International Journal of Control and Automation
Vol. 8, No. 6 (2015), pp. 221-230
http://dx.doi.org/10.14257/ijca.2015.8.6.22

Research of UCL-based Broadcast Control Technology under Complementary Architecture Network

Cong Liu[1], Xinyu Zhang[2], YiGang Diao[3] and XinGangWu[4]

[1, 2, 4] *Information Technology Center, Tsinghua, Beijing, China*
[3]*Technology Laboratory of Technical and Technology Bureau Xinhua News Agency, Beijing China*
[1]*liuc@tsinghua.edu.cn,* [2]*xyzhang@tsinghua.edu.cn,* [3]*thomasfred@xinhua.org,* [4]*wuxingang@tsinghua.edu.cn*

Abstract

According to the scale-free phenomenon of Internet and the puzzle of semantic web, we study UCL-based broadcast control technology under complementary architecture network, introduce the concept of complementary architecture network and UCL, propose UCL-based sharing-distributing system structure for web data under complementary architecture network, mainly analyze the key technologies of broadcast control based on UCL under complementary architecture network, describe the system architecture design, main implementation methods and algorithms of UCL-based super master database and UCL standard data integrated broadcast control platform, and implement a UCL-based integrated broadcast control prototype system, proving and testing feasibility and availability of integrated broadcast control platform proposed in this paper. Therefore, a complete broadcast control platform has been initially established, laying solid foundation for further researches.

Keywords: *complementary architecture network, UCL, broadcast control technology, keyword extraction, abstract extraction*

1. Introduction

With the rapid development of network technology, Internet has become an important way for people to get information, and network information resources have also shown a tendency to grow exponentially. Faced with such a tremendous amount of information resources, the existing single structure of Internet has had a great challenge. Experts and scholars in various countries have carried out researches on the structure and characteristics of Internet, and have put forward a variety of solutions. For scale-free phenomenon of Internet, academician of China Engineering Academy Youping Li innovatively proposed the idea of a complementary architecture network. The main idea is to add a special secondary architecture called broadcast-save architecture transmitting primary resources to the user terminals in the shortest path by broadcasting based on keeping the primary architecture of TCP/IP, which is fusing the advantages of Internet and broadcasting network. Academician Li also proposed using UCL (Uniform Content Locator) such a semantic tool to achieve new cultural services [1]. This service will make "people looking for information" change into "information looking for people" such kind of active services, solving bandwidth bottleneck, digital divide and other conflicts.

In the complementary architecture network, on the one hand we need to analyze and organize network information, add content indexing information, and select Internet or broadcast channel transmission according to the information characteristics. On the other hand, for the end users, ways to access information increase, structures and types of information become more complex. How to find the information they need quickly and accurately? How to shield irrelevant information and harmful information? Such issues

ISSN: 2005-4297 IJCA

must be solved in the complementary architecture network. Therefore, we need a suitable for the characteristics of complementary architecture network, efficient information broadcast control technology to support. To solve this problem, we mainly study UCL-based broadcast control technology under complementary architecture network, in order to change existing information gain pattern from user actively searching information to user getting interested information in an initiative pushing way and ensure the dissemination of health content.

2. Basic Concepts and Architecture

2.1. Complementary Architecture Network

Recent scientific studies found that the mathematical model of Internet has developed to be a power distribution scale-free model from original normal distribution random model [2]. Daily accesses of visible users have obvious cybotaxis. If we integrate thousands of popular sites, it is possible to form the mainstream of information resources to meet the daily needs of most users. According to it, complementary architecture network is proposed by academician Youping Li. The main idea is to add a special secondary architecture to transmit primary resource based on keeping the primary architecture of TCP/IP, as shown in Figure 1. Because of having network architecture and broadcast-save architecture at the same time, it can transmit primary resource to all the country by satellite broadcast, overcome digital divide, realize the goal of "getting the right information to the right people at the right time in the right format".

