Research of Marine Organism Ontology Semi-Automatic Construction

Jing Xiong[1], Yuntong Liu[1], Jianfang Wang[2,*] and Yihua Lan[2]

[1]School of Computer and Information Engineering, Anyang Normal University, Anyang, Henan, 455000, China; [2]School of Computer and Information Technology, Nanyang Normal University, Nanyang, Henan, 473061, China

Abstract: Marine creature is the most active subject in marine ecological system. The interactions among them are complex and important. Ontology as a sharing concept formal model can well express these relationships. In order to solve the role problem of marine organisms in these complex problems protean relationships, we introduced the role theory of Hoze and used it to describe the relationships. On this basis, we chose Text2Onto to realize the automatic acquisition of the ontology elements first, and then through the artificial confirmation and used Hozo to edit, thus completed the semi automation of marine organism ontology construction. Experimental results demonstrate this semi-automatic ontology building method can reach high accuracy and recall in marine organism field.

Keywords: Hozo, instance extraction, marine organism ontology, ontology learning, relation extraction, text2Onto.

1. INTRODUCTION

The 21st Century is the century of ocean. Marine ecosystem plays an important role in the global ecological system, and the marine biological activity and the interactions among marine organism have a crucial impact on the marine ecosystem. The marine ecosystem is a very complex system. There are inevitable inner links among all the ocean elements: the non biological factors and biological factors. Effect of marine ecosystem in global change is enormous. On the one hand, global change will be evident in the ocean; on the other hand, changing marine ecosystems will obviously cause or slow down the global change. In the marine ecosystem, marine organisms are the active agents, therefore study the motion law of marine organisms and the interaction is very necessary. Ontology as a formal, explicit specification of a shared conceptualization captures the knowledge of related fields, and provides common understanding in the domain [1]. It is a proper way to represent the interaction and relationships among marine organisms.

However, marine organism interaction relationship is not immutable and steady, even the role of the two sides of interaction can change each other. For example, nitrogen, phosphorus and other marine nutrients have an important impact on the marine ecosystem. On the one hand, the rapid growth of phytoplankton can provide a large amount of oxygen in the growth process by the process of photosynthesis. Phytoplankton provides abundant food resources for the consumers in the ocean, so advanced biological can grow, providing abundant fishery resources for mankind; on the other hand, if the nutrient coast excessive inflows, the ocean will present the state of eutrophication, algae sudden massive proliferation, lead to the occurrence of red tide and algae bloom phenomenon. In addition, when a large number of red tide organisms died, it would consume a lot of dissolved oxygen in water during the process of body decomposition. This leads to the massive death of fish, shrimp and shellfish because of the anoxic water mass [2].

Obviously, from different point of view, the definition of biological factor is different, and the current ontology building method can hardly describe the semantic relationships of various biological factors and non-biological factors and the interaction among them. So, to build a more comprising ontology with deep semantics, it is necessary to deal with roles. Role theory of Hozo [3] can exactly describe these complex relationships.

This paper presents a semi-automatic ontology building method to build marine organism ontology, and use role theory to describe the relations among marine organisms. After the realization of the ontology concept and relation extraction by using ontology learning technology, it needs for manual review, screening and proofing, then the ontology editor by using Hozo.

2. OVERVIEW OF ROLE THEORY AND HOZO

Role is an entity under a certain environment, which played by another entity. Research of the role theory has lasted for more than ten years, and made great achievements. Guarino [4], Loebe [5], Masolo [6], Sowa [7] and Steimann [8] have made prominent contribution in this field. But these theories do not have all the characteristics of a mature theory or model can describe the role, they also unable to process some problems such as counting problems. The researchers not only need a comprehensive theory and model to describe the role on scientific research, but also should provide an available tool and language to develop and construct ontology [9].

