Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 814 – 821

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Data, DIKW, Big data and Data science

Gu Jifa[a]*,Zhang Lingling[b]

[a]*cademy of Mathematics and Systems science, Chinese Academy of Sciences,55 Zhongguancun East Road, Beijing 100190, China*
[b]*chool of Management, University of Chinese Academy of Sciences， Beijing, 100190, China*

Abstract

In this paper we discuss the relationship between data and DIKW, that the data only evolves to knowledge, which may have some value, but if without the wisdom we still could let the knowledge be really useful to people. Now the big data occupies much attention in some extent for his volume, velocity, and variety. But in practical use the value plays more important role. Finally to judge the value for data not necessary for big, in some cases the small data also may lead to big value. So we appreciate the data science, which may consider more inherent value from data.

Keywords: Data; DIKW; Big data; small data; Data science

1. Data and hierarchy of DIKW

In the previous paper presented in ITQM2013 we had mentioned the hierarchy of DIKW that is data, information, knowledge, wisdom: a four layer hierarchy, where each layer adds certain attributes over and above the previous one (see Fig. 1). Data is the most basic level; Information adds context; Knowledge adds how to use it; Wisdom adds when and why to use it [1, 2]. A further elaboration of Ackoff's definitions follows: Data is raw. It can exist in any form, usable or not. It does not have meaning of itself. Information is data that has been given meaning by way of relational connection. This "meaning" can be useful, but does not have to be. Knowledge is the appropriate collection of information, such that its intent is to be useful. Knowledge is a deterministic process. This knowledge has useful meaning to them, but it does not provide for, in and of itself, integration such as would infer further knowledge. Understanding is an interpolative and probabilistic process. It is cognitive and analytical. It is the

* Gu Jifa. Tel:+0-0086-10-59358721; fax: +0-0086-10- 62568364 .
E-mail address:jfgu@amss.ac.cn.

doi:10.1016/j.procs.2014.05.332

process by which people can take knowledge and synthesize new knowledge from the previously held knowledge. Wisdom is an extrapolative and non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.) [3]. In the meaning of hierarchy of DIKW the data just a basic fact and raw that will be useful only let them evolve to the information, knowledge and wisdom. In the previous two levels of data and information we may use a lot of methods such as data mining, text mining, web mining and tools such as data base, data warehouse and management information system to let them to be useful [2]. In order to move forward to the third level of knowledge we should use KDD, knowledge engineering and management, intelligent knowledge [4]. Given large-scale databases, intelligent knowledge management enables to generate "special" knowledge, called intelligent knowledge base on the hidden patterns created by data mining. The process of intelligent knowledge management - a new proposition from original data, rough knowledge, intelligent knowledge, and actionable knowledge is proposed [5]. Especially in considering the expert experiences to let the knowledge extracted by some methods like data mining to be more useful, we use so-called domain driven data mining[6,7]. For the sake of fully utilize the expert experience we propose expert mining, by which people may use properly the data, information and knowledge [8, 9, 10].

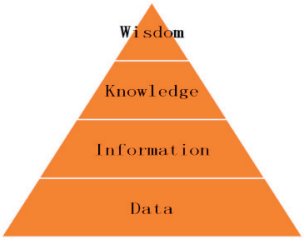

Fig. 1 D-I-K-W hierarcy

2. Big data

2.1. A short history of big data

The story of how data became big starts many years before the current tide around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion". The following are the several phases in the history of sizing data volumes plus other "firsts" in the evolution of the idea of "big data"; we adopt most information from G. Press with minor modifications [11].

1944 F. Rider estimates that American university libraries were doubling in size every sixteen years. Given this growth rate, Rider speculated that the Yale Library in 2040 will have "approximately 0.2 billion volumes (If we mention the number of bites for one book, it occupies around 10MB, then 1G=109B (equal to 100 books), 0.2 billion volumes equal to $0.2\times 10^9\times 10\text{MB}=2\times 10^{15}\text{B}=2\text{PB}$).