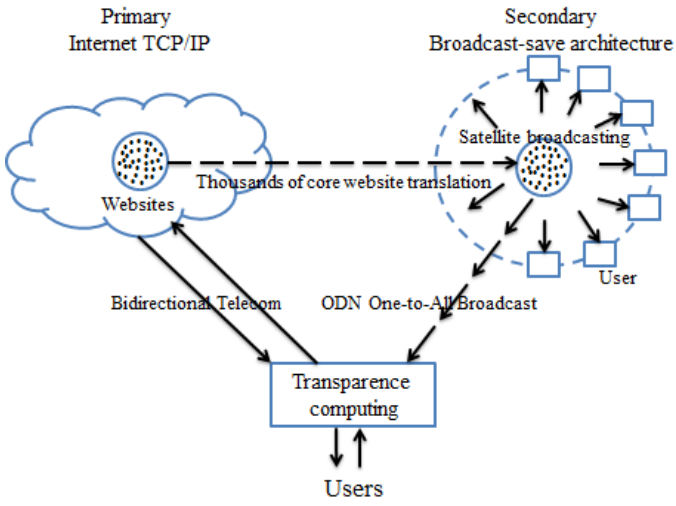

Figure 1. Complementary Architecture Network

2.2. UCL

UCL is an innovative technology idea to achieve organization management of information resources in the data broadcasting network platforms, such as cable TV HFC (hybrid fiber coax) network and IP interactive network platform, such as Internet. Its goal is to realize active personalized service according to the content of the network information resources. Technically, UCL can be used to transform any network, making it possible to become the intelligent network to understand the information content [3].

The purpose of UCL is to solve the network information discovery, search, identification, transmission, control, active service and other issues. UCL technology can automatically classify and label the resource content according to

the preset criteria in the source production stage, use language to express content requirements, establish a people-oriented cyberspace information model, and actively push the required webpage through content filtering mechanism. UCL is usually to make a multi-dimensional indexing of resource content, including category, subject, source, author, keywords, *etc*.

The UCL vector can be expressed as $U= (u_1, u2,\ldots, un)$, where u_i, $i \in N$, n is the number of UCL components and n is generally related by described objects, applications, transmission mode and user terminals.

2.3. Basic Architecture

UCL-based sharing-distributing system structure for web data under complementary architecture network shown in Figure 2, as a secondary structure and traditional IP Internet constituting a primary and secondary relationships, can effectively complement and overcome the traditional Internet difficulties, such as content security, flow distribution, etc.

In accordance with the requirements of broadcast-save theory, fully meeting the needs of the next generation technology system work of national mainstream media agencies and Internet regulators, we mainly study UCL-based broadcast control technology under complementary architecture network, including UCL-based super master database, UCL standard data integrated broadcast control platform, and UCL-based prototype system for verification test.

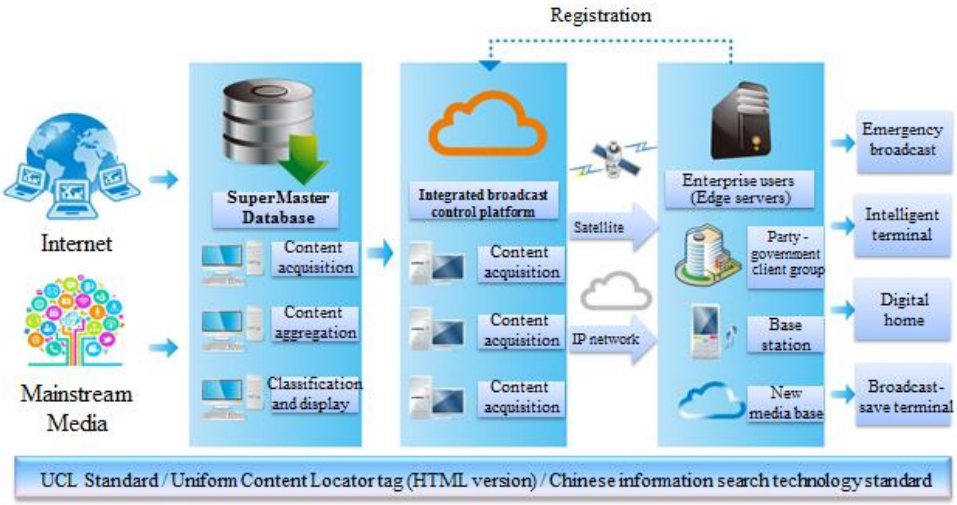

Figure 2. UCL-Based Sharing-Distributing System Structure for Web Data under Complementary Architecture Network

3. Key Technologies

3.1. Semantic Analysis of News

News event is the main way for information providers to collect, edit, release and store the news, and is a natural way for information consumers to understand the news. Thus, an event-based three-layer semantic model shown in Figure 3 is adopted. The news set, connecting with social media and public knowledge, can be described as event, topic and entity three layers. Furthermore, inter layer and inner layer have semantic association.