Mizoguchi *et al.* [10] proposed their own role theory system based on Guarino's [4] research foundation. They di-

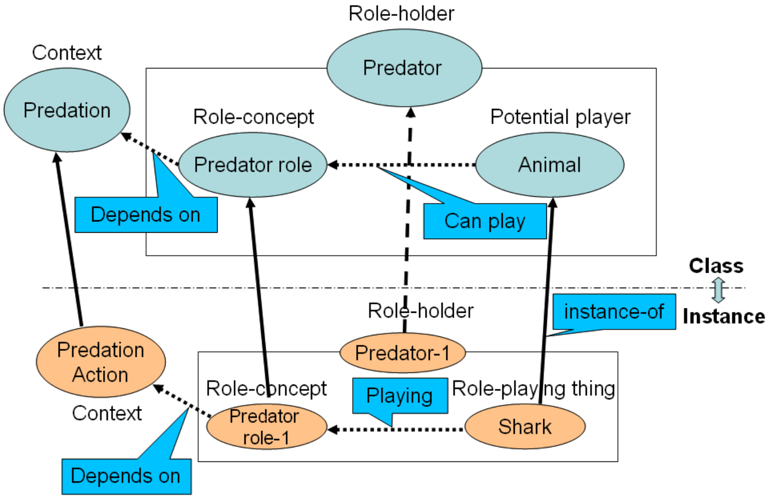

Fig. (1). A role example about marine animal.

vided the definition of concept into three categories: role-concept, basic-concept and role-holder. A role-concept represents a role which an object plays in a specific context and its definition depends on other concepts. A basic-concept does not need other concepts for being defined. Mizoguchi's role theory has been successfully used in medical ontology and the sustainable development of scientific knowledge system in Japan. In addition, by using this role theory, Lang *et al.* [11] imply unsupervised semantic role labeling by split the arguments into small clusters then merge them together to get final results. Wei *et al.* [12] represented ontology-based Home Service model to retrieve and invoke services according to user's needs automatically, and constructed two domain ontologies which are function concept ontology and context concept ontology to annotate the semantic of smart home service from different facets. In marine ecology field, Yun *et al.* [13] presented a domain upper ontology for marine ecosystem from a viewpoint of device-function, and then proposed marine ecological conceptual model and built a marine ecology OWL ontology. The ontology editor tool Hozo [3, 10] is also based on this role theory. In a word, in a context, if a potential player plays a role of role-concept, it will become a role-holder. Hozo concepts and relationships about predator role of marine animals in ocean food chain are shown in Fig. (**1**) [14].

In Fig. (**1**), the context is Predation. Context is defined as a concept which the role is recognized through a relation with. An animal can play an instance of a predator role. In particular, a hunting shark is actually playing the specific predator role named predator role-1. By doing so, it is associated with the instance predator-1, an individual predator role holder. The link from Predator-1 to Predator is a broken arrow rather than a solid one like instance-of link to show the relation is not completely same as instance-of relation in Fig. (**1**). The upper part as shown in Fig. (**1**) is class level and the lower one is instance level.

The philosophical principle of Hozo role theory is: in a context, all the parts of the whole can play their own roles. In Hozo, each concept is defined as a class, denoting with rectangles. Each class regards components or attributes as slots, which uses respectively p/o (part-of), a/o (attribute-of) to represent. Each concept node expresses a whole concept or a relation concept and its attributes. The attributes are the part concepts of the whole concept or the participant concepts of the relation concept. A part concept has three components: role-concept, class-constraint and role-holder as shown in part (a) of Fig. (**2**). The relations link the whole concept and part concept, such as p/o (part-of) and a/o (attribute-of). The two roles Predator and Prey in a Predation context are described using Hozo as shown in part (b) of Fig. (**2**) [15].

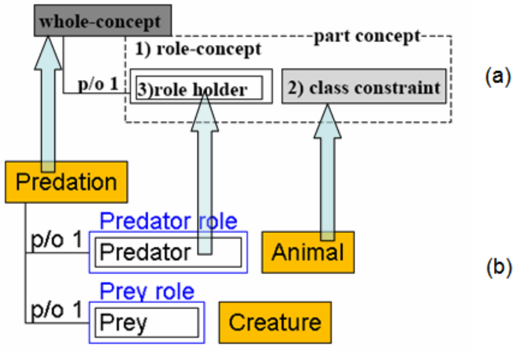

Fig. (2). Role representation using hozo editor.

