Research on Ontology Construction and Information Extraction Technology Based on WordNet

Hu Hua
School of Information Management
Wuhan University
Wuhan City - 430072, China
henryhu@whu.edu.cn

JDIM
Journal of Digital Information Management

ABSTRACT: Introducing ontology in the field of information extraction can effectively improve the performance of information extraction. Ontology of college teachers' resumes are constructed in the form three layers of ontology framework structure on the basis of exploring the related technologies and criterion for building ontology, the Racer inference engine is used for realizing the consistency and accuracy detection. WordNet similarity calculation method is improved on the basis of this, so that manual collection method can be integrated into WordNet semantic similarity, thus substantially improving this method. The test shows that the result precisions based on WordNet ontology construction and information extraction both see significant improvement, which sufficiently shows the feasibility and validity of the method.

Subject Categories and Descriptors:
I.2.10 [Vision and Scene Understanding]: Information Extraction; **I.4.10 [Image Representation]**

General Terms: Ontotology Construction, Information Processing

Keywords: WordNet, Similarity Calculation, Ontology Construction

Received: 11 November 2013, Revised 1 February 2014, Accepted 7 February 2014

1. Introduction

Effective information can be obtained from complicated information by means of information extraction technology; meanwhile, the effective information can be stored, which provide convenience for repeated use of the information [1-3]. In this article, the problem of adaptability of information extraction is the uppermost problem that information extraction faces. To solve this problem, this article describes domain knowledge by ontology information, which can effectively improve the performance of information extraction. This article constructs the ontology by means of WordNet-based semantic similarity. Meanwhile, related ontology–based theories are used for extracting the information. Words similarity is further subdivided through calculation of the extracted information, and is promoted to be the similarity calculation of synonym set. Similarity calculation of WordNet is also improved on the basis of this. The accuracy of the calculation is significantly improved as proven by examples.

2. Methods for ontology construction

This article aims at constructing ontology of college teachers' resumes, in which protege is used as the tool for ontology construction. As shown in Figure 1, ontology structure is a top-bottom ontology information structure, where the lower the information position is, the more detailed the information will be. Seen from Figure 1, in the modelling of college teachers, the ontology of the teachers' information is on the top, followed by various sub-information bellow it, with each piece of sub-information consisting of multiple events, thus forming an integral model of teachers' information. As shown in Figure 1, this article constructs the three layers of ontology of college

teachers' resumes based on the features of the domain.

In the above-mentioned figure, the name of Ontology is Domain; meanwhile, we need to represent the ontology by PCV, and represent the sub-events in the ontology by Event Concept, represent each single event by Event, and abstract each event as Event Concept. A college professor's resume may include lots of information, such as the sub-events like life experiences, education backgrounds etc. An Extended Concept can be formed by combining multiple concepts, which can be named as extended concept layer. For example, the extended concepts of college teachers' study experience include information such as "*graduating school*", "*inception*" and "*major*" etc. Substantially speaking, Extended Concept and Event Concept are both ontologies.

While constructing ontology, this article adopts the top-bottom confirmation method, in which the ontology domain is confirmed firstly, then the events in the domain are confirmed, and finally, the information in the events is confirmed. Meanwhile, they are conceptualized to confirm the relationships among them.

3. Analysis on college teachers' resumes

As shown in Figure 1, the concept of the event is allocated first. The event information is summarized by analyzing the webpages introducing the personal information college teachers in college websites, which is as follows:

1). Person Information consists of name, date of birth, specialty and post.

2). Contact of college teachers consists of classes such as telephone number, address, and E-mail etc.

3). Educations of college teachers mainly consist of enrolling time and graduating time of a school, major and inception etc.

4). Works include the classes such as specific posts, working time, and name of departments etc.

5). Interests include the classes such as the teachers' research field etc.

6). Publications include two classes of books and papers.