1997 M. Cox and D. Ellsworth in their article mentions that "data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources." It is the first article to use the term "big data."

1997 M. Lesk publishes "How much information is there in the world?" he concludes that "There may be a few thousand petabytes (PB) of information."

2000 P. Lyman and H. R. Varian publish "How Much Information?" It is the first comprehensive study to quantify, in computer storage terms, the total amount of new and original information (not counting copies) created in the world annually and stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic. The study finds that in 1999, the world produced about 1.5 exabytes (equal =1500 PB) of unique information, A similar study conducted in 2003 by the same researchers found that the world produced about 5 exabytes of new information in 2002.

2001 D. Laney, publishes a research note titled "3D Data Management: Controlling Data Volume, Velocity, and Variety." A decade later, the "3Vs" have become the generally-accepted three defining dimensions of big data, although the term itself does not appear in Laney's note.

2007 J. F. Gantz, D. Reinsel and other researchers make estimation and forecast the amount of digital data in 2006, the world created 161 exabytes of data and forecasts that between 2006 and 2010, the information added annually to the digital universe will increase more than six fold to 988 exabytes, or doubling every 18 months. According to the 2010 and 2012 releases of the same study, the amount of digital data created annually surpassed this forecast, reaching 1227 exabytes in 2010, and growing to 2837 exabytes in 2012.

2008 Bret Swanson and George Gilder estimates that U.S. IP traffic could reach one zettabyte by 2015 and that the U.S. internet of 2015 will be at least 50 times larger than it was in 2006.

2008 Cisco predicted that "IP traffic will nearly double every two years through 2012" and that it will reach half a zettabyte in 2012. The forecast held well, as Cisco's latest report (May 30, 2012) estimates IP traffic in 2012 at just over half a zettabyte."

2009 R. E. Bohn and J. E. Short publish "How Much Information? 2009 Report on American Consumers." The study finds that in 2008, "Americans consumed information for about 1.3 trillion hours, an average of almost 12 hours per day. Consumption totaled 3.6 Zettabytes and 10,845 trillion words, corresponding to 100,500 words and 34 gigabytes for an average person on an average day." Bohn, Short, and C. Baru follow this up in January 2011 with "How Much Information? 2010 Report on Enterprise Server Information," in which they estimate that in 2008, "9.57 Zettabytes of information."

2012 D. Boyd and K. Crawford publish "Critical Questions for Big Data" . They define big data as "a cultural, technological, and scholarly phenomenon that rests on the interplay of: (1) Technology: maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets. (2) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims. (3) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy." [11]

Author find that for the volume of big data, its magnitude are PB, EB, even ZB, and for the specific features now in big data there appeared 4V, the first three Volume, Velocity, and Variety are the same in 2001, but the last V some says it is Veracity, most mentions Value, probably they will add some more features, see following paragraph.

2.2. Some Comments on big data

With the rapid development in studying and using the big data we just want put some comments.

1).We should pay much attention on the complexity of the structures and contents behind of big data, as Qian X.S. et al mentioned in 1990 we had confronted difficult problems happened in so called open complex giant (OCG) systems, which originated in social system, economical system, geographical system and human body system (around ten years later Qian's colleagues Dai pointed out that the internet is also an open giant complex systems). the feature of "giant" just corresponded to the big data because of large number of elements, data and information in OCG systems, but the feature of "complex" in some extent wasn't too much been emphasized in the professionals in the fields of technical cycle, they pay too much attention to the technical treatment on big data by using advanced computer and information technology [12].

2). According to Li's opinion in 2012 we had made the small data to the big data in recent years, but now we should transfer the big data to small data discarding the dross and select the essential [13]. Qian had mentioned similar views even in 1992, that now during the time of information revolution, the quantity of information like the sea will be surging over us, as a human in 21 century you must be not drown. It means you must have a high intellectual level to manage it; otherwise you will be drowning [14]. As matter of fact we had suffered the information overload even in the beginning of 21 century [15-17].

