International Journal of Database Theory and Application
Vol.9, No.10 (2016), pp.189-198
http://dx.doi.org/10.14257/ijdta.2016.9.10.16

The Framework of Social Networks Big Data Processing Based on Cloud Computing

Liu Kewen[1,2] and Gao Changyuan[1]

[1]*Harbin University of Science and Technology, School of Management, Harbin, China*
[2]*Harbin University of Commerce, School of Computer and Information Engineering, Harbin, China*
avenn@163.com

Abstract

With the rise of cloud computing, internet of things, social networks, the type and scale of data in human society has increased at an unprecedented rate, making data from being a simple object to be process to being a basic resource. Fully mining the value of data resources that was hidden in SNS such Weibo Microblog, Wechat has become a common subject concerned by industrial circle, academic area and government departments. Although the distributed storage and analysis of cloud computing platform have been widely used in big data process of social networks, it has not been able to fully solve the problems of big data storage and process in social networks. In this paper, it proposed the big data process framework of social networks based on cloud computing. By adopting the mixing cloud model and coordinating the data storage framework and data computing framework, and regarding social networks features such as real-time, sharing, mobility, individuation, and interactivity, this big data process framework can be adopted to process large-scale massive amount of data, to research the unified management and sharing strategy of massive data, to propose data process strategy and the service application of big data such as Microblog and Wechat, and to discuss several urgent key problems in processing social networks big data.

Keywords*: Cloud computing, social networks, big data, data processing*

1. Introduction

In recent years, social networks has become one of the key subjects of big data research. Social networks services such as Twitter, LinkedIn, Facebook, Microblog from Sina, QQ and Wechat from Tecent gather large amount of participants, who contribute massive amount of content normally called UGC (User Generated Content). Big data produced by such social networks services can be divided into two classes: content data and linked data. Content data include text, figures, and multimedia data; while liked data mainly refer to the correlations between topological graph and entity of the social networks [1].

Through analyzing the massive data in social networks, it can find new business opportunities and bring huge commercial value. Deng *et al.* [2] proposed advertising recommendation system, which integrates major big data process techniques such as Mongo DB, Zarkov MapReduce framework, Mahout machine learning, artificial intelligence, and data mining, and can be used to find out consumers' behavioral model, establish advertising recommendation model and conduct trend forecast through analyzing big data of social networks such as Twitter, Yelp, Facebook.

The scale effect of social networks big data brings huge challenges to data storage, data management and data analysis. The problem in real-time data processing is the first one,

ISSN: 2005-4270 IJDTA
Copyright ⓒ 2016 SERSC

for instance. Since Hadoop-based distributed storage system cannot meet the demand for low delay, thus the real-time data cannot be obtained. Mishne *et al.* [3] proposed a customized memory processing engine, which combines the Hadoop-based analysis platform and data stream processing engine, so as to realize the conception of real-time processing of massive data. Secondly, to solve the problems on complexity and redundancy of big data processing, Choo *et al.* [4] proposed a visualized data processing method. Cloud computing is the supporting technology for data processing as well as a basic platform for big data analytical application. Conventional data management and analysis is a relational database management system (RDBMS), which cannot handle the scale and heterogeneity of big data, while cloud computing platform can serve as a basic facility of big data system and meet specific requirements on cost effectiveness, flexibility, scale-up or scale-down capability.

Currently, big data process is mainly conducted based on DFS (distributed file system) and NoSQL database [5-6], which is suitable for massive data processing with characteristics of persistent storage and scheme free. MapReduce programming framework can achieve great success in processing group-aggregation task like website ranking [7], while Hadoop, which can provide systematical solutions by integrating data storage, data processing, system management and other modules, has become the major technology for big data process [8]. Based on these technologies and platforms, a flexible big data application can be established, so as to meet the dynamic requirements and intelligent management of big data processing in various fields. Simmhan *et al.* [9] proposed an expandable software platform. With the application program for dynamic response requirement, this platform can help to realize the intelligent management, which can release the peak load of smart grid as well as provide the adaptive information integration technology.

Although some research findings have been achieved, data processing problems associated with massive data in social networks such as uniform management, security sharing, effective storage are still remained unsolved. To obtain valuable information from social networks big data, high efficiency data processing should be conducted. In this paper, regarding the massive social networks data, a combined method of distributed cloud computing and mixed cloud safe storage technology was adopted for data analysis and data mining, so as to propose the processing framework of social networks big data based on cloud computing with the purpose of processing big data from social networks, realizing key functions such as high efficiency storage and security access in processing social networks big data.

