


Keywords co-occurrence mapping knowledge domain research base on the theory of Big Data in oil and gas industry

Lin Zhu¹  · Xiantao Liu¹ · Sha He¹ · Jun Shi¹ · Ming Pang¹

Received: 12 July 2015 / Published online: 14 August 2015
© Akadémiai Kiadó, Budapest, Hungary 2015

Abstract Taking the theses' keywords in China from 1986 to 2014 as the research materials, use the basis concept of the Big Data Theory to further study the keywords which related to oil and gas industry. Analyze the keywords frequency of the theses in oil and gas industry and its co-occurrence frequency pair, and then use the theory of mapping knowledge domain to visualize the keywords co-occurrence network in petroleum industry so as to make further research of the heated issues that mapping knowledge domain has shown. According to the research we can see that the application technology R&D (research and development) predominate the oil and gas industry, featuring a high concentration and long tail phenomenon (which means various researches focus on different kinds of things, the scale of the research is large).

Keywords Oil and gas industry · Keyword co-occurrence · Mapping knowledge domain · Big Data · Data mining

Introduction

The history of human development is a process of knowledge accumulation, and we human being constantly make progress by standing on the shoulder of the ancient wiser. Large amount of knowledge are maintained in various forms, in which one of the crucial ways to store these knowledge is document. Along with the evolution of human civilization, we have already stepped into the new era of Big Data (Lohr 2012), with 50 % or more data accumulation per 2 years. We need to explore and find out new rules when we are facing such a giant flood of complex data running in front of us, and keep moving. A report released on Davos economic forum named Big Data, Big Impact said that as a new kind of

✉ Lin Zhu
messagescn@gmail.com

¹ School of Economics and Management Science, Southwest Petroleum University, Chengdu, China

economic fortune, data is as valuable as current money and gold. Moreover, Big Data has already infiltrated into every aspects of our lives, helping us to make better choices and waiting for us to make deeper mining. Among all the Internet giant companies, Google and Facebook are the outstanding leader in the application of Big Data.

As a vital department in our national economy, oil and gas industry play an important part in maintaining security and stability of our country. Under the rough situation of finding oil and producing oil, scientific technology is the guarantee for petroleum industry to realize sustainable development on the account of this non-renewable resource industry a typical capital-technology intensive industry. Especially the shale gas revolution has set up a good example to improve the whole structure of social energy system. Great emphasis on science and technology research help oil and gas industry accumulated massive achievements which represent intelligence essence of the researchers and workers of this field and maintain further study. And the achievements help us to better understand the development process and heated issues in petroleum industry.

Literature review

Big Data

Big Data is the abbreviation of an advanced technology for helping us to know the world better and make decision. It not only means great amount of data resources and tools for operation, but also a transition of our thoughts and researches (Mayer-Schönberger and Cukier 2013). Decisions based on the analysis of Big Data have already been widely recognized. Moreover, it gained lots of attentions from the business to academic cycle (Boyd and Crawford 2012). Because of the wide range and large applied field, Big Data theory research is still under an irregular situation at present. People analyze the Big Data from every aspect in order to cope with the problems that Big Data Theory is facing and realize sustainable development. This is a problem that holds every industry, even every corner of the world's breath (Marx 2013; Boyd and Crawford 2012). Therefore, we should find out how to make further progress under the circumstance of the Big Data. MIT Sloan Management Review and IBM have carried out an investigation shown that the top companies have extremely attach importance to data analysis, from which half of the researchers considered data collecting and processing are crucially important to the development of a company (LaValle et al. 2013). Up till now, the Big Data research in social science field are mainly focus on the difference between Big Data technology and tradition database technology and where is the challenge; next, to see how business and academic field cooperate with each other in order to solve the challenge; the last problem is the management of Big Data and the construction of applied ecosystem. Compare to tradition database technology, Big Data, with features of large volume, in time and variety, aim at creating value so as to transform them into competition strength.

Mapping knowledge domain

Mapping knowledge domain, also call knowledge graph or knowledge visualization is the important part of knowmetrics and social network analysis research. Knowledge visualization has a long history starting from expressing the feeling by drawing in ancient time to nowadays complex but beautiful relationship network (Bertschi et al. 2011).

