TELKOMNIKA, Vol. 11, No. 8, August 2013, pp. 4766~4773
e-ISSN: 2087-278X

Website Resource Monitoring Platform Supporting Tibetan and Uyghur Language Based on Semantics

Lirong Qiu
School of Information technology, Minzu University of China, Beijing, China
e-mail: lirongqqq@163.com

Abstract

With the development of the Internet and the increasing minority language websites, people of ethnic minorities begin to browse the news, comments and other content on the Internet according to their own interests. At present, there is still no websites to provide the Chinese–Tibetan or Chinese-Uyghur co-occurrence search engine in China. In this paper, a platform for Tibetan and Uyghur website monitoring is proposed. The functions and characterizes are proposed in detail.

Keywords*: natural language processing, search engine, semantic ontology, Tibetan language*

1. Introduction

China is a multi-ethnic country, the population of the national minorities making up about 8 percent of the country's total. The government has always respected and guaranteed ethnic minorities' right to use and develop their own spoken and written languages. The presence of different races, languages and cultures has made the world so much more interesting and the languages and dialects of minority races do have their own small roles to play. And many of the minorities have radio, film, television, books, and websites in their own languages.

With the development of the Internet and the increasing minority language websites, people of ethnic minorities begin to browse the news, comments and other content on the Internet according to their own interests, and obtain the useful information they need. The resources in websites become so huge that it's really difficult and time-consuming to collect, read and evaluate them [1].

For example, according to the 29th China Internet Network Development Statistical Report released by China Internet Network Information Center, by the end of 2011, there have been 8,820,000 netizens in Xinjiang province, with 40.4% in popularizing rate and 7.7% in growth rate, and the popularizing rate of Xinjiang province ranked 4th among 31 provinces across the whole country. And most people in Xinjiang province use Uygur language. As seeing through this statistic data, in order to understand the Uygur people's need in time and maintain the stability of Xinjiang region, it will be of great importance in analysis of Internet public opinion in Uygur language.

Tibetan/Uyghur language resource monitoring plays an important role in minority language resource research. It's intended to adopt the intelligent analysis calculation and network information search to solve the public opinion monitoring, data analysis, news propaganda and other application problems.

The key techniques of this platform mainly concentrated on:

(1) Tibetan/Uyghur website search based on the Internet;

(2) Collection, download, transcoding and other techniques in pre-processing Tibetan and Uyghur language network resource [2];

(3) Tibetan/Uyghur language automatic segmentation, annotation and automatic classification;

(4) Distributed management software of Tibetan/Uyghur language resource;

(5) Real-time network monitoring in Tibetan/Uyghur language resource;

(6) Chinese-Tibetan or Chinese-Uyghur website search;

Received January 17, 2013; Revised May 24, 2013; Accepted June 3, 2013

(7) Tibetan/Uyghur language text analysis, including hot word and sensitive word statistics.

There is no unified format to express different points of view. In order to distinguish whether the opinions are positive or negative, it should be based on the language semantics to analyze them.

As a relatively new field in China, the semantic orientation analysis and emotion recognition has great practical value. Thus it is necessary to overcome technical difficulties and make a breakthrough in Tibetan/Uighur language research. At present, there is still no websites to provide the Chinese–Tibetan or Chinese- Uyghur co-occurrence search engine in China.

In the work of Tibetan and Uyghur language processing field, some scholars have researched many applications in the website.

2. Methodology

In recent years, with the development of social and cultural diversification, the minority language websites have increased rapidly in quantity, rich in content, and involving culture, education, politics, economy and other aspects. They have overlapping parts with Chinese or English websites in content, such as earthquakes, social movements and so on. Moreover, they have their own distinguishing features, such as publishing festivals and activities of the minority nationalities.