3). In the first paragraph we had mentioned that as the data it should evolve into D-I-K-W, without transferring the data to knowledge, especially useful knowledge, it just nothing. Like a people have a lot of bricks, cement, wood and steel if you don't have help from a good designer you couldn't transfer them into a real construction. Then we also mention between the knowledge and wisdom there exist understanding. It is the process by which people can take knowledge and synthesize new knowledge from the previously held knowledge. So also if you have big data,

you should also try to let other people to understand, to utilize it, this is also the view by Kong Y.H chief manager from Teradata [18]

4). From the point of views of knowledge science the data usually belongs to explicit knowledge, but we even pay much attention to the tacit knowledge, for that purpose Nonaka proposed so called SECI model in helping people to transfer the tacit knowledge into explicit knowledge. Then as compared with the individual knowledge we should pay attention the organizational knowledge and collective knowledge, it means sometime the same data in front of different individual there appear different understanding, but after discussion, exchange and analysis a group of experts will generate more correct and useful knowledge [19]. Even more Nonaka emphasizes the practical wisdom (phronesis), that we must pay attention the morality, it means the knowledge will be useful is closely related the goodness for the most people [20]. It just point out the Value, the forth V for big data.

5). About the value of big data we wish illustrate two examples related to such kind of problems: how can we use the big data. The first example is related to the project of ADVISE (Analysis, Dissemination,Visualization,Insight and Semantic Enhancement）system, the ADVISE system run by DHS of USA and supported by a large group of experts and researchers in research institutes and universities in US in order to collect and analyze enough big data starting from 2004, they use various line for collecting the data and information from the internet, telephone, border custom and sensors from various detecting equipments etc., which will be useful to detecting the terrorists information. We especially appreciate them by using the complex network and semantic analysis (sometimes they also used the data mining instead) and try to let these big data useful to identify the terrorists, it is beneficial to government and staffs who are working in DHS. Especially in this project they proposed the idea of data science. But with the rapid growth of number of data and information collected from terrorists and other common people by this system, because of the reason of violating privacy of common people the US congress finally prohibited this system to work anymore in the August of 2007. But as matter of fact that was just a surface phenomena for US government, they had to force to stop ADVISE, but has found the another similar projects to collect all information which they assume will be useful to US [21]. The second example is the Snowden event happened in last year. It is very interesting to exhibit “PRISM” started from 2007 by Snowden in 2013 that US not only monitor and collect the data and information from internet and mobile phone in their own country but also from the other countries, even including their friend countries, like UK, Germany etc.. To this event US government assumes that Snowden is a traitor, but some other countries assume that he is a hero. So the same data handled by different people will have different value

3. Data science

The term “Data Science” has emerged only recently to specifically designate a new profession that is expected to make sense of the vast stores of big data. But making sense of data has a long history. The following timeline traces the evolution of the term “Data Science” and its use, attempts to define it, and related terms.

3.1. A short history of data science

G. Press has made a good and comprehensive description about the short story of data science; here author just wishes select part of them as a more short history of data science [22].

1962 J. W. Tukey writes in “The Future of Data Analysis”. In 1977, Tukey published Exploratory Data Analysis, arguing that more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis “can—and should—proceed side by side.”

1974 P. Naur says “Datalogy, the science of data and of data processes and its place in education, “and “the term ‘data science’ has been used freely.” Naur offers the following definition of data science: “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”

1977 The International Association for Statistical Computing (IASC) ‘s mission is to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

1989 G. Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in Databases (KDD) workshop.

1996 Members of the International Federation of Classification Societies (IFCS) for the first time, the term "data science" is included in the title of the conference.

1996 U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth publish "From Data Mining to Knowledge Discovery in Databases." In our view, KDD [Knowledge Discovery in Databases] refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data… the additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data.

