International Journal of Grid and Distributed Computing
Vol. 9, No. 7 (2016), pp.289-300
http://dx.doi.org/10.14257/ijgdc.2016.9.7.29

Research on the Key Technologies of Base Station Server Content Delivery Based on Broadcast-Storage Architecture

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

Focusing on Internet's emerging paradigm of ubiquitous content sharing, following the Broadcast-Storage concept, we study the key technologies of base station server content delivery based on broadcast-storage architecture, introduce a general architecture model of broadcast-storage network, propose a technological framework of base station server content delivery based on broadcast-storage architecture, analyze the key technologies and application algorithms included in the foundation technology layer and the key technology layer. In the key technology layer, we mainly research on Satellite-Base Station data processing technology, base station content service platform, content security management-control, topic security management-control, and user management. Finally, we summarize the nonfunctional requirements of base station service platform. Based on the above research, a base station server content delivery system based on broadcast-storage architecture is designed and implemented, which is docked with the integrated broadcast control platform, integrated management-control platform and clients, completing application demonstration in important typical business. Therefore, a complete broadcast control content delivery system has been initially established, laying solid foundation for further researches.

Keywords*: Broadcast-Storage architecture, content delivery, base station server, UCL, content security management-control, topic security management-control.*

1. Introduction

In recent years, the inadaptability between the Internet architecture and the needs of reality application is increasingly apparent, and the behavior of netizen accessing the Internet presents scale-free phenomena [1-3]. Several major problems appeared on the Internet, such as bandwidth bottleneck, information garbage, semantic barriers and digital divide, are very difficult to eliminate entirely on the Internet itself, which brings opportunities and challenges to the research and development of new generation information network technology. Accordingly Li Youping academician proposed a new network structure that is to add a network of Broadcast-Storage architecture on the basis of TCP/IP primary structure of Internet [4,5]. This kind of Broadcast-Storage network is used for broadcasting those popular information with insensitive interactivity and large capacity. On the one hand it can reduce the pressure of Internet bandwidth, on the other hand combining with UCL (Uniform Content Locator) technology it can realize the intelligent application of network information. The combination of Internet and broadcasting network, each playing its advantages, is both forming complementarity in the structure and transmission contents.

ISSN: 2005-4262 IJGDC

In the Broadcast-Storage 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 must be solved in the Broadcast-Storage network. Therefore, we need a suitable for the characteristics of Broadcast-Storage network, efficient content delivery technology to support. To solve this problem, we mainly study the key technologies of base station server content delivery based on broadcast-storage architecture, in order to achieve the ideal target namely the hottest content stored closest to the user and 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. General Architecture Model and Technological Framework

2.1. General Architecture Model of Broadcast-Storage Network

As the term suggests, physical broadcast supporting radiative transmission from point to surface and corresponding ubiquitous content storage are the two key elements to construct the physical infrastructure of Broadcast-Storage network. Broadcast-Storage network using natural one-to-many ability of broadcast transmission can ensure that content is delivered once and scalefree users can receive, which achieves content copied unlimitedly in spatial dimension and from the basic physical transmission model satisfies delivery requirements with one-to-many radiation type of content sharing application. Meanwhile, Broadcast-Storage network in the receiving part introduces ubiquitous storage widely deployed in various content receiving end, making the sending and receiving of the broadcast content decoupled, ensuring that any receiver can flexibly receive and cache the content information sent by the broadcasting source, realizing unlimited copying content in time dimension, so as to effectively support asynchronous personalized demands of user access to content. General architecture model of broadcast-storage network is shown in Fig.1.

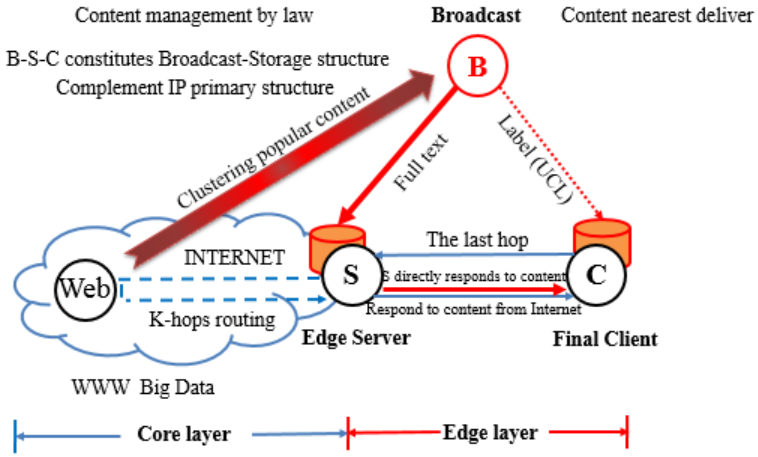

Figure 1. General Architecture Model of Broadcast-Storage Network.