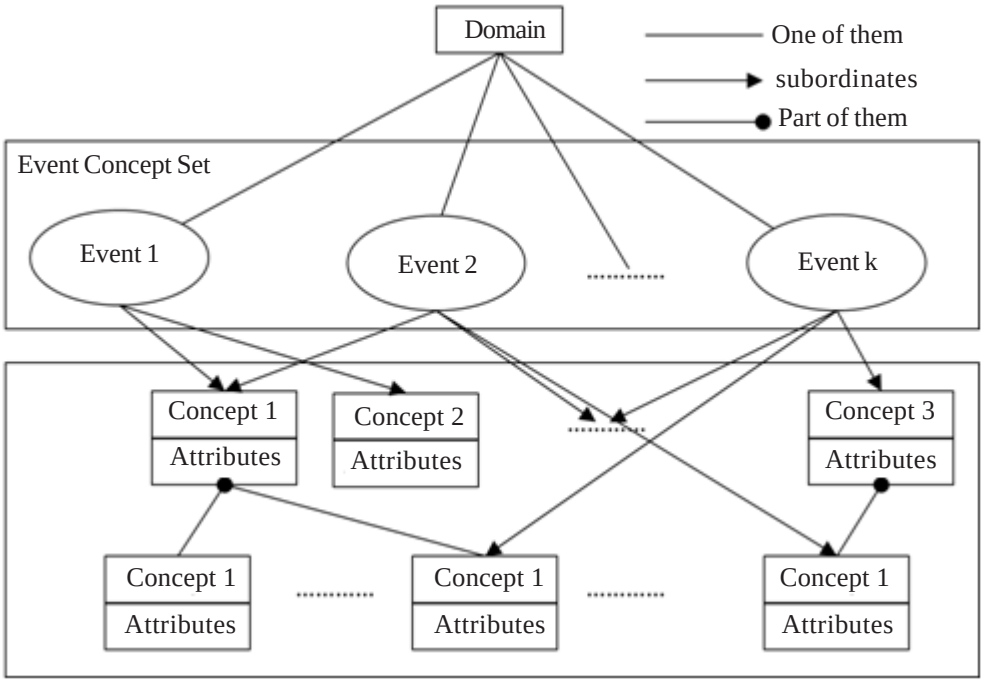

Figure 1. Ontology structure diagram

7). Awards include the awarders, awarding time, awarding reason etc.

8). Social Activity includes the activity time, name and specific work that the teacher takes charge of.

In a piece of ontology consists of multiple sub-events which form a piece of ontology. Meanwhile, the relationships among various information concepts are established. The following diagram shows the concepts in the ontology and the relationships among them constructed in this article.

3.1 construction of ontology

In this article, Protégé tool for ontology construction is introduced for the construction work to save the ontology in OWL language. The graphical interface of protégé tool is as shown in the diagram, in which all the classes are represented by the oval circles on the left, and the attribute of class "*Education*" is represented by the dotted line squares on the right. As referred to in the above paragraphs, the data type of date has been included in the data type property during the process of ontology construction. Therefore, Date is deemed as a data type property under a related concept instead of a single Class or concept.

