International Journal of Security and Its Applications
Vol. 10, No. 12 (2016), pp.141-152
http://dx.doi.org/10.14257/ijsia.2016.10.12.12

Research on Information Security Audit Base on Semantic Web Ontology and Improve Vector Space Model

Zhanjiang Wang[1], Shuoning Wang[2] and Ling Wang[1]

[1]*Harbin University of Commerce, Harbin 150028, China*
[2]*Heilongjiang Institute of Tourism, Harbin 150086, China*
[1]*zhongj_w@126.com*, [2]*1517889500@qq.com*

Abstract

Now the network has become the main source of the information where people gain from. However there are various network information, include healthy and helpful information, at the same time; also include the bad and useless information. How to protect the security and quickly and exactly find the user need from this vast information has become the hot research. This article use the improve vector space model to filter bad information and use the semantic web technology to build a computer ontology. Aim at this area to information retrieval, through this ontology to handle users' input, thus to improve the recall level and the precision rate.

Keywords: *vector space model; semantic web; domain ontology.*

1. Introduction

With the rapid development of Internet technology, network had a blood of information. While, due to the characteristics of openness, interaction and sharing of network information made the information content complicated, good and bad. There was useful, healthy information, and at the same time, there was also a large number of violent, reactionary adverse information [1]. According to the statistics by China Internet Illegal Information Reporting Center from December 1st to 31st in 2011, online reporting platform and telephone hotlines received a total of 57040 various public reports. Most of the information came from the Network. Users use the search engine (SE) to get information from this Information Source. But use the traditional SE; it's hard for users to find the information, which was needed, rapidly and accurately. This was mainly of several reasons here:

Firstly, it was not accurate enough to understand users' question. So there was much uncorrelated information in returned results, which made hard for users to find the useful information. This problem is generally the shortcoming of retrieval system. The processing of retrieval keywords is seemingly simple, in fact, very complex. The extension of the retrieval keyword using the information on the surface of the word or semantic information, the retrieval results will be different.

Secondly, processing of the information content is mostly used one side of a coding process of pretreatment technology or some kind of full-text analysis techniques, only to reflect the content;

Thirdly, users' questions could not be fully consistent with the content of the Information Source, and could not guarantee that the content and user mentioning an exact match [2]. The cause of this phenomenon is that users themselves are not clear their

ISSN: 1738-9976 IJSIA

wanted information or users enter the retrieval key words that are not in conformity with what they want.

Only to solve the above problems, could the user retrieval information safely, accurately and rapidly. Existing research show that base on Semantic Web Ontology and improve vector space technology was one of the methods to solve these problems.

In the light of the problem, that the scale of the current information retrieval system is too large and too complex, its function cannot meet the rapid growth of mobile services in the process of exchange of information on the accuracy of the requirements, paper [3] proposed a mobile business information retrieval method based on ontology. They analyzed the features of the mobile business, put forward a suitable model for mobile business and constructed the mobile business domain ontology by giving out a feasible construction plan on the basis of the existing research results. Finally, the system is realized, and the experiment is carried out.

In order to solve the low efficiency problem of traditional information retrieval model and provide a better quality of information retrieval to satisfy users personalized information retrieval request, paper [4] established ontological user personalized model based on domain ontology and proposed personalized information retrieval method. Based on the analysis of the key technologies and algorithms of model running, they expatiated on the operational mechanism of personalized information processing. At the last, the experimental results show that their method has better recall and precision rate that compared with the traditional retrieval system.

About the research of information retrieval, paper [5] used different method with the literature mentioned above. They consider that different users have different background and different knowledge, their hobby and interest is also different. Those differences leads to their information needs are also different. They consider different user visit on the Web to go to only a specific subset of the resources which is a particular area; but most of the retrieval system uses keywords to search the information, these system use the same standard interface and the same retrieval methods for all different users; those system devised the same search results for all different user. They established a model for user profile based on ontology, after considering these problems that users can not accurately express themselves and they cannot express information on specific areas which they interest in and need.