1999 J. Zahavi is quoted in "Mining Data for Nuggets of Knowledge": "Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Scalability is a huge issue in data mining.

2001 W. S. Cleveland publishes "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." It is a plan "to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'" Cleveland puts the proposed new discipline in the context of computer science and the contemporary work in data mining.

2002 Data Science Journal was launched by the Committee on Data for Science and Technology (CODATA) of the International Council for Science (ICSU).

2003 Launch of Journal of Data Science: "By 'Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling… yet the most important part is its applications–all sorts of applications

2005 T. H. Davenport, D. Cohen, and Al Jacobson publish "Competing on Analytics," describing "the emergence of a new form of competition based on the extensive use of analytics, data, and fact-based decision making… Instead of competing on traditional factors, companies are beginning to employ statistical and quantitative analysis and predictive modeling as primary elements of competition. "

2005 The National Science Board publishes "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century."

2007 The Research Center for Dataology and Data Science is established at Fudan University, Shanghai, China. In 2009, Y.Y. Zhu and Y. Xiong, publish "Introduction to Dataology and Data Science," in which they state "Different from natural science and social science, Dataology and Data Science takes data in cyberspace as its research object. It is a new science." The center holds annual symposiums on Dataology and Data Science.

2008 The JISC publishes the final report of a study it commissioned to "examine and make recommendations on the role and career development of data scientists and the associated supply of specialist data curation skills to the research community."

2009 Harnessing the Power of Digital Data for Science and Society is published. Many disciplines are seeing the emergence of a new type of data science and management expert, accomplished in the computer, information, and data sciences arenas and in another domain science.

2009 K. D. Borne and other astrophysicists submit a paper titled "The Revolution in Astronomy Education: Data Science for the Masses"

2010 M. Loukides writes in "What is Data Science?" They are inherently interdisciplinary. They can tackle all aspects of a problem, from initial data collection and data conditioning to drawing conclusions. They can think outside the box to come up with new ways to view the problem, or to work with very broadly defined problems: 'here's a lot of data, what can you make from it?'"

2010 H. Mason and C. Wiggins write in "A Taxonomy of Data Science": "…what a data scientist does, in roughly chronological order: Obtain, Scrub, Explore, Model, and iNterpret…Data science is clearly a blend of the hackers' arts… statistics and machine learning… and the expertise in mathematics and the domain of the data for the analysis to be interpretable… It requires creative decisions and open-mindedness in a scientific context."

2011 P. Warden writes in "Why the term 'data science' is flawed but useful": "There is no widely accepted boundary for what's inside and outside of data science's scope. people tend to work beyond the narrow specialties that dominate the corporate and institutional world, handling everything from finding the data, processing it at scale, visualizing it and writing it up as a story. They also seem to start by looking at what the data can tell them, and then

picking interesting threads to follow, rather than the traditional scientist's approach of choosing the problem first and then finding data to shed light on it."

2011 D. Smith writes in "'Data Science': What's in a name?" I think 'Data Science' better describes what we actually do: a combination of computer hacking, data analysis, and problem solving."

2011 M. J. Graham talks about "The Art of Data Science". He says: "To flourish in the new data-intensive environment of 21st century science, we need to evolve new skills… We need to understand what rules [data] obeys how it is symbolized and communicated and what its relationship to physical space and time is."

2011 H. Harris writes in "Data Science, Moore's Law, and Moneyball": "'Data Science' is defined as what 'Data Scientists' does. What Data Scientists do has been very well covered, and it runs the gamut from data collection and mugging, through application of statistics and machine learning and related techniques, to interpretation, communication, and visualization of the results. "

2012 T. Davenport and D.J. Patil publish "Data Scientist: The Sexiest Job of the 21st Century" in the Harvard Business Review. [22]