Since the specific specialty range is larger, it's difficult to

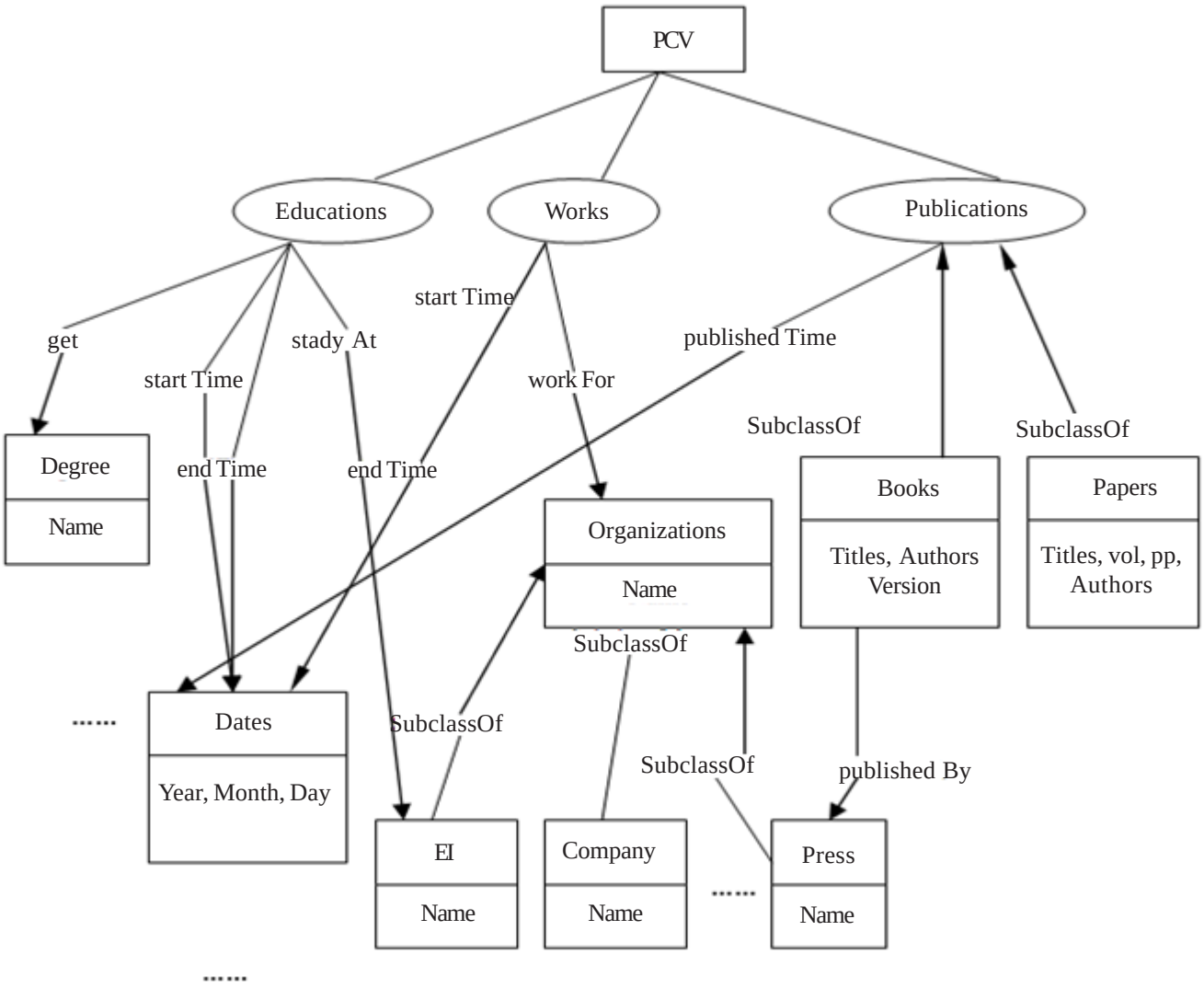

Figure 2. Relationships among the concepts of ontology

divide the domains; even being subdivided, it will not completely represent the professors' information, so these events are not further subdivided. Moreover, the sub-event of address in Contact is also not subdivided, because many contacts are not complete.

3.2 Detection of ontology

Currently, there are not so many reference materials related to ontology, and research in this respect is relatively lagging, so there are not so many available resources [4-6]. This article detects the constructed ontology based on Racer inference, and proves the consistency of ontology construction.

This article detects the consistency of the ontology system based on Racer inference. This system is a knowledge frame that can directly convert ontology into logic mode, so that we can conduct inference on the ontology system by the logics generated from Racer [7-8]. Racer inference carries out inspection from the process of concept relationship, concept consistency, the relationship of the upper and lower layers of the concepts, and whether there is contradiction among concepts etc. As shown in Figure 3, the PCV ontology constructed by this article is detected by means of Racer inference. Racer inference shall detect the ontology one by one in the sub-processes of detection proposed in the text above. From Figure 3, we can clearly see that the detection results on the left and on the right are completely consistent, which indicates that the ontology constructed in this article are consistent.

3.3 Access of Ontology Instances based on Improved WordNet

As can be seen from the constructed ontology above, it has a clear structure, with a logical relationship between various concepts. As for proper nouns in the ontology, it will be quite helpful for identification of traditional relationships to identify the special instances.

In this paper, the two methods to access instances are separately manual statistics and semantic similarity calculation based on WordNet. For example, instances for the concept Specialty are accessed through manual statistics. A total of 1321 English specialty lists is taken from websites designed for university information