The above mentioned reference is the research on personalized information retrieval based on ontology. In the face of Internet vast information ocean, personalized retrieval is very important and effective strategy. But at the same time of information retrieval, filtering out harmful information is also a part of the performance of information retrieval system. Paper [6] made a brief review information filtering history, followed detailed describes the classification of information filtering system, evaluation, application and information filtering and other information processing technology the difference. They also introduced the domestic and foreign bad information filtering the present situation of the research; it will be used as reference. Paper [7] proposed a method of information filtering based on natural language understanding and neural network technology. They established three rule base of the domain knowledge, designed a prototype of the filter system through the analysis of practical application. The established system has the filtering and monitoring capabilities of particular content information on the network, and achieved good practical effect.

In domestic ontology used in information retrieval research started late, but has been significant progress. But in foreign countries, ontology's applied research in retrieval information had been relatively mature. The research subjects were mainly a variety of

research institutions and large companies, so there produced some large ontology project. In this paper, we take use of the improve vector space model to filter the bad and useless information, use the semantic web technology to build a computer ontology and information retrieval. Aim at this area to information retrieval, through this ontology to handle users' input, thus to improve information retrieval the recall rate and the precision rate.

2. Ontology Technology

2.1. Ontology

Ontology was originally a branch of philosophy, the nature of existence of objective things, after being used in many fields, for the concept and organization of knowledge sharing and reuse. The ontology was the abstract and standardized description of the domain knowledge, to describe the semantic relationships between concept and concept, with strong formal and logical reasoning ability [8].

According to the different division method, ontology could be divided several categories. In accordance with the ontology formal attainment, formal ontology can be divided into semi-formal ontology, the structure of non-formal ontology and completely non-formal ontology; in accordance with the research level division can be divided into top-level ontology, domain-specific of ontology, task ontology and application ontology; in accordance with the research topic, ontology can be divided into general or common sense ontology, knowledge representation ontology, domain ontology, linguistics ontology and task ontology; in accordance with the logical reasoning ability, can be divided into heavyweight ontology, middleweight ontology and lightweight ontology; in accordance with the level of detail and areas dependent on the degree, can be divided into reference ontology and shared ontology [9]. Domain-specific of Ontology was to implement the research of information security audit.

The so-called domain-specific of ontology was to describe the kind of subject concepts, including the concept in the discipline, the concept of property, and the relationship between concepts and attributes, and relationships constraints. As knowledge had significant areas of the characteristics of the domain ontology can be more reasonable and effective representation of knowledge.

2.2. Ontology-building method

There were a lot of methods and technologies to build ontology. Currently, the ontology construction method using the more:

One was the "TOVE", developed by the University of Toronto Enterprise Integration Laboratory, and used first-order logic predicates to integrate;

The second one was Skeletal Method, designed to build enterprise ontology, it only provided guidelines for the development of ontology;

The third one was Meth ontology, dedicated to build chemical ontology construction;

The fourth one was IDEF5, obtained through the use of graphic language and details of the description language on the objective existence of concepts, attributes and relationships, and formal as the main framework of the ontology.

The fifth was Seven Steps Method, mainly used to build the domain ontology. The five of ontology building methods from different aspects of the analysis were shown in Table 1.

Table 1 showed that the Seven Steps Method had the better aspects in life cycle, related technology, and application field and method details.

2.3. Second-order headings

A wide range of ontology editing tool and some are commercial products, some schools and government on the outcome of the ontology technology topics, as well as a small portion of the software tool. Among them, the commercial ontology editing tools included independent editing tools specifically designed to build the ontology in a particular area, as well as an integral part of the transmission integrated enterprise solution designed for the business software group, these latest products have the language classification ability and stochastic analysis capabilities, helped to extract useful information from the disorder in the information content [10].