Broadcast-Storage network fully embodies the content-centric design concept, and adopts UCL (Uniform Content Label) to achieve the identification, delivery, caching, navigation and adaptation of the contents. In short, UCL can be expressed as two-tuples

<UCL_Code,UCL_Properties>, where UCL_Code is the UCL identification code with the fixed structure, and UCL_Properties is the description of content attributes with flexible variable-length. UCL can more comprehensively describe the rich semantic information about the content, can closely relate to the readers, authors and administrators of content, and can provide the foundation for content-based security. The organic combination of broadcast delivery and ubiquitous storage, and UCL acting as a bridge between the two, are the creative primitives to together constitute Broadcast-Storage network architecture.

2.2. Technological Framework

The technological framework of base station server content delivery based on broadcast-storage architecture is shown in Figure 2. The design framework can be divided into foundation technology layer and base station service platform. The key technologies of each part are described in the following sections.

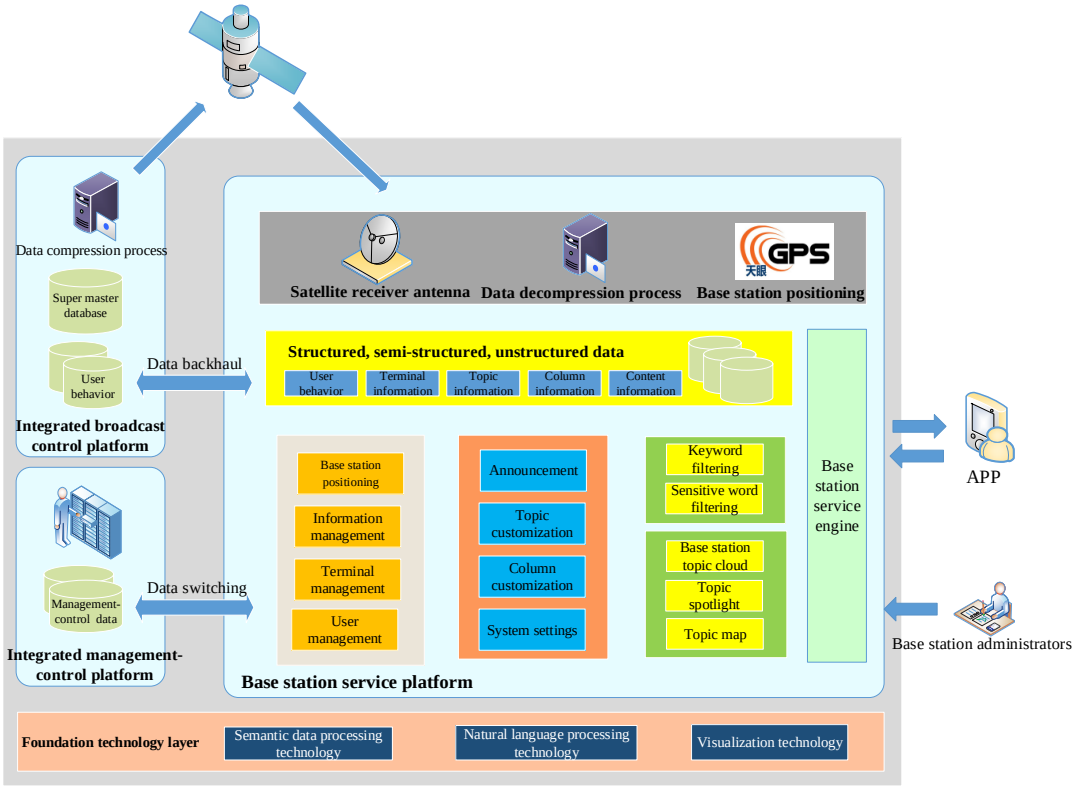

Figure 2.The Technological Framework of base Station Server Content Delivery based on Broadcast-Storage Architecture

3. Analysis on the Foundation Technology Layer

The foundation technology layer contains the relevant generic technologies, mainly including semantic data processing technology, natural language processing technology, and visualization technology.