It's common that most minority language websites, such as Tibetan/Uighur websites, publish information in both Chinese and Tibetan/Uighur, and allow users to register or login with Chinese and Tibetan/Uighur name. Such websites usually contain both Chinese web pages and Tibetan and Uighur web pages. It's also possible that the alternate use of Chinese and minority characters appear in the same webpage. For example, a professor in Xinjiang University can release both Chinese and Uighur information with the Chinese name "艾克拜尔•吐尔逊" or Uighur name "ئەكبەر تۇرسۇن".

This monitoring platform is intended to provide Chinese–Tibetan and Chinese-Uyghur co-occurrence search engine.

(1) Vicious information monitoring in Tibetan/Uyghur network resource

Some domestic separatists and international anti-China forces make full use of the Internet to beautify themselves, distort the facts and incite the masses. And then they conduct the ideological and cultural infiltration into our country, carry out activities to subvert regime and destroy unity, which heavily threaten the people's security and stability. Take Tibet independence an example, as Indian SIFY.COM reported, "The Tibetan government-in-exile have made the Internet a strong weapon to attract Tibetans against China in the past few years. A large number of websites advocating Tibet independence have been set up. Some Tibet separatists living in the Himalayan border keep in touch with Tibetans through specific websites and broadcast their actions".

According to the characteristics of vicious information in Tibetan websites, define its type and scope which need monitoring and construct vicious information library.

According to the vicious information characteristics and Tibetan grammatical features, design different regex, programme to achieve vicious information monitoring system based on regex.

(2) Sensitive word discovery in Tibetan/Uyghur network resource

Sensitive information means the information which is harmful to the interests of the state and the people, such as the company's commercial secrets, the country's political and military secrets, netizens' personal privacy and so on. Sensitive words are vicious vocabulary that the country or institutions have restrictions on their usage.

It has been developed to collect and analyse the commonly-used sensitive words in Tibetan/Uyghur, establish the Tibetan/Uyghur sensitive word list, classify them and define the sensitivity levels. Automatic addition and retrieval tracking of Tibetan/Uyghur sensitive word has been realized. Based on the sensitive word list, it can been effectively discriminated whether the network carrier contains the specified sensitive information during retrieval and found the related data to provide support for users to achieve the sensitive information monitoring.

As Figure 1 shows, Tibetan/Uyghur language resource monitoring platform can be divided into nine modules.

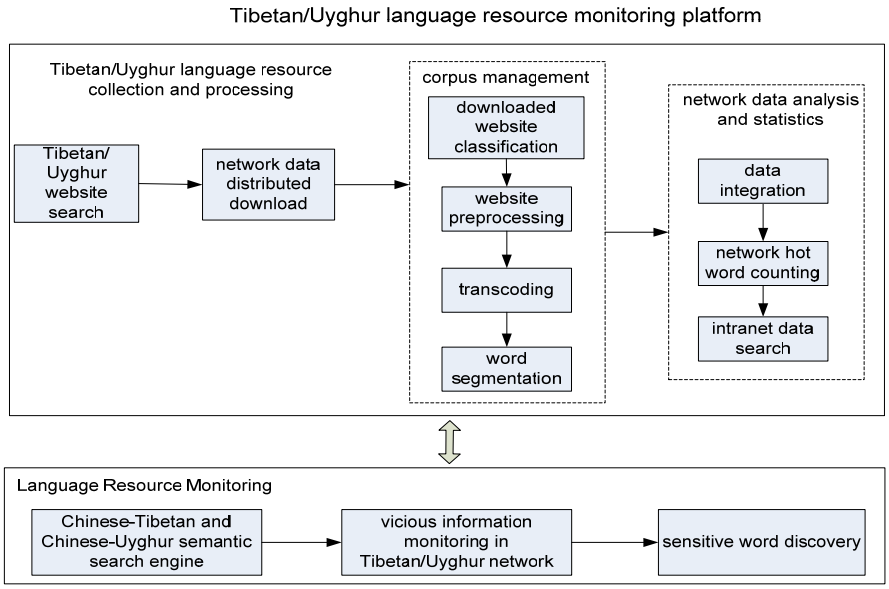

Figure 1. Magnetization as a function of applied field, note how the caption is centered in the column

3. The Characteristics and Innovations

3.1. Characericstics of the Platform

No matter which language we choose to release information, characters are just the media to convey semantics, and semantic expression is the core part in communication. The public opinion monitoring in ethnic minority users is able to combine public opinions expressed in different languages, such as Chinese, Mongolian, Tibetan, Uyghur and so on, find the focus issues and compare them synchronously. As a pivotal problem to be solved, it is required to master some important techniques in ethnic minority language processing and public opinion monitoring.