Table 1. Analysis Of Ontology-Building Method

Building Method	Life Cycle	Related Technology	Applied Field	Method Details
TOVE	Fuzzy	Uncertain	One	Less
Skeletal Method	No	Uncertain	One	Seldom
Methontology	Have	Incomplete	Dedicated	Detailed
IDEF5	No	Uncertain	Multiple	Detailed
Seven Steps	Fuzzy	Have	Multiple	Detailed

Which belong to the common ontology editing tools, including OntoEdit, Ontolingua, OntoSaurus, WebOnto, OilEd and Protégé. Both OilEd and Protégé open up source code, Ontolingua, OntoSaurus and Protégé have strong scalability, especially Protégé has extended API interface. In addition to the Ontolingua not have the reasoning ability, the others have.

A variety of ontology editing tool has its own superiority, but due to the limitations of IT development, the current ontology editing tool limitations. After many compare filter, Protégé was selected, which structure was more complex but more open, easily extended, support better the standards and support of OWL. In addition, its interface is very simple and friendly. Therefore, by Protégé ontology editing tool to create ontology.

3. Build Ontology and Improved Vector Space Model Information Filter System

In order to cover all areas of knowledge Ontology was too difficult or even impossible, therefore, a more realistic approach was to create a field of ontology, and use it to solve the specific information retrieval problems. The computer field, for example, the establishment of a computer domain ontology.

3.1. Second-order headings

Ontology-Building was a very complex system which needed the correct development of thought and development tools to assist. So it’s need to do the preparatory work for the following four aspects.

Ontology formal description language of choice: Ontology language allows the user to write a clear domain model, the concept of formal description of a direct impact on the expression of the ontology model and scalability. So, it should meet the following

requirements: a well-defined syntax; a well-defined semantics; efficient reasoning support; sufficient expressive power; convenience of expression. Formal ontology language had a lot, the mainly were RDF, RDF-S, OIL, DAML, OWL, KIF, SHOE, XOL, OCML, Ontolingua, Cycl and Loom。After the comparison, selected OWL (Ontology Web Language).It was the standard recommended by W3CSemantic Web Ontology Language, its advantages were Web resources that described the object to decidable logical reasoning, and semantic features. Such areas of OWL ontology built at the same time have a good ability to perform powerful reasoning ability

The choice of ontology development tools: At present, at home and abroad has many mature ontology development platform software to choose from. Above has been described in detail, chosen Protégé as a development tool. It was developed by the Stanford University Medical Informatics Research Group, was a knowledge modeling tool based on open architecture open source Java environment. Extended OWL plug-in is currently the most powerful OWL ontology building tools. Not only has good scalability and a simple and flexible user-customizable interface also supports graphical ontology editing mode, support for database storage model, based on the OWL database of more than the development model and support logic detection. Which greatly facilitate the ontology construction, learning and problem solving

To determine the ontology construction method: There are many ways of ontology building, shown in Table 1, after comparison, using a seven-step method for the construction of computer domain ontology.

Knowledge in the field of collecting: Domain ontology construction requires a lot of expertise in the field. Computer developers had rich ontology knowledge and a strong development capability, but on the specific domain knowledge knew very little, it was very difficult to establish the ontology model for specific areas. To build ontology should have the participation of experts in the field. Authority areas of knowledge are as a reference.

3.2. Second-order headings

According to some of the more recognized ontology construction project, summary of the following steps to build domain ontology.

Determine the areas of expertise and scope of the ontology: That clearly establishes the reasons, the scope and the range of users of the ontology. Building computer domain ontology therefore should have the following two objectives:

①Use of ontological thinking organization and description of "Computer" in the field of knowledge;

②Build the logical and scalable ontology library.

Built computer domain ontology should have the following two fundamental characteristics:

①Simple and good concept hierarchy:

②Resources based on the notion of entailment axiom scalability.

List the important terms in the ontology: List all the concepts in the field, to enumerate all the possible attributes, each attribute has a corresponding attribute value.