In Fig. (**2**), the Predator and Prey are two part concepts of the whole concept Predation. The former is composed of Predator role, Predator and Animal. Predator is the role-holder of Predator role. Concept Animal is the class constraint of Predator role because one instance of Animal can play the Predator role; the latter is composed of Prey role, Prey and Creature just like Predator. The class constraint Creature means the instance playing the Prey role can be an animal, a plant or a microbe.

3. ONTOLOGY LEARNING TECHNOLOGY

Ontology construction is a very complex process that requires the participation of experts from various fields. Although the ontology editor tools become more and more mature, but the bulk of the building is still a very tedious manual process. Can ontology construction be achieved automatically or semi-automatically is the key to solve this prob-

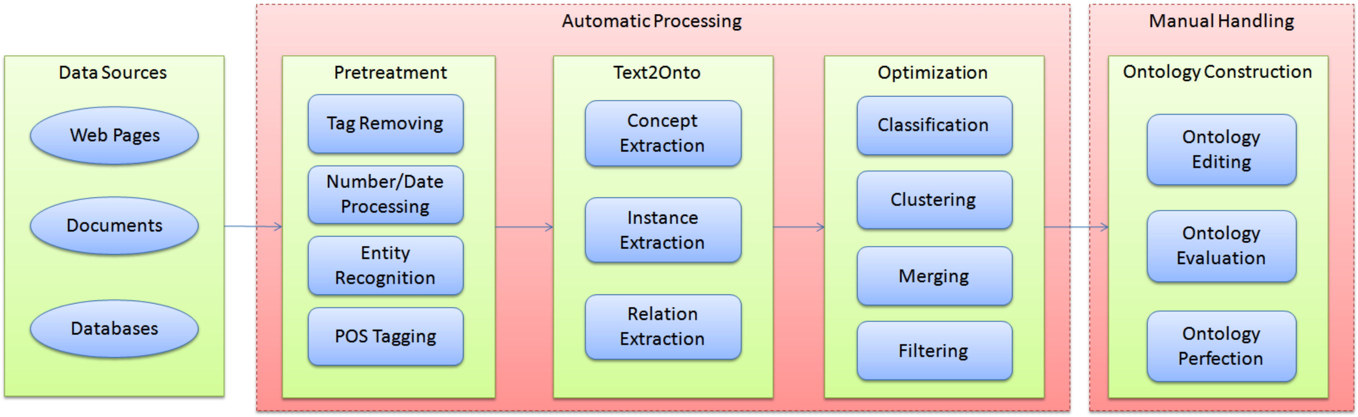

Fig. (3). Marine organism ontology construction flow.

lem. The related research of this study is very active abroad, and the related technologies named ontology learning witch discussed in literature [16]. The goal of ontology learning is to use computer technologies such as machine learning and statistics, obtain the desired ontologies automatically or semi-automatically from existing data sources.

Ontology learning includes ontology generation, ontology enrichment, ontology mining and ontology extraction and so on, which based on natural language processing and machine learning. It has formed its system framework. For different types of data sources require different ontology learning techniques, it is generally structured according to the data source level. The ontology learning techniques can divide into three categories: based on text, based on structured data and based on unstructured data.

The main task of ontology learning is to extract concepts and relationships from the data source. At present, several ontology learning systems are developed, such as Text2Onto, OntoLearn, ASIUM, Mo'k Workbench, OntoLT, Adaptiva, SOAT and DOGMA etc [16]. These systems focus on solving the problem of semantic heterogeneity using ontology. Among these tools we choose Text2Onto [17] to realize the automatic extraction of ontology concept and the relation between concepts.

Text2Onto is developed by the Institute AIFB, Germany. It is one of the most influential open-source ontology learning tools. It is an ontology learning framework based on ontology probabilistic model, using statistical methods, cutting techniques and association rules. It can find concepts, classified relations and non-classified relations from natural language texts. The core component of Text2Onto is POM. POM stores the results of the different ontology learning algorithms. Algorithm is initialized by the controller. The controller is mainly used for the pretreatment of language and executes algorithms with appropriate sequence. Algorithm change requests are applied to the POM. POM's operating authority is controlled by the limiting algorithm, so that it can farthest ensure the transparency and scalability of algorithms. The implementation of each algorithm consists of three stages: the first one is notification phase, the algorithms learn the changes of the text set; the second one is the calculation phase, these changes are mapped to the corresponding reference library which stores the reference relationship between data and ontology. The last one is generation phase. The new change requests for POM will be produced from reference library.