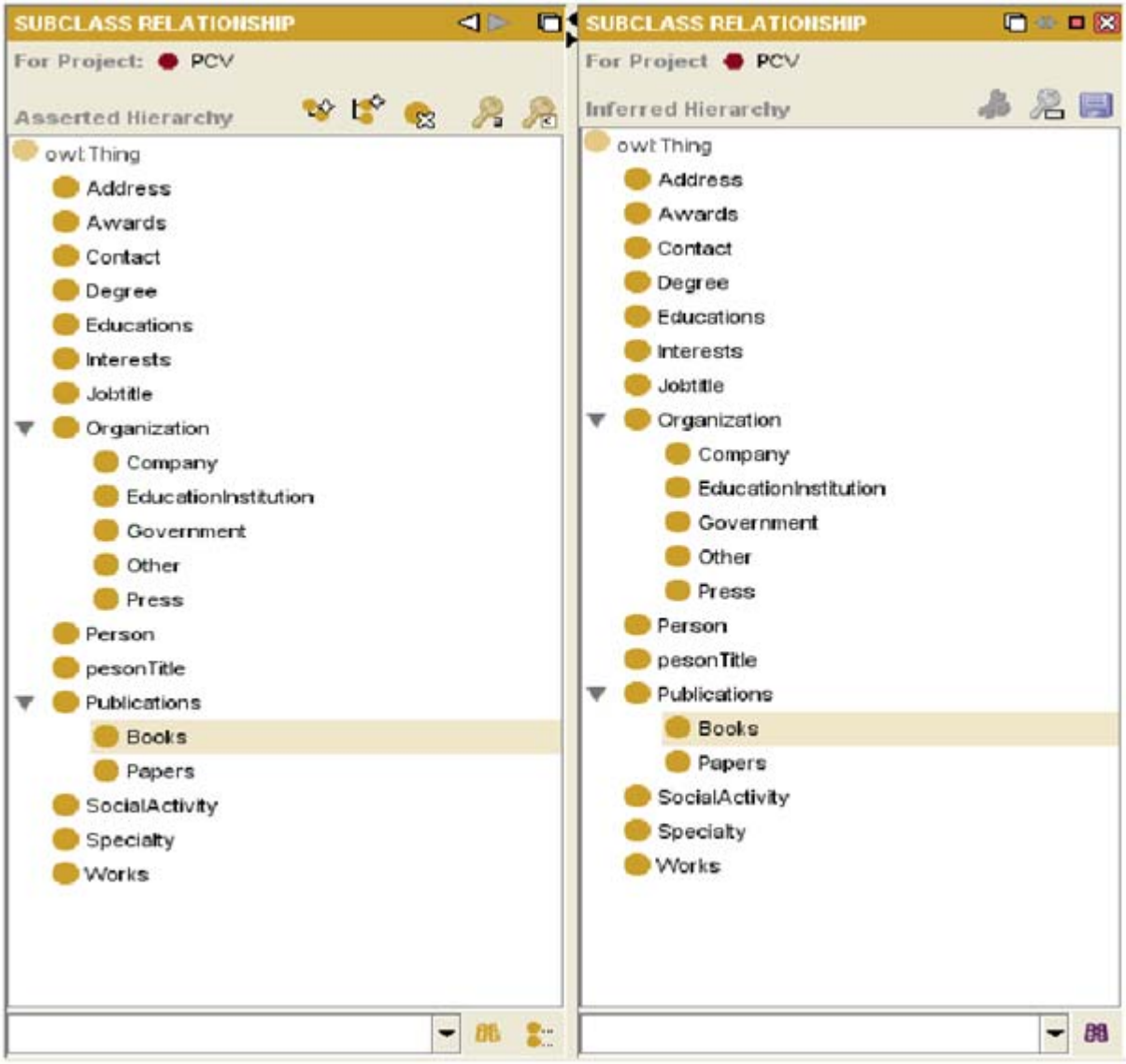

Figure 3. Realization of ontology detection

introduction. In addition, instances for Job Title have been collected in GATE's Gazetteer component, totally 1445. However, for lack of statistics for some instances, it is not feasible to perform manual statistics for Ontology instances. Therefore, in this paper, similarity transformation methods are used to access instances before manual screening, which helps improve the efficiency of instance access greatly.

Word Net indeed is an English word recognition system. In nature, it is an English vocabulary library, which can retrieve and identify the English ontology and is widely used in natural language processing [9-10]. In the library, it performs expression in the form of semantic network. There are five types of words in the WordNet, separately verbs, nouns, adverbs, adjectives and function words. They are distinguished by the identity nodes, whose relations are divided into antonyms, synonym, local and overall, morphology. Morphology shows processing between words [11-12].

(Synset) denotes the set of synonyms. In WordNet, ID index number for each Synset is unique, when the semantic relationship defined by Synsets can be deemed as a pointer existing between them. In WordNet, the word with several meanings can be expressed as a word with multiple Synsets. In this paper, a set of synonyms is used to denote the relationship between the upper and lower noun concepts, of which the former ones represent abstract meanings, while the latter ones are the underlying meanings.