Establish the conceptual structure of the target ontology: There were several possible approaches: top-down method, bottom-up method and synthesis method. Consider

automatically to the current ontology acquisition generation technology is not perfect, so we choose the synthesis method to build the computer ontology.

Defined properties: The property depicted the internal structure of the concept. Any one class of all subclasses inherited the properties of that class. According to the conceptual model was based on object-oriented features make full use of the class inherits the attributes defined. Sub-concept of common property defined in the parent concept, the sub-concept inherited all the attributes of the parent concept, and then defined their own unique attributes. This reduced the attribute redundancy, and enhanced the expressive power of the conceptual model [11].

Create the instance of the class: To define an instance of the class need to identify a class, create an instance of the class and add the attribute value.

Ontology construction project focused on instance performance, so instantiated was the workload of the largest and most tedious part throughout the development process. Protégé can automatically generate OWL syntax library files; manual creation of a large number of instances in Protégé is still very cumbersome. Because of manual entry, it was prone to error. So to all aspects of testing to discover the contradictions of the ontology concept definitions and instance attributes associated with error conditions to ensure the correctness of the ontology in logic.

3.3. Improved vector space model

In the vector space model, each document is represented as feature vectors, namely:

$$p(d) = (word1, W1(d); word2, W2(d); \cdots; wordi, Wi(d); \cdots; wordn, Wn(d)) \quad (1)$$

Where word *i* is the *i*th keyword in the document d; Wi(d), the weight of term i in document d. Vector space model was applied to information filtering theory, the hypothesis space vector B to be badness information template, the vector space D to be the template of waiting filtering information:

$$B=(W1,W2\cdots Wi\ \cdots Wn) \quad (2)$$

$$D=(W1,W2\cdots Wj\cdots W\ n) \quad (3)$$

The Similar level between the two models showed with $Sim(B,D)$:

$$Sim(B,D) = \cos\theta = \frac{B \bullet D}{\|B\| \bullet \|D\|} = \frac{\sum_{i,j=1}^{n} W_i W_j}{\sqrt{\sum_{i=1}^{n} W_i^2 \sum_{j=1}^{n} W_j^2}} \quad (4)$$

Compare $Sim(B,D)$ to determine whether adverse information, when the similarity is large, very similar to the need to filter out the information to be filtered with badness template, conversely, be filtered information and adverse template certain degree of difference can be returned to the user to browse the threshold requires a combination of the actual situation repeatedly adjusted to achieve a better filtering effect.

Structure the badness model showed with B(d), improve the vector model B(d) to:

B(d)=(word1,(-1)t*L1,W1(d);word2,(-1)t*L2,W2(d);…;wordi,(-1)t*Li, Wi(d);…;word n, (-1)t*Ln, Wn(d)) (5)

Of which: word i is the i-th keyword in the badness template d; Wi (d), the weight of term i in the badness template d; Li is the degree of similarity of keyword i replace with the word synonymous; new entrants (-1)t refers to the tendentious judgment of term i [12].

The improved method of $Sim(B,D)$ calculation as follows:

$$Sim(B,D)=\cos\theta=\frac{B\bullet D}{\|B\|\bullet\|D\|}=\frac{\sum_{i,j=1}^{n}\{[(-1)^t\times L_i]\times W_i\}\times\{[(-1)^t\times L_j]W_j\}}{\sqrt{\sum_{i=1}^{n}\{[(-1)^t\times L_i]\times W_i\}^2\sum_{j=1}^{n}\{[(-1)^t\times L_j]\times W_j\}^2}} \quad (6)$$

Calculate the two space vector template similarity value filtered by small value.

3.4. Build badness information filtering system

The literature [12] had detailed introduced the design and implementation of the badness information filtration system, where not described in detail in this paper. Use the filtration system conjunct with the computer ontology, complete to the information from the filter retrieval process. The badness information filtering system flow chart was shown in Figure 1.