4. MARINE ORGANISM ONTOLOGY CONSTRUCTION

The semi automation construction of the marine biological ontology is mainly divided into two stages: the first stage is the use of ontology learning technology to realize automatic extraction of ontology concept, instance and relationship; the second stage is the artificial intervention on the automatic acquisition of concepts, instances and relationships, and then using ontology editing tool to build the ontology based on these determined concepts, instances and relationships.

The whole process flow is shown in Fig. (**3**).

We can see from Fig. (**3**), there are two main stages in the whole building process: automatic stage and manual stage, which is why we call it semi-automatic construction. There are various types of data source. The data sources need pretreatment before they inputted to Text2Onto such as tag removing, number or date time processing, entity recognition and POS tagging. During marine organism ontology building process, we only consider three ontology objects: concept, instance and relation. Some important methods are used to organize and handle the results from the previous flow including classification, clustering, merging and filtering. All above flows are processed by computer automatically. After manual examine, verify and confirmation, we use Hozo to edit the ontology. Before the ontology is applied it should be evaluated. Ontology perfection is always based on the ontology evaluation. So the ontology building is a repetitive, spiral and improving process.

4.1. Relativity Calculation

The relativity between documents and ontologies can be calculated from the following aspects:

Based on concepts. For the concepts and terms in POM, the higher the probability, the greater the relativity.

Based on relations. If there are some related concepts appear frequently in POM, we can determine that the document belongs to the field modeling.

Based on both concepts and relations. There are many terms and related concepts belong to several fields. It probability get misclassification if we only use one of them. Therefore, we consider a comprehensive mix of concepts and relations, and import concept-relation association set based on POM.

Definition 1: For word set $S = < S_C, S_R >$, define the concept-relation association set as $\{< c, r > | c \in S_C, r \in S_R, c = domain(r) \; or \; c = range(r)\}$ denoted by $CR(S)$.

Definition 2: For POM word set S and document d, there exists:

$$relation(S,d) = k \cdot \frac{|S_C \cap d|}{|S_C|} + l \cdot \frac{|S_R \cap d|}{|S_R|} + m \cdot \frac{|CR(S) \cap (d \times d)|}{|CR(S)|} \quad (1)$$

In this paper, we set $k = 0.2$, $l = 0.3$ and $m = 0.5$. Given a word set S, for a document d, if the $relation(S,d) < \omega$, d is an irrelevant document. In order to simplify the experimental model and optimize the experimental results, we select $\omega = 0.7$.

4.2. Concept Automatic Extraction

The task of concept extraction is to extract representative term set from data sources. We used a mixture method based on linguistic and statistical to extract the concepts from pretreatment corpus. The words and phrases which their occurrence frequency is too high or too low should be filtered out from the primary focus.

Roberto Navigli proposed a screening method which introduced in literature [18] based on domain correlation and domain consistency. Based on his research, we used domain coverage instead of domain consistency.

Assume that given domain set $\{D_1, D_2, ..., D_n\}$, the term t is an element of primary term set. So for D_k, the relativity of t can be calculated as:

$$R_{t,k} = \frac{P(t \mid D_k)}{\max\limits_{1 \le k \le n} P(t \mid D_k)}$$

(2)

where the conditional probability $P(t \mid D_k)$ can be estimated:

$$E(P(t \mid D_k)) = \frac{f_{t,k}}{\sum\limits_{t \in D_k} f_{t,k}}$$

(3)

where $f_{t,k}$ means the frequency of t in D_k.

The domain coverage is the number which term t included in the document d for D_k. It abbreviates as $Q_{t,k}$ which can be calculated:

$$Q_{t,k} = \frac{\sum\limits_{d_j \in D_k} q_{t,d}}{\sum\limits_{d_j \in D_k} l}$$

(4)

where d is a document of domain D_k, $q_{t,d}$ means whether the term t appeared in document d or not, if d contains the t, then $q_{t,d} = 1$,otherwise $q_{t,d} = 0$.