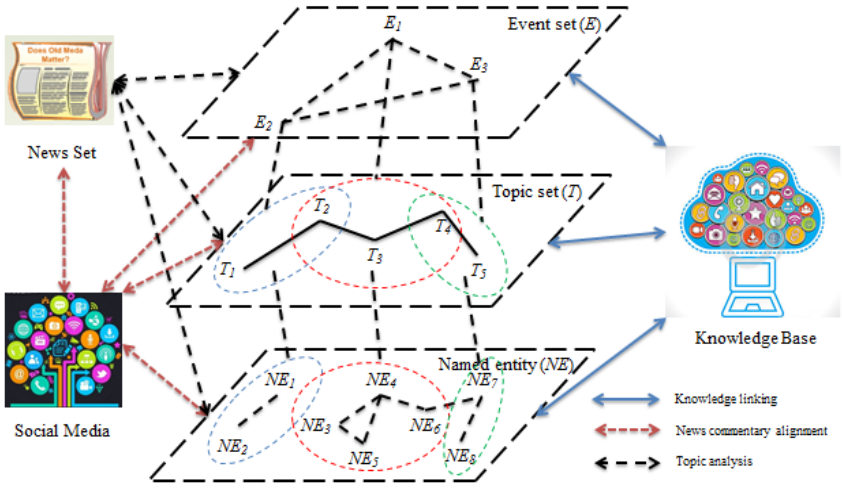

Figure 3. An Event-Based Three-Layer Semantic Model

The research shows that the link possibility of named entity is positively correlated with the context similarity of their appearance, and semantically related entities often appear together. Firstly, we generate candidate links through entity name. Secondly, we represent named entity as an eigenvector composed of context content and co-occurrence entities, and compute vector cosine similarity to generate the optimal matching. Finally, with the help of the possible existent structured information in a knowledge base, we optimize the results and generate the final link results.

3.2. Automatic Classification

Network characteristics such as high-capacity, strong timeliness and high propagation speed make the discrete document organization form difficult to meet the demands that is for user to quickly browse the information content and to explore the deep connotation. This technology of automatic classification, through topic mining, entity relationship analysis and topic sentence extraction for the document set, gets a lot of statistical data and semantic description related to the content event, and based on the above information automatically matches UCL classification criteria to generate corresponding classification attribute tag.

3.3. News Clustering

Through in-depth analysis and content clustering for news big data, we can realize intelligent processing and semantic analysis of mass media resources, break through the key technology of media content understanding, significantly enhance the use value of news big data, and provide technical support for semantic-based content search of new media resources, regulatory services of news resources, content aggregation and presentation services of focus news, data association and mining services of news and other new media integration services related to business links. Furthermore, we can achieve topic-based news clustering, help news editors to quickly comb report ideas, quickly organize news information products, and build highly reliable news production auxiliary platform based on the key technology of news content aggregation.

3.4. Digital Fingerprinting

Because the characteristics of digital products are easy to modify, copy and disseminate, copyright protection problem becomes more and more important. Digital fingerprinting is a new technique for multimedia copyright protection in recent years. The so-called digital fingerprinting is a feature sequence to distinguish between similar things

and to be processed by a computer. Because it has characteristics of uniqueness, robustness, invisibility, in the distribution of digital products, the sellers can add digital fingerprinting in the copies of digital works, identify traitor making illegal copies and provide the court trial evidence after discovering illegal copies, so as to achieve the purpose of copyright protection.

3.5. Keyword Extraction

Keyword extraction is a process to extract some of the words most associated with this news significance from the UCL-based news, abstract and title. There are two main methods for keyword extraction, i.e. keyword matching and keyword extraction. Given a keyword thesaurus, keyword matching is to find related words from the given thesaurus as the article keywords. Whereas keyword extraction is to directly extract some words from the article as the article's keywords. At present, most domain-independent keyword extraction algorithm and its corresponding libraries are based on the latter. Logically speaking, the latter is more meaningful than the former in actual use.

Keywords extraction results are also divided into two types. The first is simply to extract words, such as FudanNLP [4], jieba [5], SnowNLP [6]. The second is to extract the conjunction of words and phrases, which needs to increase the phrase extraction this step, such as ICTCLAS [7], Ansj [8]. For clustering or classification, it is very obvious that phrases are more valuable than words.