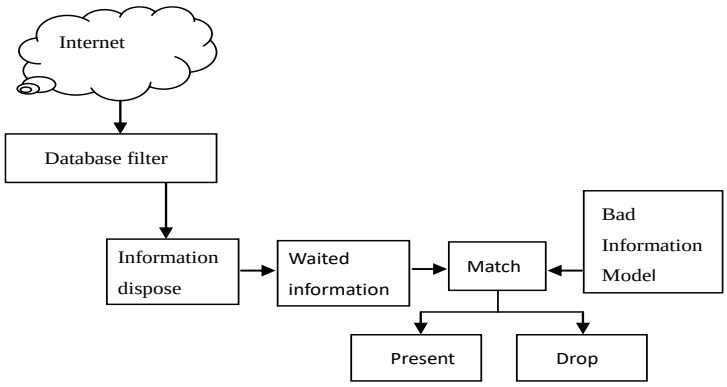

Figure 1. Filtering System Flow Chart

4. System design and implementation

4.1. Information security audit system process design

The information security audit system flow chart was shown in Figure 2.

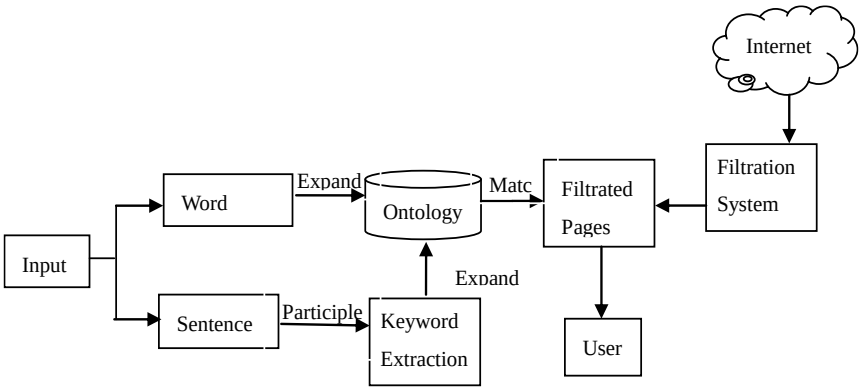

Figure 2. Information Security Audit System Flow Chart

The main flow of information security audit system is as follows: on the one hand, the network's website to pass the information filtering system for filtering, the other hand, the user through the search box to enter the problem, pretreatment user input, extract keywords. The key word extends through the established domain ontology. Expansion matches with the filtered pages and then feedback to the user.

4.2. Information audit and the key technology

4.2.1. Word segmentation: Information retrieval recall rate and precision rate is not high, one reason for the user's input can not exactly match the contents. Users the vast majority is not professionals, so the user's input that computer does not fully understand, and even diametrically opposed. Therefore, the first pretreatment user's input: generally user's input can be divided into two kinds, the one is input for words and the other one is input for sentence. When user input is the word, directly to the ontology library for keyword expansion. When user input for the sentence, it need to extract keywords. Use the parsing technology for the user input sentence.

In this paper, the ICTCLAS(Chinese Academy of Sciences Institute of Computing Technology, Chinese Lexical Analysis System) [13] on the user input sentence segmentation. ICTCLAS is one of China's current high levels of lexical analysis, word accuracy rate of more than 98%. Key features include the Chinese word segmentation; speech tagging; named entity recognition; new word recognition; supports both user dictionaries. Its main advantages are the following:

①combination optimal performance;

② calculate the theoretical framework of a unified language;

③full support application development in a variety of environments;

④changed by needs, tailor-made.

So this paper chose this system to parse the user input sentence.

For example, a user enters "计算机的特点是什么？" Segmentation results as follows:

"计算机/n 的/u 特点/n 是/v 什么/r ？/w ".