Quantify the relevance and coverage of terms in the field, the important degree of term t in D_k denoted as $T_{t,k}$ can be expressed:

$$T_{t,k} = \alpha \cdot R_{t,k} + \beta \cdot Q_{t,k}$$

(5)

where $\alpha, \beta \in (0,1)$, in our experiments $\alpha = 0.7$ and $\beta = 0.3$.

4.3. Semantic Relation Automatic Extraction

Semantic relations link the concepts of domain. According to the semantic relationship between concepts, the taxonomy relation can be created. There are many relations between concepts including inheritance, synonyms, antisense and other semantic relations in a particular scene. The key issue of relation finding is calculating the correlation between concepts. If two concepts are close the correlation is high, otherwise the correlation is low. When Text2Onto is used to extract the relations, there are many algorithms to choose. But those algorithms are not good enough to gain the relations especially on handling Chinese ontology semantic relations.

Concept hyponymous relation is the core semantic relationship of domain ontology [19]. Mo *et al.* [20] proposed a hyponymy extraction method of domain ontology concept based on Cascaded Conditional Random Field (CCRF). The conditional random field model is:

$$P(Y \mid X) = \frac{1}{Z_X} \exp\left(\sum_{j=1}^{N} \sum_{i} \lambda_i f_i \left(y_{j-1}, y_j, X, j \right) \right)$$

(6)

where $X = x_1, x_2, ..., x_N$ means the words and their POS order; $Y = y_1, y_2, ..., y_N$ means the predicted simple entities sequence; $Z_X = \sum_{y \in Y} \exp\left(\sum_{j=1}^{N} \sum_{i} \lambda_i f_i \left(y_{j-1}, y_j, X, j \right) \right)$ is the normalizing factor; λ_i is the weight of characteristic function f_i.

The Viterbi dynamic programming algorithm is used to calculate the remark sequence:

$$Y^* = \arg\max_{Y} P(Y \mid X)$$

(7)

Non-taxonomic relation extraction is the emphasis and difficulty of ontology learning. The VCC (n)-transactions

Table 1. Marine organism ontology objects extraction result.

Ontology Objects	Total Number				Precision (%)	Recall (%)	F1 (%)
	Documents	Reference	Extraction	Correct			
concepts	100	8660	8002	7058	88.2	81.5	84.7
instances	100	1798	2205	1663	75.4	92.5	83.1
relations	100	12839	10206	9308	91.2	72.5	80.8