In WordNet, sets of synonymous are used to integrate all words, and perform similarity conversion. The semantic similarity calculation formula is shown as follows:

$$sim(W_1, W_2) = \underset{S_{ij} \in s_1, S_{ij} \in s_2}{MAX} (sim(s_{1_j}, s_{2_j})) \qquad (1)$$

In the formula, the similarity of the words $W1$ and $W2$ is donated by Sim (W_1, W_2), and the similarity of Synset S_{1i} and S_{2i} is donated by Sim (S_{1i}, S_{2i}). In the word set, we use S_1 (S_2) to represent W_1 (W_2) with word concentration.

In Ontology, when a specific Synset is accessed for a uniquely determined concept, that's, statistics on specific meanings for a concept (attribute) in a special ontology are performed, similarity transformation is carried out on the basis, to improve the quality of instances. Semantic similarity access mode in the WordNet is improved to address such actual needs.

Specific approach is shown as follows. Firstly access instances for some concepts (or attributes) manually. Then, perform similarity transformation to words in the instances and concepts based on the WordNet semantic similarity calculation method, and record the Synset with the largest similarity transformation value, i.e. with the most key words in the ontology. After accessing specific keywords, perform training on similarity between words in the text and the keywords, calculated using the following method:

$$sim(Key, W) = \underset{s_i \in s}{MAX}(sim(K, S_i)) \qquad (2)$$

In the formula above, specific meaning of the keyword Kev in the ontology (Synset) is donated by K, and Synset set contained in W is represented as S.

In this paper, the similarity between the two Synsets is accessed using the following equation.

$$(sim(K, S_i)) = -\ln \frac{\alpha Dis(K, S_i) + \beta \Delta Depth}{K} \qquad (3)$$

In the formula above, $Dis(S_i, S_j)$ means the distance between S_i and S_j; Δ Depth represents the distance between S_j and K, as well as the distance difference between K, S_j and Synset; K is a constant; as for constants α and β, $\alpha + \beta = 1$. What is to be emphasized here is that there are a lot of similarity calculation methods. As for Synset calculation, at the very beginning, only the distance between two Synsets in the semantic WordNet tree is involved, gradually taking into account the depth and density of Synset in the semantic tree. In this paper, since there has been a certain value about the depth and density, the later development of depth and density will be in a disadvantageous situation, which is a flaw in this paper.

4. Testing and Analysis

Take the concept Degree as an example. Following tests are performed in this paper. Wherein, values of α, β and K are separately 0.1, 0.9 and 2. Take 0.5 as a threshold value. On this basis, firstly, extract the words related to basic information of 80 university professors. Perform similarity transformation between it and Degree. The test results are shown in Table 1.

	total number of instances extracted	Number of instances	Number of accurate instances extracted	Accuracy rate of instance extraction	Recall rate
Test result of Formula (1)	83	24	20	24.10%	83.33%
Test result of Formula (2)	23	24	20	86.96%	83.33%

