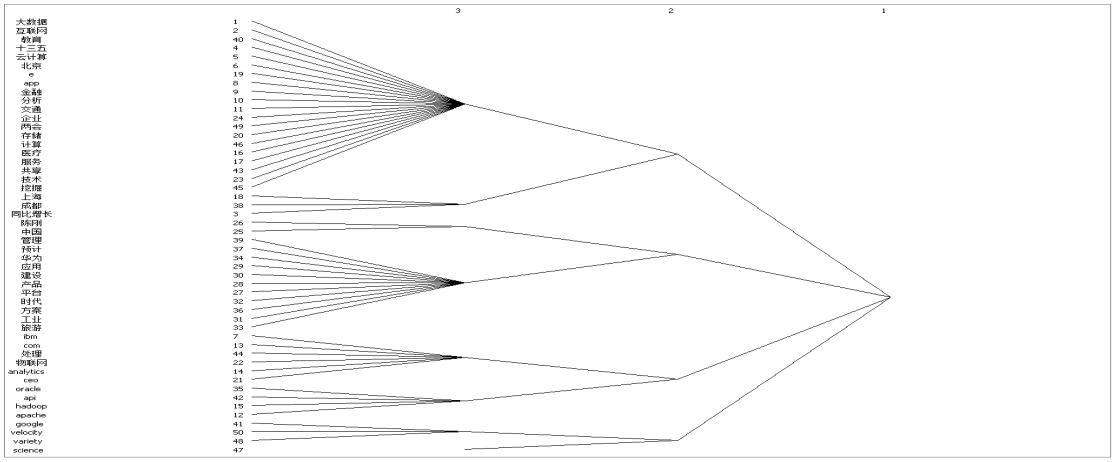

Fig 2 Clusters of the big data network

4 Implementation

Exploring the big data from a perspective of Internet will benefit us to get a better understanding of the development of both technology and research on big data. For the researchers, finding the hot discussed topics of big data may be helpful for the theoretical construction of big data. For the business institutions, finding the main area of big data will benefit the designing and implement of big data related programs.

5 Conclusion

Present study shed light on the networks, hot topics and clusters of big data via Social Network Analysis based on the retrieval of Internet. The findings of this study include: a) The term 'big data' has strong ties with service, platform, analysis, technology, application and enterprise. b) The technologies in relate to big data is a hottest topic on Internet nowadays. Internet is both the closest word and the most influential word to big data in the network. And, c) 8 clusters were found in the big data network and each cluster showed the specific aspect of big data network.

Acknowledgment

This article is an outcome of the project "Reliability and Validity of Co-words analysis among Scientometrics" (No. 71563042) supported by National Nature Science Foundation of China

References

[1] Freeman, L. G. Centrality in social networks: I. conceptual clarification. Social Networks, (1979)1, 255-239.
[2] Friedkin, N. E. Theoretical foundations for centrality measures. American Journal of Sociology, (1991)96, 1478-1504.
[3] Hakimi, S. L. Optimum location of switching centers and the absolute centers and medians of a graph. Operations Research, (1965)12, 450-459.
[4] Sabidussi, G. The centrality index of a graph. Psychometrika, (1966)31, 581-603.
[5] Snijders, T. A. B. The degree variance: an index of graph heterogeneity. Social Networks. (1981) 3, 163-174.
[6] Snijders, T. A. B. Maximum value and null moments of the degree variance. TW-report 229. Department of Mathematics, University of Groningen, 1981.