The goal of semantic data processing technology is to transform the World Wide Web into a data web, not a documentation web [6]. The essence of semantic technology is to mark the key data, pictures, tables, summaries, conclusions and references of each piece of information in the writing, recording and publishing process, and to make these tags

associated to other related information by automation and stylization links. Information dissemination, integration and interaction design based on semantic technology has gradually become the forefront of scientific communication and system development. It enhances the association between data and information through improving the access efficiency of data and information, so it is more convenient for users to find, learn and understand the complex knowledge system.

Natural language processing technology is the use of computer to achieve processing and operations on text pronunciation, form, meaning and other language information, including input, output, identification, conversion, compression, storage, retrieval, analysis, understanding and generation for word, vocabulary, phrase, sentence and chapter. It is a discipline-crossing interdisciplinary subject of linguistics, computer science, cognitive science, mathematics and so on. In this system, natural language processing technology is mainly used for news text word segmentation, recognition and understanding, which is the basis of the key process including abstract extraction, entity recognition, relation extraction and opinion analysis, and is the basis of building ontology knowledge base and processing news semantic.

Visualization technology refers to using graphics technology to help people understand and analyze the news data. It can take the objects interested by people in the news presented to the user in a graphical way visually and vividly, which is easy for users to quickly understand and accept.

4. Analysis on the Key Technology Layer

The key technology layer is composed of *Satellite-Base Station* data processing technology and base station content service platform.

4.1. Satellite-Base Station Data Processing Technology

The massive news data of super master database is the basic work of the study. The massive data of super master database via satellite are compressed as upstream data and are decompressed as downstream data. The processing content includes information content, topic content, keyword content, and other contents in super master database. *Satellite-Base Station* data processing flow is shown in Figure 3.

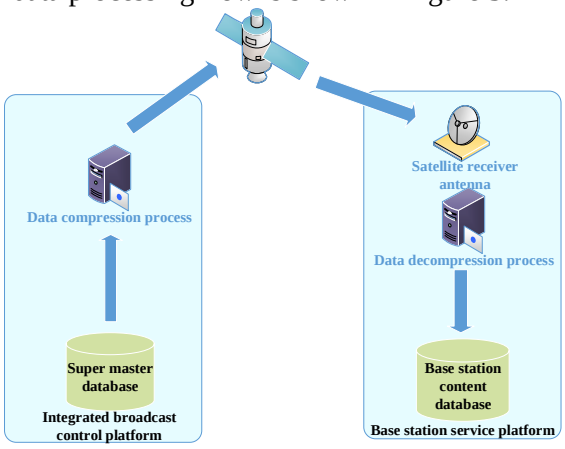

Figure 3. Satellite-Base Station Data Processing Flow

Most of the news data in super master database are static data, so we adopt SDT method in our research. Because SDT data compression algorithm is a common method in piece wise linear method, due to its outstanding advantage that the algorithm is simple and fast [7].

4.2. Base Station Content Service Platform

Base station content service platform mainly provides services for the clients and the base station administrators. The data of the platform includes the decompressed contents of super master database, the topic management, the column management, the client user data and the client user behavior data. Base station content service platform provides the user data and the user behavior data passing back to broadcast control platform, provides content data with UCL, topic matching, column pushing and security operation management-control to the clients, and provides content management, data statistical analysis and data visualization to the base station administrators.

Base station data management services mainly include news content information management, column information, topic information, terminal information and user behavior data. News content information mainly refers to media files with UCL label in super master database, including its extraction theme, keywords and abstract. User behavior data is the backhaul from the clients about their behavior data.

Base station application service mainly provides services for the clients and the base station administrators through the base station service engine.

The service function of the client mainly includes topic recommendation, security operation and data return.

Security operation is mainly relying on the content security management-control and topic security management-control in the whole network of security operation to provide management-control services for the clients.

The service function of the base station administrators mainly includes the following functionalities.

- *Announcement*. Send system announcement to all users who access to the base station.
- *Topic & Column customization*. Within the range specified by the master server, news sent from the base station to the user's mobile phone terminal can be customized on the level of topic and column.
- *System settings*. Make system parameters settings to the base station, such as viewing GPS coordinates automatically obtained and locating Province's Administrative Region, for adapting the topic scheme customized by the master server.
- *Other settings*. Manually set Province's Administrative Region, applying for those conditions where GPS coordinates cannot be obtained automatically.