Knowledge visualization is now rapidly developing, which has already become a new research field that combined graph theory, bibliometrics, statistics and many other theories (Wang and Jacobson 2011). It materializes the abstract knowledge, making them easy to observe, as a result, helping more people to grasp the knowledge better. There are quite a lot research studied knowledge visualization, such as Zhao and Zhang (2011), who utilized mapping knowledge domain method to analyze e-library from 1994 to 2010; Hong and Haiyue (2012), who applied CiteSpace to dissect the earning management theses from 1988 to 2009 by using knowledge network co-occurrence method; Yue (2014) used CiteSpace to mapping analyze 548 core theses and master theses in order to find out the frontier theory; Hua et al. (2013) analyzed the co-occurrence of authors and keywords in Chinese Medical Equipment Journal from 2001 to 2010; Zheng et al. (2013) visualized the theses concerning government information released by China National Knowledge Internet (CNKI) for finding the heated issue from 2002 to 2012; Xiuling (2012) used CiteSpace to analyze 2916 information resource management theses in Web of Science with the purpose of summing up the heated issues; Wei (2013) described the features of the document distribution related to semantic network from 2003 to 2012 in CNKI; Zhichao (2012) used UCINET to visualize 14 high quoted authors aiming at mapping knowledge domain in order to find out how they collaborate with each other; Zins and Santos (2011) studied the classification system of the library through mapping knowledge domain; Cobo et al. (2011) utilized knowledge network tool to review and summarize; Wang et al. (2011) used knowledge visualization method to resolve the learning process for the sake of helping the learners to solve cognitive overload and concept lost; Pollack et al. (2014) took the knowledge graph of the complex system theory over the past two decades as the basis to analyze the current condition of it; Osinska and Bala (2014) studied the classification mode from the theses in terms of relevant researches; Somekh et al. (2012) used knowledge graph method to figure out what caused knowledge gap in molecular biology system; Eppler and Pfister (2013) said that knowledge visualization could promote the knowledge management level of the whole police department; Petra et al. (2012) utilized knowledge visualization network to classify the face recognition algorithm for better development; Biloslavo et al. (2012) emphasized on the importance of knowledge visualization in strategy decision; Womack (2014) highlighted that knowledge visualization method is a core skill to a library manager and a teacher. From the above researches of the knowledge visualization method we knew that the point to knowledge visualization is to take a keyword to search through a certain field, and use computer software to analyze the data for the rules of how knowledge transferred. As a result, we can say that knowledge graph mainly aim to help make reasonable choices for any organization by chasing the course of knowledge.

Data and method

Source of the data

The data of this thesis derived from CNKI database, which is the biggest continuous dynamic updated database in the whole world. Containing 6100 different kinds of CSSCI and professional journals, the database not only plays an important role in CNKI, and national knowledge infrastructure, but also be widely accepted by all walks of life.

According to Big Data Theory we know that sample data is hard to truly represent the features of all data. Therefore, take all data into calculation is the only way to fully grasp the rules of the real world, and gain the knowledge that were impossible to have before, rules that were never been discovered. In order to study the knowledge network and analyze the heated issues better in oil and gas industry, we take the total theses data in oil and gas engineering and oil and gas chemical engineering as our research objects in CNKI database, from which we found 363,458 theses. Among them there were 44 theses did not have accurate published date and 96,361 theses did not have keywords. The reason for these errors is because we did not have a standard at that time. Hence, the research objects that fit our requirements are 267,097, see Fig. 1.

Ways to research

The way to analyze the keywords in theses is: firstly, extract the keywords and then separate them, classify the keywords so as to calculate the frequency. The way to figure out keyword co-occurrence is to extract the keywords in the theses first. It must be certain relation within different keywords in the same articles. We suppose that each keywords are equal important to the article, then we can get the weight of the keywords are all 1. Assume that the number of keywords in one article are N , then we can generate C_n^2 keyword co-occurrence pairs. We can get keyword co-occurrence matrix after summarization, which lay a solid foundation of data analysis.

The way to analyze core network of keyword co-occurrence in theses is to segment the network according to the frequency distribution of the keyword co-occurrence pair and degree. Select the highest 100 sub-network as research subnet, take 28×2 as boundary value to cut off the boundary, chip the edge and used k -Core theory to divide the chipped network. Then analyze 3-core network and 4-core network in k -Core on the basis of the concept of Group.