3.2. Some Comments on data science

In China there is a series of science conference named Xiangshan science conference, which is the general designation of a small-scale academic workshop series, is supported by some very important organizations related to the science and technology, such as Ministry of Science and Technology, Chinese Academy of Sciences, NSFC, and Ministry of Education etc. The topics selected by scientists according to its interdisciplinary cooperation and integrated studies in various areas of excellence and exploring new frontiers. And in the 278 session contributed to "Data technology and knowledge economy" (May 22-24, 2006), Gu had attended this session and express two opinions based on his own practices: 1) from data mining to expert mining and 2) from data technology to data science, by the way in that conference Zhu Y.Y. also mentioned the dataology similar with data science [23, 24]. Based on these two opinions we wish make two comments.

1). From 2005 Gu had engaged in two projects, first project related to the social system problem, in this project we have to deal with the huge amount of data and information, so at first like our US colleagues in ADVISE project we pay attention to data mining very much, since in that time in many presses in US they just related the ADVISE project as an application of data mining. In 2006 we read the document on "Data sciences information technology for homeland security information management and knowledge discovery" written by Sandia National Laboratory and Lawrence Livermore in January of 2005. Even in that time Gu also appreciate the terminology of data science very much. For Gu it means that they didn't want stop in the data mining, instead of use the data science, since they also found that only stop on the dealing with data is quite not enough, instead, they pay much attention to the knowledge management. In that time Gu had applied the Meta-synthesis system approach to solve the social problem. This approach not only pays attention to utilize the data, information, but also the experiences and wisdom from experts and decision makers. So Gu had published the paper related so call expert mining, which pays attention on expert experience, judgment, understanding and making decision [8]

2). In second project related to the Traditional Chinese Medicine (TCM), we wish inherit the experiences and idea from some veteran and famous TCM doctors in China. In this project we may use the number of data and information, just a few, from statistics point that is a small sample problem. Again we use the expert mining to solve the problem [9]. Here really we just wish mention that we met not a big data problem, but a small data. But we think the data science may be will resolve also these problem, if we set the data science, including both big data and small data, only we should use different methods and techniques, e.g. We should introduce the intelligent knowledge management [5], domain driven data mining [6, 7] and expert mining [25] etc.

In May of 2013 there is another Xiangshan science conference, 462 session, titled in "Data science and big data" chaired by Shi Y from the research center on Fictitious Economy and Data Science, Chinese Academy of Sciences and Zhu Y.Y from Research Center for Dataology and Data Science and other two professors P. S. Yu and Li J.P.. In that conference Chinese scholars just wish to perfect the definition, contents, methods and applications for Data science. After this conference many speakers have published series of papers in the Journal of Science and Technology for Development in the beginning of 2014, Gu also publish his paper in this journal related to the issues on big data [17].

3) The data not necessary comes from real world, now with the development of computer technology we may create a lot of data from computer. Now the numerical simulation may generate enough data as you required in your problem. Especially now when we deal with the social problems, since it is hard to collect the real data, sometime it involves the privacy of common people, sometime data also relates to the money, you have to pay a lot of money to buy some real useful data and information, so we often have to use the data generated from computer.

4. Conclusion

Author just wish explain some relationships within the data, D-I-K-W, big data and data science. From one side we want emphasize the utilization of data that must evolve to knowledge and wisdom, from another side we wish to point out that too much emphasizing the data "big", sometime may not solve all problems. In some problems we really have only possibility to collect just a few data and information. So we wish to develop the data science as soon as quickly to cover and solve much more issues which we had met already. Certainly to establish an entire theory for data science requires us all to make more effort and deep consideration.

Acknowledgements

We wish to acknowledge Shi Y. and his colleagues from the research center on Fictitious Economy and Data Science, Chinese Academy of Sciences for their kind support, freely exchange and discussion of views on data mining, expert mining, big data and data science.