As the fourth medium following the newspaper, radio and TV, the network has been recognized as one of the main media reflecting the social public opinion. The primary sources of public opinion in websites include news commentary, BBS, blog, Rich Site Summary (RSS) and so on. Network public opinion expression is characterised by its efficiency, multivariate information and interactive way, which has lots of advantages that the traditional media can not match. According to its openness and virtuality, public opinion on the Internet has the following characteristics:

(1) Directness. Via BBS, news comments and blog sites, Internet users can express their own views directly to others. Network provides the people with a more smooth way.

(2) Burstiness. The formation of the public opinion on the Internet is very quick. A hot issue and some sentimental views can lead to a blockbuster in public opinion.

(3) Prejudice. Due to the hiding of netizens' identity and the lack of restrictions and effective supervision, the network becomes the space for netizens to vent their negative emotions, such as setbacks in real life, one-sided understanding of social problems and so on. Therefore, it's easier to arise vulgar and gray remarks on Internet.

Network public opinion monitoring makes full use of search engine and data mining. Through automatic acquisition and processing, sensitive word filtering, intelligent clustering classification, subject test, project focusing and statistical analysis on basis of website content, it can satisfy different needs of agencies towards network public opinion supervision and management, and generate corresponding analysis reports [3]. It provides an analytical basis for the management team to get a comprehensive grasp of public opinion dynamically and make the right decision to guide them.

3.2. Key Techniques and Innovation Points

Chinese/Tibetan/Uyghur language resource monitoring platform has several key techniques and innovation points as followed.

(1) Innovation in the semantic search results

Through the comprehensive understanding towards search semantics, this system can combine the results in different search engines, fully resolve them, and show the top 50 search results. According to the special requirements in semantic monitoring, the results can be reordered. For example, reordering by the accepted time and showing relevant information. It has been developed in semantic search results to facilitate the Chinese/Tibetan/Uyghur language resource monitoring [4].

(2) Innovation in semantic search results display

This system can automatically implement comprehensive search and show the results in the semantic search. For example, when a user input "西藏山南", Tibetan "ཞུ་ནས་" can be shown as the search result in the same webpage, which is very convenient for users to see. Similarly, when the user input "ཞུ་ནས་", "西藏山南"can also be shown. It acts as a beneficial attempt in multilingual resources monitoring.

(3) Innovation in background support during semantic search

It's necessary to integrate a large number of Chinese, Tibetan and Uyghur corpus during semantic search. The sources of corpus are varied, and the collection process is full of difficulties, so it needs to mobilize all forces to gather them. After the collection, corpus is integrated to facilitate the computer processing, such as integrating data with various formats into the same database by using all kinds of text processing techniques. The number of corpus is as high as 390000, therefore, speeding up the search processing rate will improve the retrieval efficiency.

(4) Corpus encryption and retrieval utilization technology

Because of the high cost and the great practical value of corpus, it's required to encrypt the data. Then even if the database leak, nobody can decrypt the data stored in the corpus resources. At the same time, the decryption should be expedient so that the semantic matching search can be carried out during the semantic retrieval. Therefore, this system adopts AES encryption/decryption technology, and uses the method of separately storing "Chinese plaintext-Tibetan ciphertext" and "Tibetan plaintext-encryption Chinese", which offers the convenient corpus retrieval function in guarantee the corpus safety.