2. Cloud Computing and Social Networks Big Data Processing

2.1. Application of Cloud Computing in Processing Social Networks Big Data

Cloud computing is a newly developed computing mode in recent years, which reflects an information service pattern with features of resource dynamic allocation, on-demand computing and dynamic information flow. Key technologies involved are: distributed programming and calculation, data storage technology, virtualization technology, and data management technology.

Due to the effective data processing ability, cloud computing applied in social networks can make data processing faster and more convenient, providing new possibility for the emergence of new application products and the better user experience brought by the product. Through mining the data in social networks, we can quickly know current hot topics, control emergencies, guide public opinion direction, analyze user behavior, and find new commercial mode.

Specific applications involved in social networks big data processing:

- Friend recommendation feature on Microblog or Wechat. Help users to find the people they are interested in by relation prediction method.

- Public opinion analysis which mainly involves identification technologies of computer text information such as text classification, text clustering, orientation identification of opinion, topic detection and tracking, automatic summarization. At present, researchers conduct analysis mainly in following aspects:
- Hot topic detection. According to the parameters of topics from Microblog or Wechat such as authority of topic source, number of comments, intensive degree of speaking time, hot topics within certain period of time can be identified.
- Orientation analysis. Sentiment analysis on the contents and related comments on Microblog and Wechat.
- Topic tracking. Analyze whether there have been the same topics with the newly released topics on Microblog or Wechat.
- Automatic summarization. Perform the automatic summarization of various topics on Microblog or Wechat, so as to help understand the core sense of the topics.
- Trend analysis. Predict the development trend of topic by analyzing its degrees of received attention in different time periods.
- Analysis of emergencies. Through emergency analysis, the overall picture of the emergency can be obtained and thus the development trend the emergency can be predicted.
- Alarm system. Timely detection of emergencies and sensitive topics associated with public or personal safety and call the police.
- Statistical report. The report is concluded based on results of public opinion analysis, and provide information retrieval function.

2.2. Features of Social Networks Big Data Processing

According the requirement of processing time, the processing method for big data can be divided into two types: stream processing and batch processing. Batch processing follows the working pattern of store-then-process, which stream processing is a method of straight-through process. For example, stream processing is used for the Storm system of Twitter and the Kafka system of Linked in stream processing method is mainly used for on-line applications [10], where it work in the second or millisecond level. In stream processing, the potential value of data is the freshness [11], therefore the data should be processed as soon as possible and the processing results should be obtained quickly. In such processing method, the data arrive in the form of stream. During the data arriving process, there is only small part of stream data being stored in the limited memory even though the stream contains large number of data. The MapReduce program model from Google is the representative of batch processing pattern, of which the core design thought is to divide the problem into small parts and then calculate them, so as to effectively avoid the large amount of communication overheads in data transfer process. This model is widely used in the field of bioinformatics and text mining.

Seleh *et al.* [12] pointed out that data connection is the confluence of social networks and cloud computing, while big data provides solutions for social networks and cloud computing. Based on the two dimensions of universality and executive capacity, social networks can be divided into universal type social networks and executive type social networks. Moreover the universal type social networks can be divided into integrated social networks and specialized social networks; while executive type social networks can be further divided into information type social networks and executable social networks. For example Facebook and LinkedIn are integrated social networks, while Amazon Elastic Compute Cloud (EC2) belong to executable social networks.

The fast-developing social networks greatly improve the information dissemination in terms of amount, speed, and range. Through social networks platforms such as Facebook, Tecent QQ, Sina Microblog, people express their comments, sharing information and interact with each other. Social networks applications are equipped with real-time feature, sharing function, mobility, personalization and interactivity [13]. The key component elements of social networks are users, content generated by users, and social platforms, based on these three cores, relation between users, user message transmission, service and application provided by social networks platform are developed. Currently hot research topics on social networks include:

- Social networks analysis, such as statistic analysis and complex network analysis. Through social networks analysis, to further find verify the features and attributes of social networks.
- Community detection, which is evolved from the cluster analysis of complex network, is mainly used to explore the network structure and user relation.
- Researches of user behavior are various such as user influence research, user's behavior prediction, recommending for users.
- Information dissemination. Social networks is basically the carrier for information. Especially after the emergence of real-time social networks services such as Microblog and Wechat, information are spread faster and to greater range, making a great impact on the entire social opinions, therefore, it is very important to research the information dissemination pattern in social networks.
- Network dynamic evolution. With the explosive growth of social networks, in network dynamic evolution, researches focus on the development trend and growth model of social networks.
- Visualization. Social networks can be presented more clearly to decision makers; in addition, people can know and find the information in deeper level of social networks with visualization function.