References

1. Wikipedia, DIKW, http://en.wikipedia.org/wiki/DIKW 15 December 2011
2. Gu Jifa, Data, Information, Knowledge, wisdom and meta-synthesis of wisdom-comment on wisdom global and wisdom cities,**,** *Procedia Computer Science*, Volume 17,2013: 713 – 719, First International Conference on Information Technology and Quantitative Management (ITQM2013), Beijing, China
3. Bellinger G., Castro D., Mills A., Data, Information, Knowledge and Wisdom, http://www.Systems-thinking. org/dikw/dikw.htm, 2004.
4. Leondes C.T., Intelligent Knowledge-based Systems: Business and Technology in the New Millennium, Springer, 2010
5.Zhang L.L., Li J., Li A.H., Zhang P., Nie G.L., Shi Y., A New Research Field: Intelligent Knowledge Management ,International Conference on Business Intelligence and Financial Engineering, 2009 (BIFE '09), Beijing,24-26 July 2009 , 450 - 454
6. Cao L.B Zhang C.Q., Domain-Driven Actionable Knowledge Discovery in the Real World, PAKDD 2006, LNAI 3918, 2006:821-830
7. Zhu Z.X.Toward Domain-Driven Data Mining, International Symposium on Intelligent Information Technology Application Workshops, 2008, IITAW '08. 21-22 Dec. 2008, Shanghai, 44 - 48,
8. Gu J.F., Expert mining for discussing the social complex problems, MCS2006, Beijing, September 2006
9.Gu J.F, Song W.Q., Zhu Z.X, Gao R. and Liu Y.J., Expert mining and TCM knowledge， KSS2008， Guangzhou， December11-13， 2008
10. Gu J.F., Liu Y.J., Zhu Z.X. Expert mining and Meta-synthesis, Science Press, Beijing, 2014(in Chinese)
11. Gil Press, A very short history of big data, http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/ Jun 6, 2012
12.Qian X.S., Yu J.Y. and Dai R.W., A new scientific discipline-Open complex giant systems and its methodology, Journal of Nature， 13(1), 1990：3-10, (in Chinese)
13. Gan X.， Academician Li G. J.: big data turns into new interesting points in IT, ， China Science Daily，2012-6-27
14. Qian X.S., 21 century is the century for the competition of information, we should grasp now the construction of information network, In Jiang L. editor. "*Ian X.S., on Systems science (collections of talks)*", Beijing, Science Press, 2011, p.101-107
15. Shi Y., Gu J.F., Information Overload, in the IEBM Handbook of Information Technology in Business (M. Zeleny ed.), Business Press, Thomson Learning, 2000, 554-560.
16. Shi Y., (2013).Three important technical issues on big data, IT Expert net, 2013-08-25， http://m.ctocio.com.cn/storage/359/12698859_m.shtml. (in Chinese)
17. Gu J.F., (2014), Some issues on big data, *Science &Technology for Development*, 10(1), 2014:20-24(in Chinese)
18. Huo N., Teradata: Big data should go to easily to use and to be popular, *Zhongguojisuanjibao*, 2013, Vol.17 p.24 http://www.zcom.com/article/107572/
19. Nonaka I. Takeuchi H., The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, 1995
20. Nonaka I., Toyama R., Hirata T., et al., Managing Flow: A process theory of the knowledge-based firm, Palgrave, Macmillan, New York, 2008.
21. Joint report by Sandia National Laboratory and Lawrence Livermore, Data sciences information technology for homeland security information management and knowledge discovery, Sandia report SAND2004-6648, LINL report UCRL-TR-208926, January 2005
22. Gil, Press, A very short history of data science, http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/2/
23. Zhu Y.Y., Xiong Y., Dataoogy and data science, Shanghai, Fudan University Press, 2009

24. Zhu Y.Y., Xiong Y., Dataology and Data Science：Up to Now, http://www.paper.edu.cn/en_releasepaper/content/4432156, 2011-6-16
25. Gu J.F.,Expert Mining: Realization of Knowledge Synthesis, *Journal of Engineering Studies* , 2(2), 2010: 100-107, (in Chinese)