Research analysis

Keyword description analysis

Due to the process of disposing keywords, we came by 233,240 keywords after extracting keywords from 267,097 articles of oil and gas industry and data mining them. Along with observing the frequency of the keywords data, we have found out that the percentage of frequency beyond 100 is about 0.67 % within the 28 years from 1986 to 2014, among which the percentage of the keyword that only appear once is about 68.52 %, frequency

Fig. 1 The total annual theses in oil and gas industry from 1986 to 2014

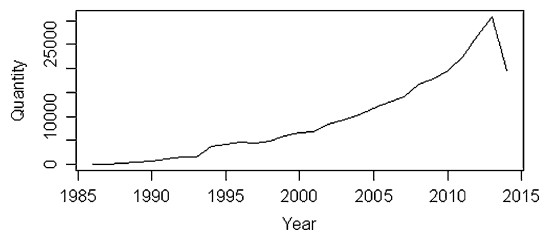
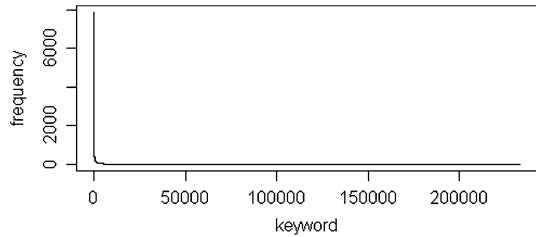


Fig. 2 Distribution curve of keyword frequency



that under 29 is about 98.69 %, you could see more data from Fig. 2. In order to explain keywords’ sequence better, we take the first 40 keywords as example to make Table 1.

From Fig. 2 we can see that keyword frequency in oil and gas industry is more distinctive, featuring of high concentration but small amount. It is quite different from what people used to think of Law 28, which has very little transition researches, and a great number of uncommon keywords. We can learn from the highest frequency keywords that petroleum industry researches have high concentration and great focus on specific problems, which boosts the development of application research by putting great emphasis on the operation of the research result. From Table 1 we can see that the keyword frequency faces sharp decline in oil and gas industry. In the past 28 years’ research, oil and gas industry are mainly focused on horizontal well, natural gas, well erosion, heavy oil, oil cracking, piping, crude oil, permeability, carbonatite, recovery rate, reservoir, fracturing, coalbed methane, well cementation, pay zone, crack in terms of petroleum techniques. To separate them in more specific aspects we could get four different directions: one is permeability, fracturing and cracking; the second is about oil and gas transportation such as prevent pipe from erosion; the third is drilling technique research such as horizontal well, well cementation; and the last aspect is rock traits as the characteristic of carbonatite. In the emphasis of optimizing the energy structure, we can know that natural gas, bio-diesel oil, coalbed methane, energy preservation are the heated issues; quantitative research such as numerical simulation, mathematics model accounts for large percentage in terms of research method; and evaluation and solution are popular in administration.

Keyword co-occurrence frequency analysis

Base on keyword analysis, we get some basic knowledge about keyword distribution in oil and gas industry, which laying a solid foundation of keyword co-occurrence frequency analysis. According to the method of matching the keyword, we pair the keyword by co-occurrence, and then data mining the paired keyword in order to get distribution diagram of keyword co-occurrence frequency. From the distribution diagram we can see that there must have some changes in comparison of single keyword frequency through matching the keyword and adding limited requirements. For example, single keyword high-frequency distributions are all basically up to 1000, and the high-frequency distribution of matched keywords are up to 100.

From Table 2 we can see that keyword co-occurrence pairs are regular, which can effectively display the key problems of the research and the practical problems that oil and gas industry are facing during the research and development process. We can see from Figs. 2 and 3 that the distribution curve of keyword pairing frequency and keyword frequency are both display as hyperbolic curve, similar to the distribution in the first quadrant of rectangular coordinate. From Fig. 3 we know that the keyword co-occurrence frequency

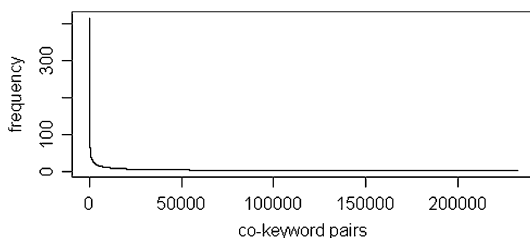
Table 1 The first 40 keywords in frequency