Massive amount of data contains massive amount of information. Just obtaining the data is not enough, what is more important is to obtain the useful information from data, to extricate knowledge and information that can be used in market, thus the commercial competition advantages can be obtained. Processing big data of social networks is mainly to solve the problems on diversity of data source, on abundant data acquisition technologies and on diversity of data storage form, therefore for social networks big data processing, the key is not only about the massive data storage capacity, but also about the rapid and efficient computing capacity.

3. Social Networks Big Data Processing Framework Based on Cloud Computing

For the features of low-latency and high-throughput in big data analysis processing, the coordination between data and computing framework is the most important [12]. The high efficient data storage and management is the basis for data computing and processing. Cloud storage is a cloud computing system where data storage and management serve as core. Through cluster technology, thousands of storage devices can be virtualized into a big storage platform via network, so as to realize the coordinated and effective work. According to the differences in owners and operating modes, Cloud storage platforms can be divided a three classes: public cloud, private cloud, and hybrid cloud.

With advantages of big capacity, lower cost, high scalability, and high reliability, cloud storage provide effective method for big data management. Cloud storage environment, composed by thousands of basic facilities with relatively lower price, provides enough space for large-scale data storage and data computing. Normally social networks are of features such as magnanimity, diversity, real-time, spatiality, sensibility and heterogeneity. All these features play important role in determining the high efficient management of social networks big data. Considering the distinguished features, this paper adopted the hybrid cloud which is more suitable for social networks big data storage.

As one of mature technologies for massive data storage and computing, Hadoop is especially suitable for off-line big data processing (a completely open source project). In operating large-scale computer cluster, the two key sectors are HDFS for data storage and MapReduce for data computing.

The whole workflow of big data processing includes the extraction and integration of widely heterogeneous data source, unified data storage based on certain standard, analysis of stored data using proper data analysis technologies, so as to extricate useful knowledge within and present the results to end users in a proper ways, which can be concluded as 3 major sections including data extraction and integration, data analysis, and data evaluation and interpretation. Combing the features of social networks big data and the distributed framework of Hadoop, the social networks big data processing framework based on cloud computing was constructed which is shown in Figure(1).

3.1. Data Extrication and Integration

For big data processing, the first process is to extricate and integrate the data from needed data source, so as to extricate the relation and entity within. After association and aggregation, these data are stored in a unified structure. In addition, during data extrication and integration, it is needed to clean data to guarantee the quality and reliability of the data. Data cleaning must be carefully conducted, because the subtle useful information is mixed in the massive amount of data. If cleaning data over grindingly, the useful information is easily filtered out; while if cleaning data too coarsely, the genuine cleaning effect cannot be realized. Therefore a careful consideration and balance should be conducted between data quantity and quality. For example, there are large numbers of splogs on Microblog and Wechat, the identification of such splogs can enhance the efficiency of data cleaning.

3.2. Data Analysis

Data analysis is the core process in the whole big data processing, because the value of big data is produced in data analysis process. The data extricated and integrated from heterogeneous data source constitute the original data for analysis. According to the requirements of different applications, it is suggested to select the whole or parts of these data for analysis.

With abundant information on semantic ontology, social interaction, community-media, geographical map and multimedia content, multimedia data in social networks are structured. After analysis and mining, such information can be applied in commercial fields including mobile position retrieval, landmark detection, scene reconstruction, and recommendation of scenic spots.

3.3. Data Evaluation and Interpretation

Through presenting the results of data analysis to end users and involving users in the evaluation process, it can provide information for further data analysis. Normally visualization technology as well as human-computer interaction are effective methods of interpreting big data.