But this only put sentences into a word, all these word cannot as keyword for information retrieval, which also includes the particularly high frequency words but does not make sense, such as "的、了、吗〞etc, which often to drown out the real meaningful words. One of the ways to solve this problem is to create a stop list to filter out these words. However, all the remaining words are not all contain useful information, so the

need to choose some of the large amount of information contained in the words of the query feature can be better. One of the most effective is that some nouns, verbs and so on. These words are added to the search terms in the table.

4.2.2. Keywords expansion of technology based on ontology: Literally, taking into account the issues raised by users, often contain a relatively small vocabulary, so moderate expansion the problem keywords with semantic knowledge. Keyword expansion is including expansion of the expansion of the areas of vocabulary and synonyms. Among them, the word of the extension of the field, including two extensions:

Firstly, the areas of vocabulary alias expansion, there will be a lot of words in the vocabulary of the computer field has an alias, such as "计算机" there is a "电脑"and "PC" as an alias for this vocabulary expansion can increase the recall rate;

The second is the expansion of close ties between the areas of vocabulary. Have a close relationship between many words in the computer field. Query keywords according to the word in the ontology to extend the relationship between the areas of vocabulary data mining method of digging out large amounts of data. The following three formulas show that the calculation of the relationship between the two words [14].

$$\mathrm{Support}(X_k \Rightarrow Y_k) = \frac{\left|\{t_i \mid X_k \cup Y_k \subseteq t_i\}\right|}{n} \tag{7}$$

$$Confidence(X_k \Rightarrow Y_k) = \frac{\left|\{t_i \mid X_k \cup Y_k \subseteq t_i\}\right|}{\left|\{t_i \mid X_k \subseteq t_i\}\right|} \tag{8}$$

$$RLA(X_k \Rightarrow Y_k) = \sqrt{(\mathrm{Support}\ (X_k \Rightarrow Y_k))^2 + (\mathrm{Confidence}\ (X_k \Rightarrow Y_k))^2} \tag{9}$$

Among them, the support means the ratio of transaction accounts for all transactions contain Xk ∪ Yk, confidence is defined as when Xk appear, Yk emergence of probability. The words' relationship RLA is through a combination of support and confidence to calculate. Accordingly the field of vocabulary expansion, setting up the threshold exceeds the threshold value of the words added to the sequence of query words.

5. Experiment and Simulation

My eclipse platform, tested on windows 7 environment, and to download a large number of pages of information templates to filter out bad information, bad information filtering system. Use the seven-step method to build computer domain ontology in Protégé editing tools assisted, and through a series of processing user input, extract the user search keywords, keyword expansion through ontology, after the match, and then back to the users.

In order to validate the test, the system downloaded a lot of web pages containing the bad information and bad information filtering system filters, the filter part of the effect shown in Figure 3.

Table 2. The efficiency comparison of two method

Retrieval Method	Retrieval Results				
	LC	AD	DR	P/%	R/%
Convention	220	270	160	59.2	72.7
System's	220	230	220	95.6	100