Keyword (1–10)	Frequency	Keyword (11–20)	Frequency	Keyword (21–30)	Frequency	Keyword (31–40)	Frequency
Application	7889	Heavy oil	2231	Measurement	1866	Petroleum enterprise	1658
Horizontal well	5341	Design	2180	Tarim basin	1862	Optimization	1653
Natural gas	4400	Catalytic cracking	2146	Prediction	1821	Technology	1644
The numerical simulation	3799	Catalyst	2136	Recovery rate	1819	Coalbed methane	1635
Analysis	3018	Energy saving	2107	Drilling	1759	Fracturing	1623
Corrosion	2801	Biodiesel	2107	Oil field	1754	Source rock	1616
Drilling fluid	2508	Affecting factors	2052	Permeability	1748	Well cementation	1521
Mathematical model	2328	Pipe	1942	Carbonatite	1738	Reservoir rock	1516
Research	2306	Crude oil	1870	Solution	1722	Technique	1515
Ordos basin	2290	Evaluation	1867	Pay zone	1720	Cracking	1489

Table 2 The first 20 keyword co-occurrence pairs in frequency

Keyword co-occurrence pairs (1–10)	Frequency	Keyword co-occurrence pairs (11–20)	Frequency
Biodiesel-transesterification	415	Biodiesel-ester exchange reaction	250
Research-application	406	Ordos basin-diagenesis	249
Porosity-permeability	377	Temperature-pressure	246
Extending group in Ordos basin	342	problem-measurement	245
Cementation-slurry	338	Horizontal well-numerical simulation	226
Catalytic cracking-catalyst	320	Oil-natural gas	224
Application-to develop	301	Tarim basin-carbonatite	217
Pumping unit-energy saving	258	Application-principle	208
Application-technology	256	Horizontal well-well track	205
Corrosion-protective	255	Application-structure	199

Fig. 3 Distribution diagram of keyword co-occurrence frequency



pairs are highly concentrated, and the researches go downward sharply when running away from the heated issue. The keyword-pair percentage that appeared 5 times is about 88.68 % within the past 28 years. As a result, oil and gas industry do have some significantly important research point, but the keywords with low frequency account for 90 % of the petroleum researches.

Keyword co-occurrence network analysis

We got the keyword co-occurrence matrix with the scale of $232,908 \times 232,908$ through analyzing the keyword frequency, and then draw the keyword co-occurrence symmetrical network diagram. There are 232,908 vertexes and 1,650,052 sidelines in the diagram. Among all the sidelines, 1,415,872 of them valued as 1, accounting for 85.81 % of the total network, and 234,180 of them valued bigger than 1. From the above data we can figure out the keyword co-occurrence network is a sparse network with the density of 0.00006084 and high dispersion. The relationship between different nodes is not very strong and imbalance distribution with the average degree of 14.17. Some of the nodes connected tightly but some scattered widely, but the average result is still quite high, which support the analysis of the core network.

In order to find out the heated issues of the network, we categorize the co-occurrence network on the base of the degree among the nodes. It has no difference between input degree and output degree as the network is nondirective, then we can get 995 networks

Table 3 The node network distribution on the base of different degree (the top ten)

Sub-network	Frequency	Frequency %	Cumulative frequency	Cumulative frequency %
1	1826	0.7840	1826	0.7840
2	28,251	12.1297	30,077	12.9137
3	52,938	22.7291	83,015	35.6428
4	47,711	20.4849	130,726	56.1277
5	24,690	10.6008	155,416	66.7285
6	13,808	5.9285	169,224	72.6570
7	9744	4.1836	178,968	76.8406
8	7017	3.0128	185,985	79.8534
9	4972	2.1347	190,957	81.9882
10	3813	1.6371	194,770	83.6253

with different degree. The basic distribution of co-occurrence network is shown in Table 3, from which we see that the percentage of the nodes lower than 10 is about 83.63 %; nodes lower than 29 is about 93.66 %. Therefore, the research in oil and gas industry has a trend of high concentration with large amount of keywords have a very low frequency. We can almost ignore the network of low degree as the researchers tend to focus on same objects. From Fig. 3 we can see that there are 1826 nodes with the degree of 1, accounting for 78.4 % of the total network. These nodes must be on the edge of the whole network diagram. 52,938 nodes have the degree of 3, accounting for 22.73 % of the entire network. From the above data we know that different research issue have total distinctive characteristic. As a result, in order to improve the core network, we choose the first 100 nodes as example to analyze its features, and the degree of the bottom 100 nodes is about 1611, that is to say, it has 1611 related keywords.