(5) Hot search words statistical techniques

When customers use the monitoring platform for search, the platform can record the statistical number of search keywords and demonstrate the hot search words in Tibetan and Uyghur, which can directly reflect the monitoring role it plays. While lacking boot data and launching "Cold Boot", keywords can be specified by the platform for user-friendly retrieval. Meanwhile, it can realize the real-time statistics to adapt to the change of users' intention, which is convenient for administrators to process data and show current hot search words to users.

(6) Complete foreground and background resource monitoring display platform

In this system, the display style of foreground is similar to Outlook, allowing users to be clear of the function and module layout of the website. The platform also provides background data support technology, through which website administrators can easily add and manage the data, analyse the user's accessing log to acquire their intention, and then further modify website data for enhancement of the website usability.

4. Fuctions of the Platform

This platform uses JSP + Servlet as the main technical framework and constructs a dynamic website to provide convenience for administrators in data management and the users in retrieval. The technical protocol is mainly divided into the following two parts.

4.1. Foreground Display

It contains the following six parts as shown in Figure 2.

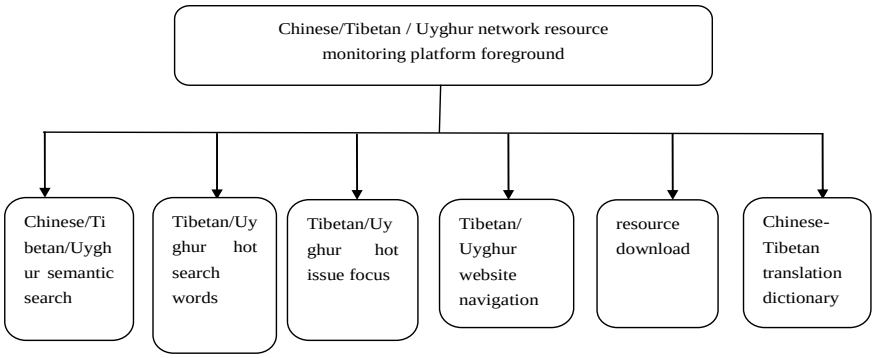

Figure 2. Foreground Function of the Monitoring Platform

Use flash for scrolling pictures in foreground and display customizable images in the website.

Use Jquery in sub-module switching to provide dynamic effect and good interactive experience

Use CSS to set the whole website layout and be compatible with IE6, IE7, IE8, Chrome, FireFox, Safari and other browsers, which is easier for customers to access.

4.2. Background Control Mechanism

It contains the following five parts as shown in Figure 3.

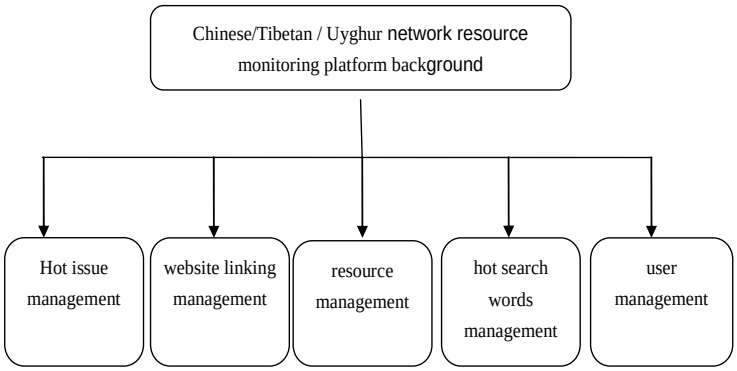

Figure 3. Background Control Function of the Monitoring Platform

There are Addition and List two sub-modules at the bottom of each module, which is convenient for managing every module.

Use Frameset to divide the background into several modules, and thus can be centrally controlled in the background to facilitate the operation of the administrator.

Be compatible with IE6, IE7, IE8, Chrome, FireFox, Safari and other browsers, which is management barrier-free.