3.6. Abstract Extraction

Automatic text summarization mainly uses machine learning related model to extract abstract from the text. Different from keyword extraction, abstract should contain the core content of original text or user interested content, and is output as the form of paragraph or even chapter with semantic coherence. The goal of abstract extraction is committed to making comprehensive concise text information directly present to user, improving the efficiency of user access to information.

The techniques of automatic text summarization can be divided into extraction and abstract two kinds of methods. Most implementation technologies of extraction is based on statistical methods and shallow semantic understanding, whereas the implementation technologies of abstract is based on the natural language understanding technology. The content of statistical-based extractive abstract is completely the part copy of the source document. The semantic-based understanding abstract is composed by restructuring or alternative of terms and sentences to form a summary using natural language processing techniques.

4. System Design

4.1. UCL-Based Super Master Database

Based on the existing point-to-point and multi-hop addressing Internet with a single structure, UCL-based super master database is established, mainly including oriented mass network media data aggregation, tag extraction and content distribution. The architecture of UCL-based super master database is shown in Figure 4.

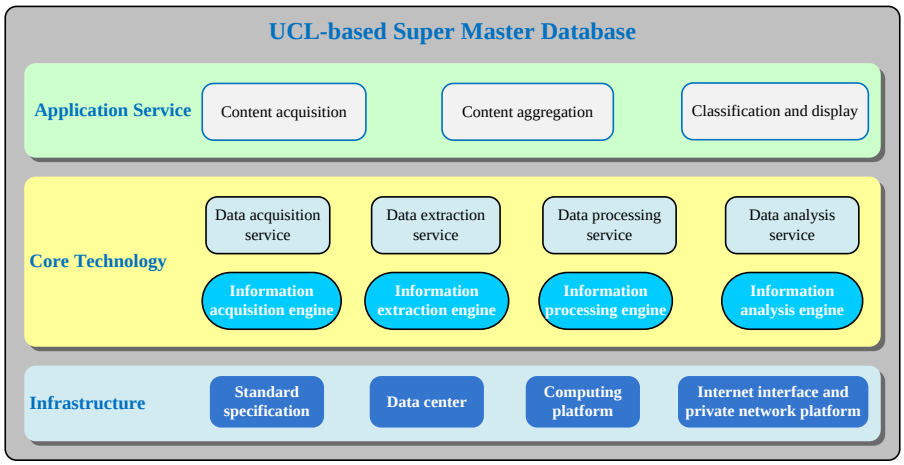

Figure 4. The Architecture of UCL-Based Super Master Database

4.2. UCL Standard Data Integrated Broadcast Control Platform

UCL standard data integrated broadcast control platform integrates the industry-leading technologies, such as content security, license management, and other management control methods, for establishing semantic computation links between network media data resources and users. Unified integrated broadcast control technology could ensure that the broadcast control platform has powerful capabilities of content verification and content audit, from the dimensions of content security, tamper-proof and anti-attack to manage and distribute Internet content. This platform has explored the UCL application mode, the unified management of licenses, topics and risk control for broadcast control data. Integrated broadcast control focuses on the timing and the priority towards distributed full text information and UCL abstract information providing broadcast control. The data flow of UCL standard data integrated broadcast control platform is shown in Figure 5.

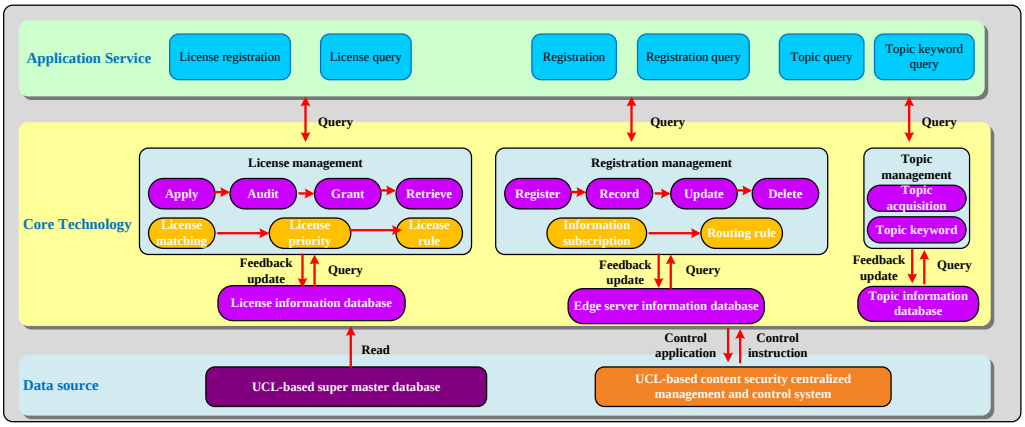

Figure 5. The Data Flow of UCL Standard Data Integrated Broadcast Control Platform

4.3. Key Algorithms and Implementation Methods

In this section, we describe the key algorithms and implementation methods in this paper.