4.3. Content Security Management-Control

The content security management-control module of base station service platform is responsibility for managing and controlling news information of base station, providing word-level coarse-grained content filtering. The purpose is to deal with the rapid response to public sentiment, in particular including real-time collection, rapid analysis of public sentiment information, hotspot capture, grasping the direction of public sentiment, predicting the crisis level, and then helping the administrators of base station service platform make feedback at first time.

The module accommodates hybrid information filtering model shown in Figure 4, which is mainly divided into two-layer structure including keyword-based matching filtering and UCL-based sensitive word filtering.

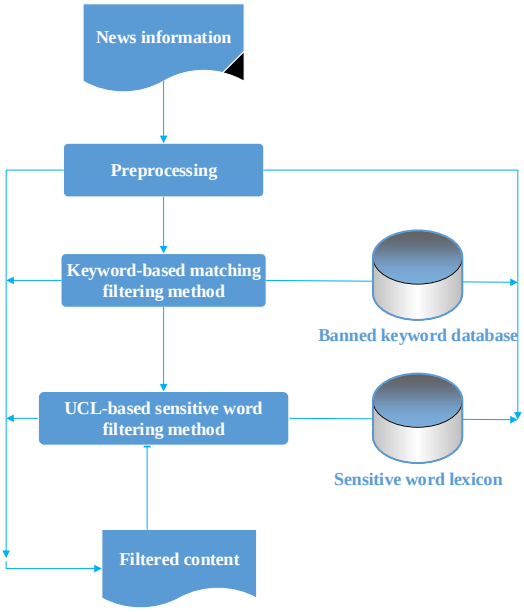

Figure 4. Hybrid Information Filtering Model

4.3.1. Keyword-based Matching Filtering Method: The principle of keyword-based matching filtering method is described as follows. First, prepare a keyword lexicon and preselect some keywords from the lexicon. When make a judgment on a news content, the news content is matched with the words in keyword lexicon one by one. If there are one or more keywords that can be matched with the text content of news information, that is keywords appearing in the text, it is said that this webpage belongs to management-control information required to be filtered. Otherwise, it is said that this webpage does not need to be filtered.

The advantage of keyword-based matching filtering method is the algorithm with high accuracy, high efficiency and fast running speed. Its disadvantage is that only relying on simple word matching is unable to understand semantic information of the news text and is unable to grasp the news connotation in depth, so it is difficult to make accurate judgments on the news content.

4.3.2. UCL-based Sensitive Word Filtering Method: In order to supplement the shortage of keyword-based matching filtering method, we propose the use of sensitive word lexicon combined with emotional analysis to implement secondary filtration. Sensitive words refer to those words needed to be prohibited or controlled in the news content, which often would bring extremely negative influence to society. However, such news may be positive news to combat these negative information. So it needs to further analyze the emotional tendency of news to determine whether or not to prohibit or strengthen management-control of this kind of information.

Sensitive word lexicon can be maintained by the administrators, which is constantly updated as demand changes. Depending on the influence of sensitive words, our study divide sensitive words into three priorities. According to the ranking order matching filter, and take different management-control ways in each sensitivity level.