5. Semantic Information in the Platform

With the digital development, Tibetan and Uyghur information processing flourishes in various fields, covering letter, word, phrase, sentence, chapter and other multi-level technology research. Display, coding and input techniques have also been solved perfectly. Tibet University, Northwest University for Nationalities, Qinghai Normal University and other research

institutes have made a fruitful achievement in Tibetan information processing, and made relative breakthroughs in letter, word, phrase and sentence processing.

5.1. Semantic analysis in Tibetan and Uyghur

For those minority languages, such as Tibetan, Uygur, Mongolian, comparing to Chinese, there are huge differences in grammar, pronunciation, spelling, vocabulary, which is increased the difficulty of interoperability between those languages with regard to information retrieval, information extraction and automatic translation.

The concept itself defined in the dictionary is not ambiguous; it can be associated with the real-world entity or object uniquely and accurately. Word (binary data for the computer) is only a medium of semantics, and semantics is the core and critical part of communication.

According to the language research history, research on the semantic aspect originated from philosophy. While the semantic study as an independent discipline was in nineteenth Century, subsequently, studies on the semantic were carried out in linguistics, logic, natural language processing and other fields [5].

Semantic analysis is a way of formal representation that can reflect sentence meaning according to the syntactic structure of the sentence and the sentence in each notional word. Semantic analysis can seize the essence of the sentence through the changing syntactic form.

HowNet is a common sense knowledge base which is put forward and established for Chinese by the teacher named Zhendong Dong, which has collected more than 100,000 Chinese words, referring to more than 30,000 concepts. HowNet puts emphasis on reflecting the concept's commonness and individuality, which greatest feature is to describe various relations among the concepts and the attributes of concepts [6]. It describes not only the semantic features of Chinese content words' concepts, but also a variety of semantic roles acted by Chinese function words.

However, the great differences between Chinese and those minority languages lead to great difficulty in automatic mapping from HowNet to Tibetan and Uyghur.

Both Mongolia and Uyghur belong to Altaic language family, and are characterized by agglutinative languages. While writing, their letters are linked together and vary in written forms. Tibetan belongs to Sino-Tibetan language family. While writing, the letters are superimposed with each other, and there are no specific delimiters between words or sentences.

Tibetan name has only the first name and no surname, which is quite different from the Chinese name. In addition, it's very common to use natural species in Tibetan names. For example, "ཉི་མ" can be found in nature as "the sun", and also can be used as a Tibetan name "Nima". Therefore, Tibetan name recognition do not have the same characteristics with the Chinese name which can use the characteristics of the surname or even the first name for identification

Uyghur has the following characteristics: the special writing direction which is from right to left, automatic selection in Uyghur text, Uighur grammar deformation; interval between Uyghur words; rich Uyghur affixes, especially verbal affixes. According to the current statistics, there are more than 13,000 Uyghur verb affixes and around 30 noun affixes.

5.2. Named Entity Recognition based on semantic ontology

Named entity recognition in social network relationship has two difficult parts. The first one is how to identify the person or organization entities in text and extract nodes in a social network schema from them. The second one is how to extract the relationship from the named entities.

The combination of grammatical structure and semantic mode is a powerful way to improve performance in natural language information processing system. This project will combine Tibetan and Uyghur syntactic analysis and semantic ontology to achieve Tibetan and Uyghur named entity recognition in social network.

On the basis of the semantic ontology, the technical route can be divided into three parts, (1) creating the social network relationship ontology, (2) named entity recognition, (3) entity relationship mining, just as the following diagram shown.

The steps are shown as Figure 4.