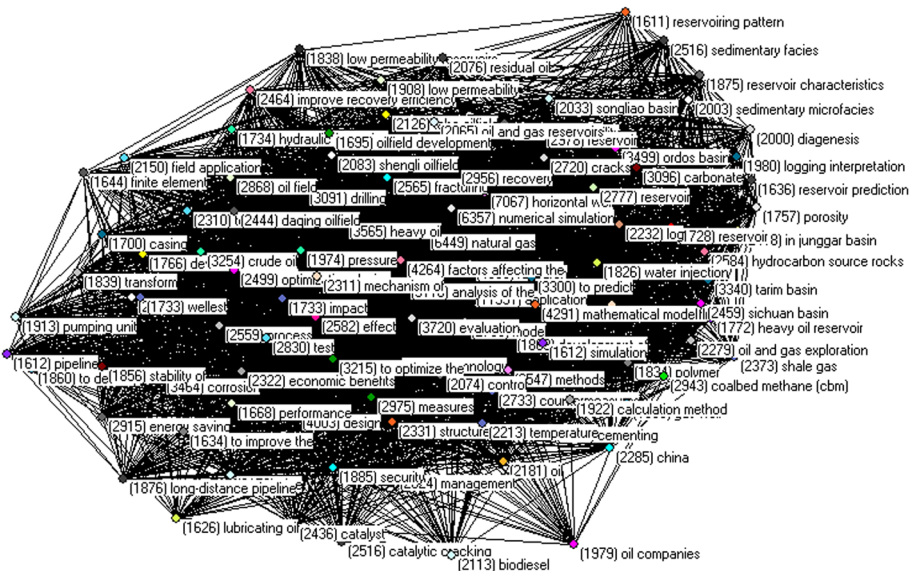


Fig. 4 The classification of keyword co-occurrence network on the base of degree (top 100)

By the analysis of core degree network, we can get 104 nodes, see Fig. 4. From Fig. 4 you can see that the high complex keywords co-occurrence network have clearly shown us the heated issues and core researches in oil and gas industry, which provide a rough view for further research. And the relationship between different keywords is very complicated, as from the perspective of one parameter in the network we know that at present it is about 674 sidelines value as 1, 3187 sidelines valued larger than 1, the density of the network is about 0.72, and the average degree of the network is about 74.25. The nodes are closely connected with each other.

The concept of Degree in nondirective network diagram means the amount of nodes' neighbors. In order to seek the core node, we simplify the network in this high density network. Delete the sidelines of value lower than 56, which means the concept should form solid connection with other nodes rather than occasional connection. And then re-organize the network according to the concept of *k*-Core, we can get the result of 5 nodes when *k* = 4, and 27 nodes when *k* = 3. These 32 core concept are the key issues in oil and gas industry. The reason why we put up 3 and 4 is because *k* = 3 is the basic boundary of dividing group.

Figure 5 shows the 3-Core degree network in oil and gas industry, from which we know that the reservoir rock in Tarim Basin, Ordos Basin and Sichuan Basin are the key research spots. Numerical simulation mainly uses to develop fracturing technique and improve recovery rate, and the natural gas exploiting technique and evaluation in Sichuan Basin and Ordos Basin is also a special feature. Prevent oil pipe from erosion is a problem that bother the researcher in oil and gas industry for a long time, so it became a heated issue. Secondly, some resources which need some extra efforts to recover like remaining oil and heavy oil are also heated issues, therefore, the whole industry put lot of attention to increasing recovery rate. From Fig. 5 we know that universal keyword is the medium to connect different node, such as numerical simulation, mathematical model, evaluation and analysis. These keywords hold different core concepts together. Moreover, regions, integrated with other core keywords, work as a crucial element to different key researches, and thus benefit the development in regional oil and gas resources.