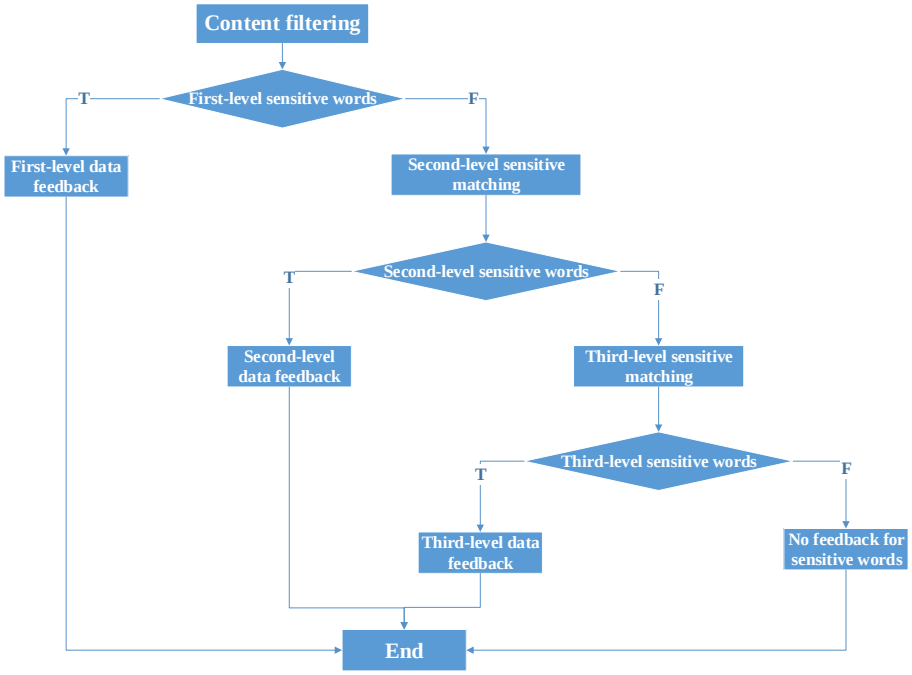

Figure 5. Sensitive Word Filtering Process

Sensitive word filtering process is shown in Figure 5. Due to sensitive words divided into three levels, in order to improve system efficiency for keyword filtering, the matching can be carried out in three steps. Loading and scanning is in accordance with the order of first-level, second-level and third-level. If there is a sensitive word in the upper level, the scan stops. Otherwise, the scan continues. This can reduce the scan time and improve the efficiency.

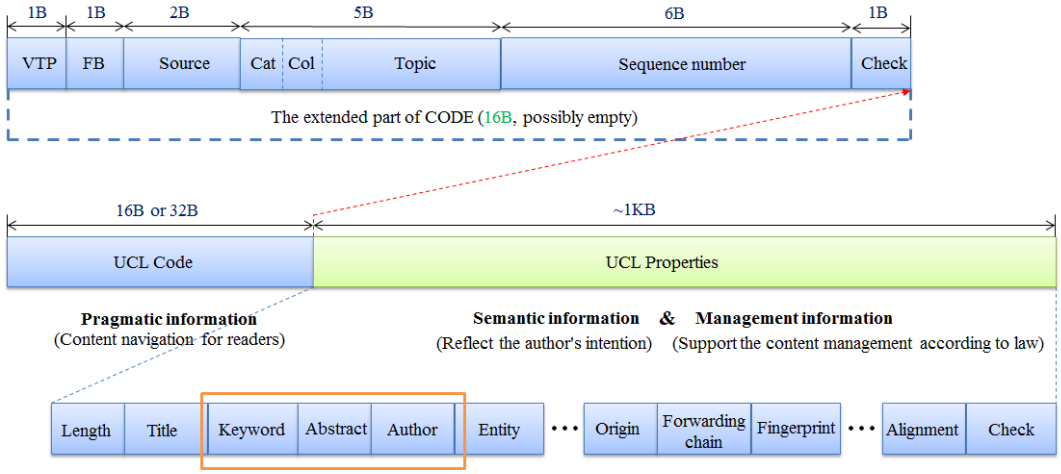

Figure 6. UCL-based Sensitive Word Filtering Method

Before filtering sensitive words, we need to make necessary preprocessing for news text, such as removing special characters, converting traditional and simplified Chinese characters, and word segmentation.

After the above-mentioned technical means to obtain sensitive word lexicon, we combine UCL-based news with title, keywords and abstract shown in Figure 6, and use multi-level sensitive word string matching algorithm to accomplish security management-control of sensitive words.

4.4. Topic Security Management-Control

The new media content in line with UCL standards has been accomplished topic extraction and model calculation in advance. UCL label generation technology will make news content describing the same content within a certain period extracted a topic.

Topic recommendation refers to establishing analysis and identification system on user's past behavior and reading interest based on the binary relation constructed between the client users and news content with UCL label. The recommendation system is composed of three parts including recording module, analysis module and recommendation algorithm module. The recording module is responsible for collecting user information behavior. The analysis module is responsible for analyzing user preference model. The recommendation algorithm module is the most core part. Our study adopts the combination of collaborative filtering algorithm and content-based recommendation algorithm.

Topic security management-control of base station service platform mainly includes visualization technology of topic cloud, hot topic management-control technology and topic map management-control technology. The specific schemes are described in the following sections.

4.4.1. Base Station Topic Cloud: Based on users and data of base station service platform, base station topic cloud is built. Topic cloud is an effective tool of topic visualization. Figure 7 is an example of topic cloud formed to tag combination of events, where text color and size respectively represents categories and heat.