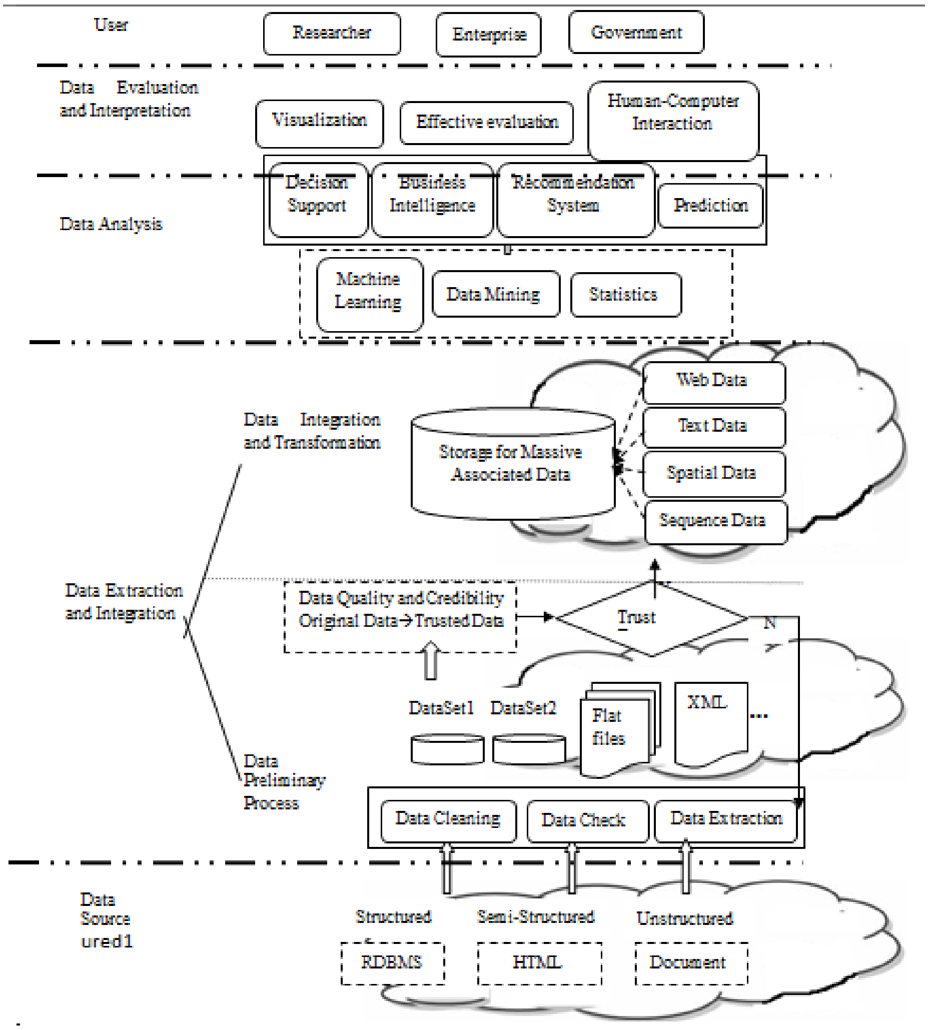

Figure 1. Social Networks Big Data Processing Framework Based on Cloud Computing

4. Social Networks Big Data Processing Strategy

4.1. Services Application of Social Networks Big Data on Microblog and Wechat

In the rapidly developing social platform such as Microblog and Wechat, there are massive users and massive information they left. Mining useful information for such massive data and realizing precision marketing and service recommendation are what all corporate expect. The Circle of Friends on Wechat and the Number of fans in Microblog can truly reflect the influence of the user. In addition the contents posted on Microblog and Wechat such as user location, text, voice, images and videos include user's life trace, hobbies and interests, liked brands, *etc*. The data processing procedure and personalized service application are shown in Figure(2).