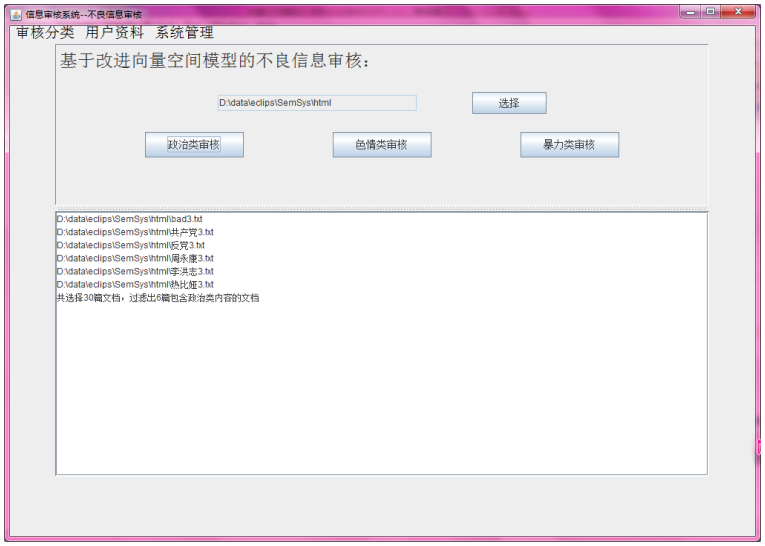

Figure 3. Filter Bad Information Part of the Renderings

Retrieve filtered through this system, the effect of greatly improving the retrieval recall rate and precision rate. To see the results more intuitive, the system test were compared with the conventional keyword search. Conventional keyword search methods and the efficiency of the system to retrieve contrast, is shown in Table 2.

6. Conclusion

Bad filtering of information by improving the technology of the vector space model, and the establishment of the computer domain ontology to retrieve specific information in the field, Effective composition of a complete information security audit system, effectively filter out bad information at the same time, be extended through the processing of user input after the establishment of the ontology and reasoning, thus improving retrieval recall rate and precision. Simulation results show that the effect is obvious. .

Acknowledgements

The Research Sponsored by Science and Technology Department of Heilongjiang Province, Science and technology project (GZ07A101; RC2009XK010001), the Nature Science Foundation of Heilongjiang province (F201243) and the Scientific Research Foundation of Education Bureau of Heilongjiang Province (12511127).

References

[1] Chen Pinglin, Liu Ting, Zhu Weiping, Tang Yao. The Application of LBS in the Scenic Visitor Management. journal of Hangzhou Normal University (Natural Science Edition). Vol. 12 No. 5 Sep. 2013: p467-473.
[2] Wang Shiwei. On Information Security, Network Security and Cyberspace Security. Journal of Library Science in China. March, 2015 Vol. 41. No. 216: pp72-84.
[3] Mao Yimei. Mobile Business Information Retrieval Based on Ontology (In Chinese). Journal of Wuhan University of Technology（Information & Management Engineering）. 34(6), 2012, pp: 699-703.
[4] Liu Yisong, Pan Chao. Research of Personalized Information Retrieval Model Based on Domain Ontology(In Chinese). Wireless Communications Technology. 22(3), 2013, pp: 29-33.

[5] Yin Hongli. The personalization information retrieval engineering research based on ontology(In Chinese).Journal of Shandong Institute of Light Industry(Natural Science Edition), 22(2), 2008, pp: 76-79.
[6] Zhou Tianqi. Information Filtering Reviewed in the Network Security (In Chinese). Microprocessors, 32(5), 2011, pp: 30-34.
[7] Liu Jian, Lv Guoying, Sun Jia. Based on the Semantic Recognition of Adverse Tendency Information Filtering System Design and Implementation(In CHinese). Netinfo Security, 2012(10), pp: 13-16.
[8] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993, 5(2): 199-220.
[9] M. Fernández López. Overview Of Methodologies For Building Ontologies [J]. Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5) Stockholm, Sweden, 1999:8.
[10] M. Grüninger, M. S. Fox. Methodology for the Design and Evaluation of Ontologies [J]. Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95, Montreal, 1995:298.
[11] Natalya F·Noy, DeborahL McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology[D]. 2001.8. http://protege.stanford.edu/publications/ontology-development/ontology101.pdf, Accessed: 2008:2
[12] http:‖Protege.stanford.edu [EB]. Accessed: 2008.2.
[13] Jena2 Interface Support [EB/OL]. http://jena.sourceforge.net/I nterface/
[14] WANG Shu-da, LI Hai-long, LIU Zhan-qing. Research of picture archiving and communication systems [J]. Journal of Harbin Institute of Technology. Vol. (17). Sup. 2010:86-89.

Author

Zhongjian Wang, Ph.D., professor. His main research interests include natural language process, Chinese sentence paraphrase, Chinese word segmentation and Information retrieval etc.