Figure 7. An Example of Topic Cloud

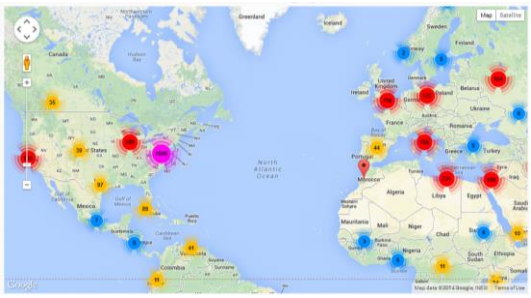

Figure 8. An Example of Hot Topic Event and Map

4.4.2. Hot Topic Management-Control Technology: Using hot topic calculation, we develop regression analysis of topic heat, dynamically maintain and update UCL hot topic system. UCL label generation technology adopts Single-Pass clustering method based on distance and geographical position to achieve news content clustering. At the same time, using the news event corpus and combining LDA model based on topic model and probability distribution, we extract possible latent semantic topic in the media content. The input is the set of news documentation fragmented by time. The output is the topic category in different time periods and statistical results of news document amount. According to the statistics, it is considered that which contains more news documents belongs to the hot event. In this way, we can get the changes of event heat over time, as well as the heat of each event category in the different time periods.

Using map visualization tool, we can simulate the location and the heat of events shown in Figure 8 as an example, where the position of circular point in the map represents event position, different color represents the heat level of events, and the number in the circular point represents the amount of corresponding news.

4.4.3. Topic Map Management-Control Technology: Topic map construction is the core part of the topic process. Its purpose is for mining the relationship between the potential named entities in the news topics to obtain the topic source and its composition.

Topic tracking. For the topics of different time slices, we calculate the correlation degree among the topics, usually based on the topic keywords, 5W1H and other key elements. Whose similarity is greater than a predetermined threshold value can be considered as the same topic. Based on this, we can find the same topic in different time slices, and can analyze the evolution process of the topic according to the changes of the topic keywords and key entities, so as to achieve the purpose of topic tracking.

Topic relationship analysis. The relationship between the topic and the entity is very important. Different entities included in different topics are often different. We can statistics the frequency of entities appearing in the topic to obtain the relationship weight of the entity and the topic.

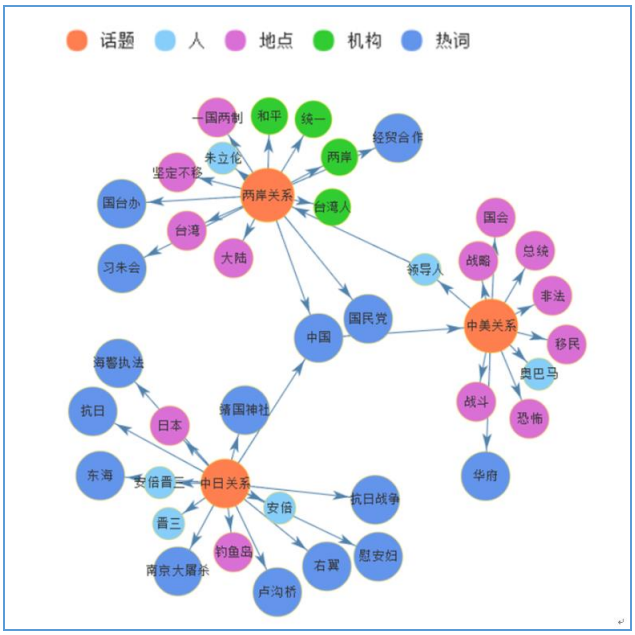

Figure 9. An Example of Topic Map

In a topic, the relationship between the entities can solved through an intermediate variable. Probability distribution relationship between the entity and the topic may be expressed by the relationship between the words with same name and the topic in the LDA model [8]. We can calculate the similarity according to the cosine distance, and can calculate the relationship weight between two named entities using the formula (1). Assume that there are K topics, where e_i, e_j represent two entities respectively, $p(z_k|e_i)$ represents the probability of e_i belonging to the topic z_k, and $p(z_k|e_j)$ represents the probability of e_j belonging to the topic z_k.

$$Sim(e_i, e_j) = \frac{\sum_{k=1}^{K} p(z_k|e_i) \times p(z_k|e_j)}{\sqrt{(\sum_{k=1}^{K} p(z_k|e_i)^2)(\sum_{k=1}^{K} p(z_k|e_j)^2)}} \quad (1)$$

The relationship between the entity and the keyword in the topic is shown in Fig.9, where we can clearly see which concepts (including people, places, institutions, hot words, *etc*.) often appear in the news topic map and the mutual relationship between the news entities.