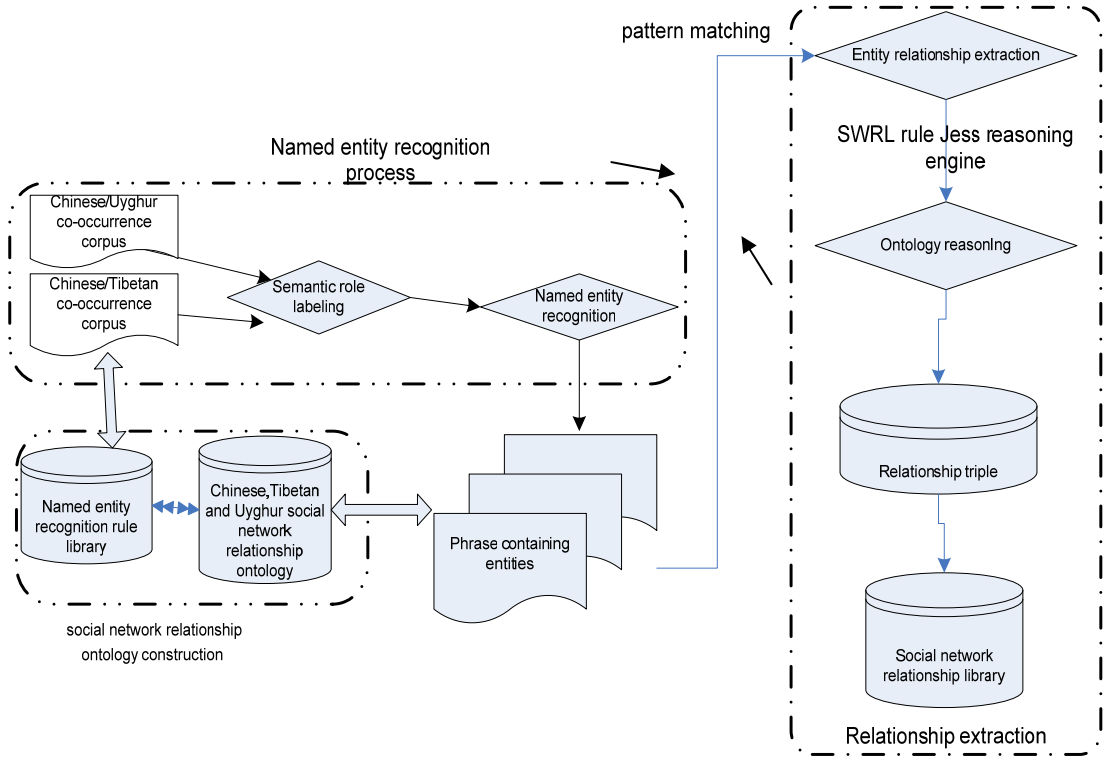

Figure 4. The Process of Named Entity Recognition

(1) Create social network upper ontology, and adapt it in Tibetan and Uyghur; clearly describe ontology based on social network relationships, and explicitly incorporate the semantic ontology structure in Chinese, Tibetan and Uyghur.

(2) Create Tibetan, Uyghur named entity recognition rule library; according to the characteristics of Tibetan and Uyghur vocabulary, let the linguists formulate rules of named entity, and the computer technicians input rules into the library.

(3) Conduct semantic role labeling with both Tibetan and Uyghur syntactic features; according to the Tibetan and Uyghur syntactic characteristics, mark the corpus with the methods of semantic role labeling, which is of great help in named entity recognition.

(4) Conduct the relationship mining between named entities; extract relationships from the events that took place between the entities, construct a network schema using the search term, and finally visualize it.

6. Related Work

Network transmission has the attributes of freedom, interactivity, openness and hiding. It can not only bring a lot of progressive, healthy and profitable positive information, but also contains some reactionary, superstitious and low ranked negative information. Therefore, the research on network public opinion analysis and forecast has become an important actual demand [7, 8].

There are some research results in English, Chinese and other languages [9-12]. Coincident with the development of digitalization, the information processing research in Tibetan and Uyghur has also gained rapid progress, covering such aspects as characters, words, phrases, sentences and chapters [13]. And in Tibetan language, Long Congjun provided semantic relations of nouns in paper [14], but his work did not touch how to retrieve the dictionary and automatic group the nouns based on similarities.