4.3.1. Keyword Extraction Algorithm: In accordance with the draft version of UCL format specification, we develop Application Programming Interface (API) of UCL using

Java to achieve uniform content label encoding for information with CNML (Chinese News Markup Language).

UCL basic format is shown in Figure 6. The part of Code is the data header, recording version number and other relevant information. The part of Properties is the content data, storing author, source, abstract, *etc*. UCL uses binary coding, in order to improve the efficiency of access and storage.

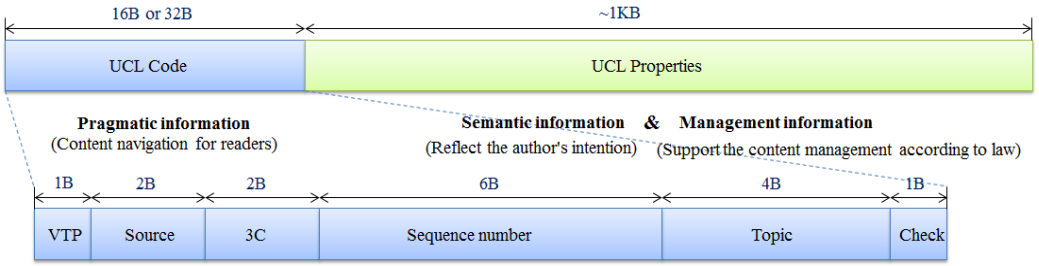

Figure 6: UCL Basic Format

The API of UCL is divided into two parts. One part is in the UCL project, as the main UCL implementation and the server. Another part is in the UCLTest project, as the client example. Code is packaged as a jar for the UCL project reference.

4.3.2. Keyword Extraction Algorithm: We use the position of word first appeared in the article for keyword extraction algorithm in the news in addition to using TF-IDF (Term Frequency–Inverse Document Frequency). The reason for using this way is that most articles especially for news text are "general-devide-general" structure. It is obvious that the probability of a word appeared in the header and tail to become the keyword is more than a word only appeared in the middle of the content words. Given different weights according to the position in the article first appeared on each individual word, we realize the keyword extraction algorithm using combination of TF-IDF and continuous data discretization method.

The principle of Ansj keyword extraction is based on part of speech, TF-IDF of word segmentation in document and the first position of keyword distribution. Ansj has implemented the keywords extraction, but the output of Ansj does not give each keyword weight. Thus, we need to manually modify the keyword class. It can be found that in the source code of Ansj weight members have been defined. Therefore we need to add a Get function in the program.

4.3.3. Abstract Extraction Algorithm: We use the combination of statistical-based extractive abstract and semantic-based understanding abstract to realize news abstract. The main steps of the algorithm are described as follows.

- First we use the Ansj to achieve word segmentation for news, and get all sentences and the various components of each sentence in the news.
- Second by means of LDA (Latent Dirichlet Allocation) model of natural language processing, we develop text processing. LDA is a very practical topic model, which is based on the news is composed of multiple topics.
- Third we can obtain the news topics and their corresponding keywords, and evaluate each news sentence.
- Finally, we take the highest weight of the sentence as the news abstract.

The topics of the text and the keywords corresponding to each topic are given in the form of Map. Each topic and keyword has its corresponding weight. After that, we find topic keyword in each sentence of the text, and use weights of existing keywords to grade

sentence. Finally, sort the sentence score, and get sentences with high scores as the news abstract.

5. Prototype System

According to the above research design and combining the actual work of Xinhua News Agency, we have accomplished key technologies of UCL-based integrated broadcast control in the important typical application. The system functions of typical demonstration application include black-white-gray list of active defense control, tamper-proofing control, anti-attack control, topic security control, and webpage content security control, as well as unified management for the license plate, topic and risk control of broadcast control data. The interface diagram of prototype system is shown in Figure 7.