Table 1. Test results for instance abstract

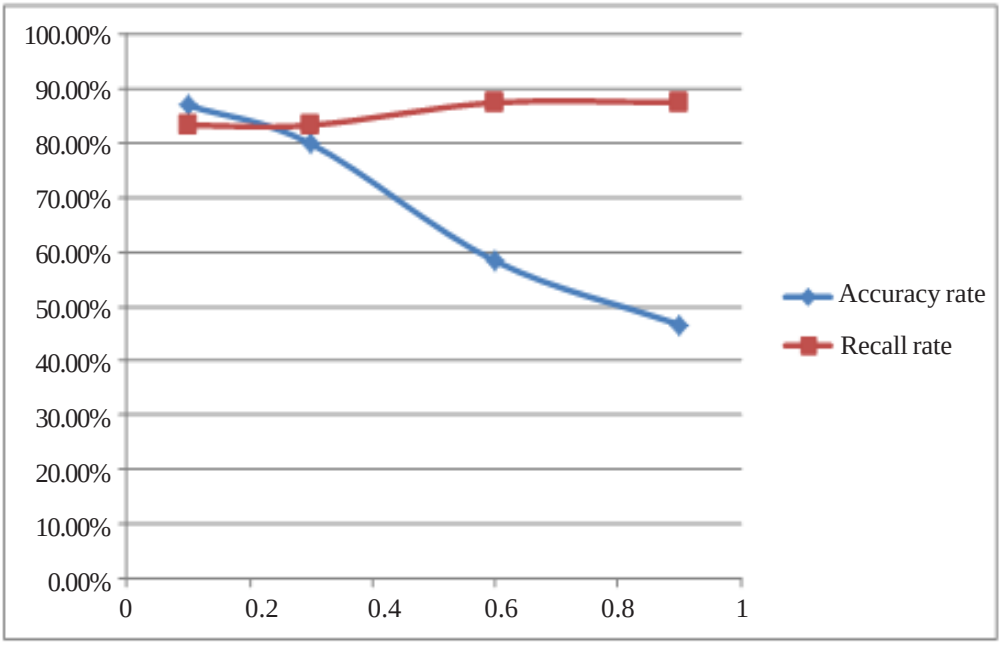

Figure 4. Coefficient of evaluation

α	Accuracy rate	Recall rate
0.1	86.96%	83.33%
0.3	79.92%	83.33%
0.6	58.33%	87.50%
0.9	46.67%	87.50%

Table 2. Coefficient changes with the result of extraction

During the tests, the extracted words are restored, so we can guarantee that each word is used only once. In this paper, the value of α is obtained through the tests. When α is in the range of 0.1 to 0.9, As shown in figure 4, we can achieve the best performance value when $\alpha = 0.1$.

It can be concluded from the results above that extraction accuracy can be effectively improved when using the formula (2). Meanwhile, in the upper relationship, five wrongly extracted words are accessed through tests. As for the lower relationship, the accessed words are consistent with our requirements. Therefore, to increase the access performance, the key words and instances should go through instance analysis.

5. Conclusions

In this paper, on the basis of discussing technologies and standards related to the ontology construction, 3 levels are adopted to construct the ontology of teachers' curriculum (PVC). Firstly, determine the event concept within the ontology, then extend the concept further, access the relationship between the extended concept ant the original concept, and achieve consistency and correctness testing with Racer inference engine. This paper combines similarity transformation and manual collection, accesses specific examples of concepts within the ontology through tests, and ultimately constructs a relatively complete ontology library combining concepts and instances.

References

[1] BAO Hong, LIU Hong-zhe. (2005). Web Services Based Virtual Antique Museum Architecture [J]. *Journal of System Simulation*, 17 (6) 1412-1417.

[2] XU Honglei, et al. (2010). Research on Chinese Event Abstraction Technology within Automatic Identification Technology Events [J]. *Mind and Calculation*, 4 (1) 34 -44.

[3] JIA Sai, QIAO Hong. (2011). Research ON Web Information Extraction Based on Ontology and Implementation of Ontology Construction [J]. *Library Studies*, (5), 31- 33.

[4] ZHONG Wei-jun, WEI Ji-cai. (2005). Study on Semantic Information Model of Crisis Situation [J]. *Journal of System Simulation*, 17 (10) 2367-2370.

[5] LIU Feng, GU Junzhong. (2009). Design and Implementation of Metadata Management Application System [J]. *Computer Engineering*, 35 (11) 29 -31.

[6] Alvarez, M., PanA, Raposo, J., et al. (2008). Extraction lists of data records from semi-structured web page [J]. *Date & Knowledge Engineering*, 64 (2) 491-509.

[7] REN Zhongchen, XUE Yongsheng. (2007). Web Structured Data Extraction Based On Web Label [J]. *Computer Science*, 34 (10) 133 -36.

[8] WEI Shouzhi, Zhao Hai, WANG Gang, et al. (2005). Complex System Situation Assessment Model and Its Ontological Method [J]. *System Simulation Journal*, 17 (5) 1200- 1202.

[9] Du, T C., Li, F., King, I. (2009). Managing knowledge on the Web-Extracting on-tology from HTML Web [J]. *Decision Support system*, 47 (4) 319-331.

[10] Liu Pengho, Che Haiyan, Chen Wei. (2010). Survey of knowledge extraction technologies [J]. *Application Research of Comruter*, 27 (9) 3223-3226.

[11] TAN Juan, LI Bo-hu, CHAI Xu-dong. (2006). Technology Research of Extenxible Modeling and Simulation Framework-XMSF [J]. *Journal of System Simulation*, 18 (1) 96-101.

[12] LIU Wei, MENG Xiaofeng, MENG Wei. (2007). Deep Web Data Integration Research Overview [J]. *Journal of Computers*, 30 (9) 1475-1489.