Copyright ⓒ 2016 SERSC

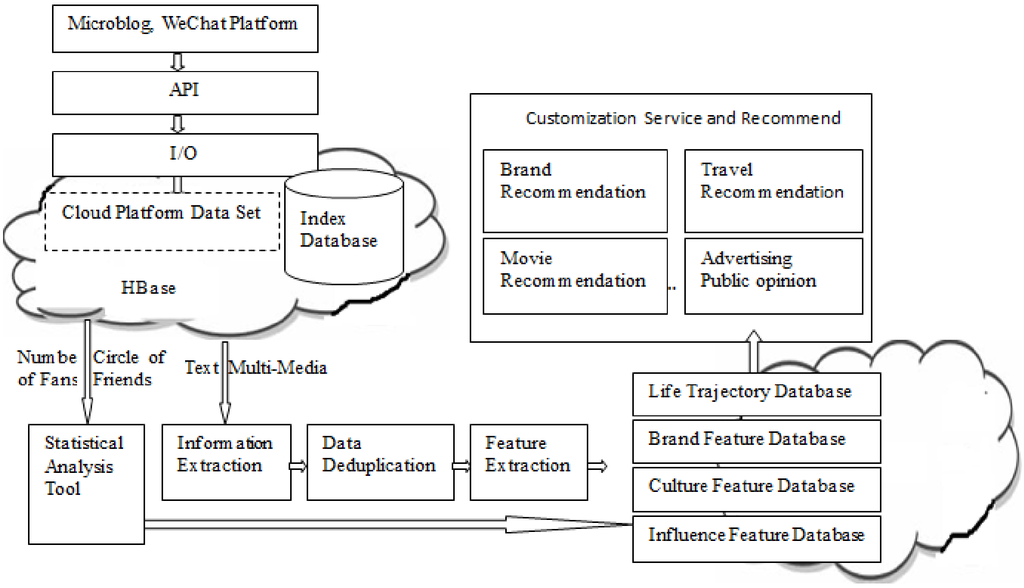

Figure 2. Data Processing Procedure and Personalized Service Application

Through the API provided by Microblog and Wechat platform, it can obtain the user's fans situation, situation of followers, information about circle of friends as well as the posted information such as location information, text, audio, images, and videos. Then store these data via data access interface into cloud platform data set, and store the information index into index database. After that the received multimedia information is converted which together with text information are calculated by clustering algorithm to extricate user's influence features, liked brand features, culture features and life trace. Subsequently, through storing these in corresponding feature library, related services and recommendations can be mined according to user's features. Data Deduplication, which adopts data compression technique, can solve replication of repeated data and effectively decrease the storage space.

4.2. Social Networks Big Data Processing Strategy

The requirements of social networks big data processing are not only reflected by the massive data processing ability, but also reflected by the formulation of effective and reliable data storage, safe and quick accessing and analysis strategy.

- ●Processing strategy of high efficient calculation. Since Hadoop data storage considers the load balance but not considers the relations between data sets, and all data stored by HDFS are placed according to the work load of Hadoop cluster, when running the Mapreduce calculation, large numbers of data migration will occur, thus increasing the I/O cost, especially the I/O cost of Shuffling. In order to increase the processing efficiency, there have been emerged many optimizing methods or strategies.By putting the data block related to social networks cloud storage platform into the same data node, the MapReduce I/O cost can be reduced. According to the Hadoop mechanism, the social networks big data was divided into data blocks before being marked such as time tag and association tag. It is worth noting that these making processing are in accordance with social networks service and specific application analysis. Through putting the associated data set into the same node, the calculation of each data set can be finished in the same node without going through the Shuffle section of MapReduce, so as to greatly increase the calculation efficiency of MapReduce.
- ●Processing strategy of incremental updating. Social networks big data include static historical data and dynamic incremental data. The strategy of upgrading based on

incremental data can be implemented without load the large-scale data into memory. In the context of distributed computing, through combining the parallel processing method and incremental information processing method, it is only needed to scan the new transaction database upon each time of incremental updating, which can save the time cost in scanning and statistics. Through constructing the index for each data block, it can reduce the time cost in scanning original transactions, and increase the incremental updating efficiency. In addition, dynamic incremental data is processed by distributed real-time stream processing pattern; while the static historical data is processed by distributed batch processing pattern.

- Strategy of sensitive data security sharing. Possessed with the features such as massive amount and multi-dimension, multi-source heterogeneous, extensive distribution, and dynamic growth like normal big data, social networks big data are also of features such as spatiality, real-time and sensibility. Due the privacy issue, the demand for security is even higher, and the access control should be consolidated. In the process of sensitive data circulating among multi-users and multi-data platforms, the sensitive data is threatened by some non-safety factors and suffer high risk of information leakage while providing values to corporate, therefore the strategy of sensitive data safe sharing is highly concerned issue. On one hand, adopting the secret key encryption strategy when data owner execute submitting, storage and extrication of sensitive data. On the other hand, to guarantee the safety in processing data, the protection method of user process under the virtual machine monitoring was adopted; By introducing trusted computing technology and constructing trusted environment and trusted channel, the application procedure and security plug-in of cloud service providers can protected free from the external interference. In addition, a consideration should be conducted on realizing a proper balance between consolidating security access control and increasing the convenience in data processing.