The 4-Core degree network in oil and gas industry is shown in Fig. 6. We can see from Fig. 6 that the most connected core issues in petroleum industry are drilling, well cementation, drilling fluid, horizontal well, and technology applications. Moreover,

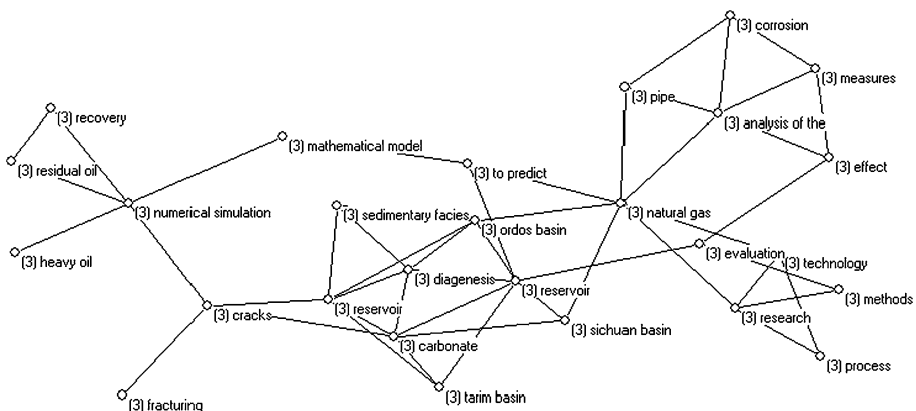


Fig. 5 The 3-Core degree network in oil and gas industry

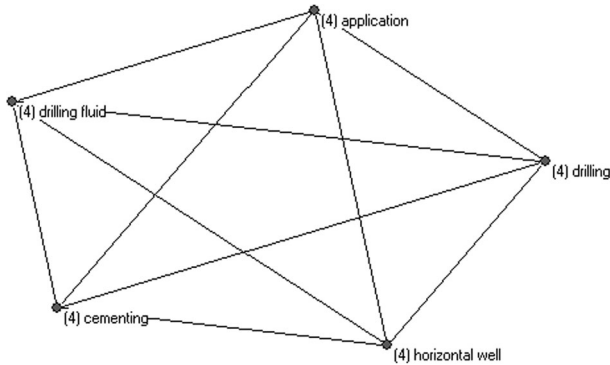


Fig. 6 The 4-Core degree network in oil and gas industry

through the development process of the industry, we find out that most of the researches are relatively suitable for the actual need of the industry. Along with the exploration of the oil and gas resources, the detected reservoir layers are hidden in a much more deeper area than before, which asking for new requirements such as advanced techniques and good application in drilling, well cementation and drilling fluid. The horizontal well drilling is showing up in recent few decades, with the features of large contact area to the oil zone. It effectively boosts the yield of the oil and gas resource and as a result, increases the benefits. All these push horizontal well drilling ranking one of the fifths basic supporting concepts in petroleum industry.

Conclusion

This article takes the keywords in theses data from 1986 to 2014 in oil and gas industry as material, using the theory of Big Data to data mining the theses. Analyze the keyword frequency and keyword co-occurrence pair frequency in oil and gas industry, and then map out the co-occurrence network of the entire oil and gas industry. The conclusions are as followed:

1. The researches in oil and gas industry are focused on practical usage. We can see from the sequence of keywords analysis that “application” ranks the first among the frequency of all the keywords. And it also plays a key node in 4-Core degree network, which means all the researches are aim at the oil services and further exploitation in oil and gas resource. And you can see from the 3-Core degree network that the researches which base on different technologies are all conducting in the Major Basin in China.
2. The researches in oil and gas industry are highly-concentrated in key issues, scattered in some unpopular issue and almost without transition keyword. You can see from the distribution curve of the keyword frequency in oil and gas industry that the curve dropped sharply without stop when the issue was not heated. And the diagram of the keyword co-occurrence pair frequency curve looks similar. These curves show that the researches in petroleum industry are emphasized in some key fields. Secondly, the heated issues are also highly concentrated, which means the whole industry are facing similar problems that need further efforts from the researchers. The curves in both keyword frequency and keyword co-occurrence pair frequency are featured of long-

tail, which means researchers may have tried numerous ways to solve the key problems.