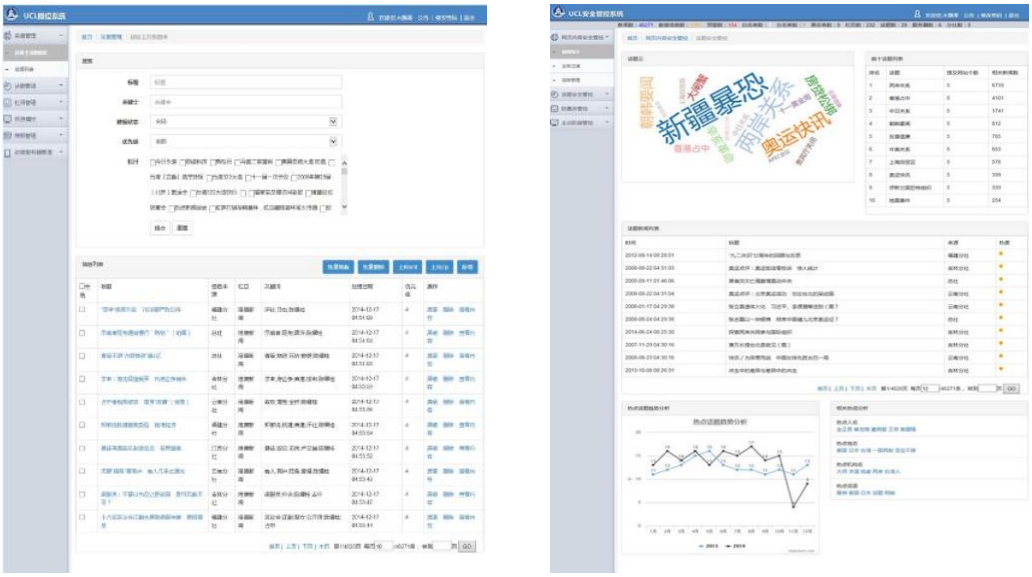

Figure 7. The Interface Diagram of Prototype System

6. Conclusions

The research direction of this paper is for understanding, representation, analysis, aggregation, filtering, and distribution of UCL-based semantic information under complementary architecture network. We mainly focused on the information characteristics of complementary architecture network, carried out efficient UCL-based experimental study of information integrated broadcast control technology, and has made some achievements. In the complementary architecture network, it is possible for the country to manage and control the pushing information resources, to reduce the information replication and widespread circulation of unhealthy information. At the same time, such network architecture will change existing information gain pattern from user actively searching information to user getting interested information in an initiative pushing way, ensuring the dissemination of health content.

To achieve more effective forwarding of broadcast control, more targeted management and control of news, we need to do further research on massive information data and user behaviors in the system application, in addition to building a scientific and rational system platform. The further research can ensure that the system operation adapts to the characteristics of business and users, in order to achieve the optimal performance of the whole system. In the system application, the core issues of performance optimization include news topics, propagation paths

and user behaviors. The next step will be to conduct a thorough research on the above issues.

References

[1] Li Youping. The secondary web of knowledge embodiment. Engineering Science, 2002, 4(8), p8-11
[2] Li Youping. Research of Complementary Architecture Network. Journal of Southwest University of Science and Technology, 2006, 21(1), p1-5
[3] McGuinness, D.L. Question answering on the semantic web. IEEE Intelligent Systems, 2004, 19(1), p82-85
[4] Information on http://www.oschina.net/p/fudannlp
[5] Information on http://www.oschina.net/question/tag/jieba
[6] Information on http://www.oschina.net/p/snownlp
[7] Information on http://www.oschina.net/p/freeictclas
[8] Information on http://www.oschina.net/p/ansj

Authors

Cong Liu is an engineer in Information Technology Center at Tsinghua University. Her research interests are major in university education informationization including electronic school affairs system and online learning. She holds an MS in computer software and theory from Northeastern University.

Xinyu Zhang is a senior engineer in Information Technology Center at Tsinghua University, a vice general secretary of Chinese Association for Artificial Intelligence and a visiting scholar at Cambridge. His research interests are major in network education and unmanned system platform. He holds an MS in Humanities and Social Sciences from Tsinghua University.

Yigang Diao is deputy director of technology laboratory of Xinhua News Agency, chief editor of "Chinese Mass Media Technology" magazine. His research area covers technology standard in Chinese media region, data mining, Natural language processing, content security protection on internet and etc. He graduated from Tsinghua University and achieved master degree in 2006.

Xingang Wu is an engineer in Information Technology Center at Tsinghua University. His research interests are in the higher education informatization including office automation and online learning. He holds a bachelor's degree of Business Administration of University of International Business and Economics.