5. Conclusion

Big data processing brings great challenges to data management methods while providing extremely great convenience to human life. As the existing data processing methods cannot meet the requirements of social networks big data development, in this paper, regarding the social networks big data's features such as scale level, real-time, heterogeneity, complexity and privacy, it proposed security sharing strategy, privacy protection strategy and access control strategy for social networks big data resource. Since social networks big data is a huge complex system, the whole data management system should be reformed from the system structural level to each specific mechanism level. Only the joint effort of interdepartmental researchers is achieved, can be solved the big data processing problem.

Acknowledgement

This work was financially supported by project of National Natural Science Foundation of China (71272191).

References

[1] C. C. Aggarwal, "An Introduction to Social Network Data Analytics", Social Network Data Analytics. Springer US, **(2011)**, pp.1-15.

[2] L. Deng and J. Gao, "An Advertising Analytics Framework Using Social Network Big Data. Conference: The First IEEE international conference on Big Data Computing Service, and Applications", San Francisco Bay, California, USA, **(2015)**.

[3] G. Mishne, J. Dalton, Z. Li, A. Sharma and J. Lin, "Fast Data in the Era of Big Data: Twitter's Real-time Related Query Suggestion Architecture", Conference: 2013 International Conference on Management of Data, ACM, **(2013)**, pp. 1147-1158.

[4] J. Choo and H. Park, "Customizing Computational Methods for Visual Analytics with Big Data", Computer Graphics & Applications IEEE, vol. 33, no. 4, **(2013)**, pp. 22-28.

[5] J. Howard, M. Kazar, S. Menees, D. Nichols, M. Satyanarayanan, R. N. Sidebotham and M. West, "Scale and Performance in a Distributed File System", ACM Transactions on Computer Systems, vol. 6, no. 1, **(1988)**, pp. 51-81.

[6] R. Cattell, "Scalable SQL and NoSQL Data Stores", Acm Sigmod Record, vol. 39, no. 4, **(2011)**, pp. 12-27.

[7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Proceedings of the 6th Symposium on Operating Systems Design and Implementation, Berkeley: USENIX Association, **(2004)**, pp. 137-150.

[8] T. White, "Hadoop: The Definitive Guide", Yahoo! Press, USA, **(2010)**.

[9] Y. Simmhan, V. Prasanna, S. Aman, A. Kumbhare, R. Liu, S. Stevens and Q. Zhao, "Cloud-based Software Platform for Big Data Analytics in Smart Grids", Computing in Science & Engineering. vol. 15, no. 4, **(2013)**, pp. 38-47.

[10] K. Goodhope, J. Koshy, J. Kreps, N. Narkhede, R. Park, J. Rao and V. Y. Ye, "Building LinkedIn's Real-time Activity Data Pipeline. Bulletin of the Technical Committee on Data Engineering", vol. 35, no. 2, **(2012)**, pp. 33-45.

[11] N. Tatbul, "Streaming Data Integration: Challenges and Opportunities", Proceedings of the 26th International Conference on Data Engineering Workshops, California, **(2010)**, pp. 155-158.

[12] I. Saleh, T. Wei and M. B. Blake, "Social-Network-Sourced Big Data Analytics", IEEE Internet Computing, vol. 17, **(2013)**, pp. 62-69.

[13] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship", Journal of Computer-Mediated Communication, vol. 13, **(2008)**, pp. 210-230.

Authors

Liu Kewen, is a doctoral student in Management Science and Engineering at School of Management, Harbin University of Science and Technology. She is also a vice professor in School of Computer and Information Engineering at Harbin University of Commerce. Her research interests are in the areas of e-commerce, management information system and coopetition.

Gao Changyuan, is a doctoral tutor in Management Science and Engineering at School of Management, Harbin University of Science and Technology. His main research area is High-tech Virtual Industrial Cluster.