3. The core keywords in oil and gas industry researches are drilling, well cementation, drilling fluid, horizontal well and application. And the main places for applying the latest research results are Ordos Basin, Tarim Basin and Sichuan Basin. The development of heavy oil and remaining oil, cracks, natural gas development, prevent pipes from corrosion and rock-forming mechanism are also the major heated issues in oil and gas industry.

References

- Bertschi, S. Bresciani, S. Crawford, T. Goebel, R. Kienreich, W., Lindner, M. & et al. (2011). What is knowledge visualization? perspectives on an emerging discipline. In *Information Visualisation (IV), 2011 15th international conference* (pp. 329–336). IEEE.
- Biloslavo, R., Kregar, T. B., & Gorela, K. (2012). Using visualization for strategic decision making: A case of slovenian entrepreneurs. In *13th European conference on knowledge management* (p. 83). Academic Conferences Limited.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.
- Cobo, M. J., López Herrera, A. G., Herrera Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology, 62*(7), 1382–1402.
- Eppler, M. J. & Pfister, R. (2013). Best of both worlds: Hybrid knowledge visualization in police crime fighting and military operations. In *Proceedings of the 13th international conference on knowledge management and knowledge technologies* (p. 17). ACM.
- Hong, Y., & Haiyue, W. (2012). The frontier analysis of earnings management based on mapping knowledge domains. *Management Review, 6*, 19.
- Hua, M., Gao, Y., & Li, Y. (2013). Mapping knowledge domains of chinese medical equipment journal. *Chinese Medical Equipment Journal, 1*, 38.
- LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 21*, 52–68.
- Lohr, S. (2012). *The age of big data*. New York Times, p. 11.
- Marx, V. (2013). Biology: The big challenges of big data. *Nature, 498*(7453), 255–260.
- Mayer-Schönberger, V., Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.
- Osinska, V., & Bala, P. (2014). Study of dynamics of structured knowledge: Qualitative analysis of different mapping approaches. *Journal of Information Science, 2000*135577.
- Petra, K., Miroslav, B., & Tomislav, F. (2012). Knowledge visualization in biometric face recognition on two-dimensional images information technology interfaces (ITI). In *Proceedings of the ITI 2012 34th international conference* (pp. 349–354). IEEE.
- Pollack, J., Adler, D., & Sankaran, S. (2014). Mapping the field of complexity theory: A computational approach to understanding changes in the field. *Emergence: Complexity & Organization, 16*(2), 74–92.
- Somekh, J., Choder, M., & Dori, D. (2012). Conceptual model-based systems biology: Mapping knowledge and discovering gaps in the mRNA transcription cycle. *PLoS One, 7*(12), e51430.
- Wang, M., & Jacobson, M. J. (2011). Guest editorial-knowledge visualization for learning and knowledge management. *Educational Technology & Society, 14*(3), 1–3.
- Wang, M., Peng, J., Cheng, B., Zhou, H., & Liu, J. (2011). Knowledge visualization for self-regulated learning. *Educational Technology & Society, 14*(3), 28–42.
- Wei, S. Y. C. (2013). Research development of grid service in China—Bibliometric and mapping knowledge domains analysis based on CNKI from 2003 to 2012. *Journal of Modern Information, 7*, 26.
- Womack, R. (2014). Data visualization and information literacy. *International Association for Social Science Information Service and Technology, 12*–17.
- Xiuling, Y. H. X. (2012). Mapping knowledge domains analysis on information resources management based on web of science. *Journal of Intelligence, 12*, 12.

- Yue, Z. (2014). Mapping knowledge domains analysis of research hotspots and front in China's digital archives. *Archives & Construction*, 6, 6.
- Zhao, L., & Zhang, Q. (2011). Mapping knowledge domains of Chinese digital library research output, 1994–2010. *Scientometrics*, 89(1), 51–87.
- Zheng, Y., Hu, C., & Ma, Y. (2013). The visualized mapping knowledge domains of the research on Chinese government information disclosure. *Advances in Asian Social Science*, 4(2), 836–843.
- Zhichao, Z. (2012). Social network analysis of high cited authors based on domestic mapping knowledge domains. *Journal of Modern Information*, 32(8), 97–100.
- Zins, C., & Santos, P. L. (2011). Mapping the knowledge covered by library classification systems. *Journal of the American Society for Information Science and Technology*, 62(5), 877–901.