Many researchers develop their Tibetan-Chinese electronic dictionary, but as far as we known, there hasn't been any research in Tibetan and Uyghur language processing on the level of website application, such as search engine and online dictionary.

7. Conclusion and Future Work

With the development of social and cultural diversification, the minority language websites have increased rapidly in quantity, rich in content, and involving culture, education, politics, economy and other aspects.

Such websites usually contain both Chinese webpages and Tibetan and Uyghur webpages. How to search the content described in Chinese, Tibetan or Uyghur is a tough problem. Chinese-Tibetan and Chinese-Uyghur search engine on basis of semantic ontology is applied in this platform, http://cmli.muc.edu.cn/ResourceManage/.

In the future, the platform will add some functions include:

(1) Query of single Tibetan entity, that is query Web information analysis table;
(2) Query of relations between entities;
(3) Evolution analysis and query for entity group's dynamic tendency;
(4) Auto discover constant to sensitive entity;
(5) Entity group develops, varies regularly and forecasts, analyses report etc.

The work of this paper is a part of our ongoing research work, which aims to provide an open website for supporting such applications, such as search engine for Chinese-Tibetan or Chinese-Uyghur. Various experiments and applications have been conducting in our current research. Future work includes how to increase the semantic annotation information in Tibetan and Uyghur language, on the basis of the information of predicate that have been marked.

Acknowledgments

Our work is supported by Program for New Century Excellent Talents in University, and the National nature science foundation of China (No. 61103161) and the "Science Fund for Youths" project of Minzu University of China (No. 1112KYQN39).

References

[1] Guha R, McCool R, Miller E. *Semantic Search*. Proceedings of the 12th International Conference on World Wide Web, ACM Press. 2003; 700–709.
[2] Qi LS. A New Resource Information Integrating Method in Semantic Concept Networks. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(3).
[3] Moldovan DI, Mihalcea R. Using Wordnet and Lexical Perators to Improve Internet Searches. *IEEE Internet Computing.* 2000; 4: 34-43.
[4] Jiang Di. The Classification of Tibetan Verbs and Relative Patterns based on Semantics and Syntax. *Journal of Chinese Information Processing*. 2005; 20(1): 37-43.
[5] Purandare A, Pedersen T. Sense Clusters - Finding Clusters that Represent Word Senses. *Proceedings of the National Conference on Artificial Intelligence*. 2004; 19: 1030-1031.
[6] Zhang PY. A HowNet-Based Semantic Relatedness Kernel for Text Classification. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(4).
[7] Li Yuqin, Sun Lihua. Hot-word Technology Based on Internet Public Sentiment. *Journal of Chinese Information Processing*. 2011; 25(1).
[8] Heydon A, Najork M. Mercator: A scalable, extensible Web crawler. *WorId Wide Web.* 1999; 2(4): 219-229.
[9] Satoshi S, Nagao M. *Toward Memory-based Translation*. Proceedings of the 13th International Conference on Computational Linguistics (COLING-90). Helsinki, Finland. 1990; 3: 247-252.
[10] Chang CL，Chen DY, Chuang TR. *Browsing News Groups with a Social Network Analyser*, Proceedings of the Sixth International Conference on Information Visualization. 2002; 2: 750-758.
[11] Sharon A. Caraballo. *Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.1999; 120-126.
[12] R Rubinstein, A Bruckstein, M Elad. *Dicionaries for Sparse Representation Modeling*. Proceedings of the IEEE. 2010; 98(6): 1045-1057.
[13] Lirong Qiu, Congjun Long, Xiaobing Zhao. *A Joint Approach for Building a Large Tibetan Corpus with Syntactic Parsing and Semantic Role Labeling.* 2012 Fifth International Conference on Intelligent Networks and Intelligent Systems. 2012; 12: 232-235.
[14] L Congjun, Z Xuewen. *The Research of the Nominal Semantics Relations.* Proceeding of the 2rd Minority Language Processing Conference for Young Scholars. 2008.