4.5. User Management

User management of base station service platform needs to provide user information management functionality and self-service functionality for users. User management module mainly includes user information management, user self-service and unified user management. Specific requirements are as follows.

4.5.1. User Information Management: User information management provides user basic information management functionality for system administrators described as follows.

- *User list*. View system user list, send notifications to designated users, and stop the specified user from accessing the base station.
- *Online users*. View instant messages of online users, and make designated users kicked out of login status, so that he needs to log in again when he visit.
- *Statistical analysis*. It refers to statistical analysis of user behavior, such as the use of operating system, columns, user access status of topics, access status of news, and accusation status of news.

4.5.2. User Self-Service: User self-service provides that ordinary users can modify their own basic information, change and retrieve their own password.

4.5.3. Unified User Management: Unified user management provides user information synchronization functionality including collection and release of user information, as well as release functionality of real-time user, and provides user information batch import, export and synchronization audit information query functionality for system administrators.

5. Nonfunctional Requirements of base Station Service Platform

Nonfunctional requirements of base station service platform needs to consider the following aspects.

- *Performance requirements*. The system should be normal operation and timely response to the operation in a multi-user concurrency accessing, and timely response to the operation.
- *Portability*. The system needs to support Single Sign-On functionality of all web system based on B/S architecture, and can be normal operation in different operating systems.
- *Workability*. It needs to ensure that the user interface is friendly, guiding users to use.
- *Scalability*. The protocols interacted between our system and external systems must be the common industry protocols, such as ODBC/JDBC, LDAP, XML, *etc*.
- *Security*. The system is a user authentication system. Transmission security must be guaranteed, which can use HTTPS to guarantee transmission security. For the security of password storage, encryption algorithm can be used to encrypt password and then to put these encrypted password stored into the database.
- *Maintainability*. The system needs to have better maintainability. When adding or removing a module should not affect the other system modules being used.

6. Conclusions

The Internet has evolved into a power-law scale-free network, and Internet's emerging paradigm is undergoing profound change to share content-centric information. The basic transmission model of Internet based on the bandwidth allocation is difficult to effectively support large-scale information sharing for wide area. The issues such as traffic explosive growth, quality of service, energy saving, content security and credibility, have plagued the development of today's Internet architecture, and have also challenged the researches on future Internet architecture. Therefore, the secondary structure called Broadcast-Storage network based on the

radiation-copy model came into being. Following the Broadcast-Storage concept, combining the advantages of broadcast radiation and content storage, we propose a technological framework of base station server content delivery based on broadcast-storage architecture, give emphasis on the involved key technologies and application algorithms, design and implement a base station server content delivery system based on broadcast-storage architecture, which is docked with the integrated broadcast control platform, integrated management-control platform and clients, completing application demonstration in important typical business. The above researches have laid a solid foundation for a more fine-grained deconstruction of Broadcast-Storage network architecture, in-depth study of Broadcast-Storage network and guidance for constructing Broadcast-Storage network system.

References

[1] Conti M, Chong S, Fdida S, *et al*. Research challenges towards the future Internet, Computer Communications, 34(18), pp.2115-2134 **(2011)**
[2] Barabasi A L. Scale-free networks. American science, 5, pp.50-59 **(2003)**
[3] Barabasi A L., Albert R. and Jeong H. Scale-free Characteristics of Random Networks: The Topology of the World Wide Web, Physics, 281, pp.69-77 **(2000)**
[4] Li Youping. The secondary web of knowledge embodiment. Engineering Science, 4(8), pp.8-11 **(2002)**
[5] Li Youping. Research of Complementary Architecture Network. Journal of Southwest University of Science and Technology, 21(1), pp.1-5 **(2006)**
[6] McGuinness, D.L. Question answering on the semantic web. IEEE Intelligent Systems, 19(1), pp.82-85 **(2004)**
[7] Bristol E H. Swinging Door Trending: Adaptive Trend Recording. ISA National Conference Proceeding. 45, pp.749-753 **(1990)**
[8] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 3, pp.993-1022 **(2003)**

Authors

Cong Liu. 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. 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. 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. 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.