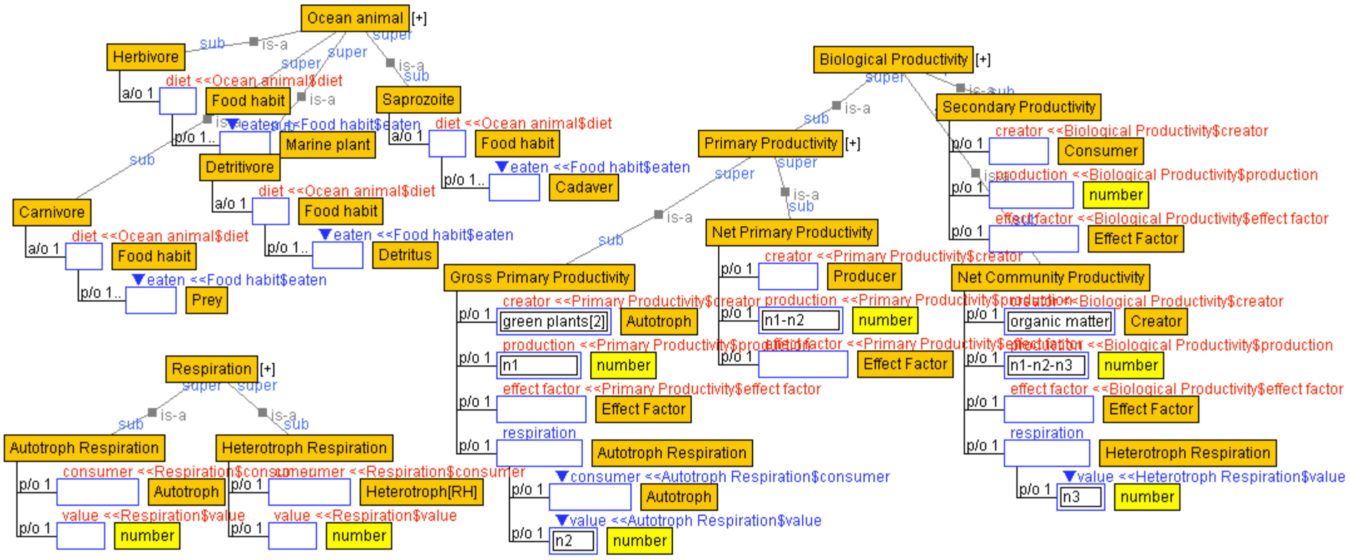
Ocean animal [+]
Herbivore
Saprozoite
Carnivore
Detritivore
Food habit
Marine plant
Cadaver
Detritus
Prey
diet <<Ocean animal$diet
eaten <<Food habit$eaten
is-a
sub
super
a/o 1
p/o 1
Biological Productivity [+]
Secondary Productivity
Primary Productivity [+]
Net Primary Productivity
Gross Primary Productivity
Net Community Productivity
creator <<Biological Productivity$creator
production <<Biological Productivity$production
effect factor <<Biological Productivity$effect factor
creator <<Primary Productivity$creator
production <<Primary Productivity$production
effect factor <<Primary Productivity$effect factor
Consumer
Producer
Creator
Autotroph
number
Effect Factor
green plants[2]
n1
n1-n2
n1-n2-n3
organic matter
respiration
Respiration [+]
Autotroph Respiration
Heterotroph Respiration
consumer <<Respiration$consumer
value <<Respiration$value
Heterotroph[RH]
consumer <<Autotroph Respiration$consumer
value <<Autotroph Respiration$value
value <<Heterotroph Respiration$value
n2
n3

Fig. (4). Marine organism ontology fragment.

method [21] can be used to extract non-taxonomic relation. It assumes that if two concepts C_1 and C_2 have non taxonomic relationships V, if and only if C_1 and C_2 are around the verb V, you can use a conditional probability to represent the correlation degree among the verb and the concepts:

$$P(c_1 \wedge c_2 / v) = \frac{\left|\{t_i \mid v, c_1, c_2 \in t_i\}\right|}{\left|\{t_i \mid v \in t_i\}\right|}$$

(8)

$$P(v / c_1 \wedge c_2) = \frac{\left|\{t_i \mid v, c_1, c_2 \in t_i\}\right|}{\left|\{t_i \mid c_1, c_2 \in t_i\}\right|}$$

(9)

where t_i means VCC (n)-transactions. Equation (8) expresses the possible correlation of the concept pair and the given verb and equation (9) expresses the possible correlation of the verb and the given concept pair [22].

4.4. INSTANCE AUTOMATIC EXTRACTION

An instance is the specific individual belonging to a certain concept. Instance extraction is the process that identifies the individuals of a concept from domain corpus. A common instance extraction method is considering it as the instance-of relation between concepts. KL distance can be used to extract the instance-of relation:

$$D(q \| r) = \sum_x q(x) \log \frac{q(x)}{r(x)} \tag{10}$$

where $q(x)$ and $r(x)$ are two probability distribution functions. The function of $D(q \| r)$ is to measure the distance between two probability distribution.

5. EXPERIMENT AND DISSCUSS

We chose 100 web pages about marine organism domain from Internet and converted them into text format. Then we used TxtWitch tool to deal with the text documents and removed the irrelevant information. At last, after finishing the tag removing, number and date format processing, entity recognition and POS tagging, we put the texts into Text2Onto to extract ontology objects. The experiment results are shown in Table **1**.

From the experimental results shown in Table **1**, it was found that the precision, recall and F-measure are all over 80%. That shows the automatic extraction can be used to help the artificial ontology construction of the next stage, and it also shows the artificial intervention is an indispensable work.

Based on the automatic extraction of concepts, instances and relationships, we used Hozo to build the marine organism ontology. The ontology fragment is shown in Fig. (**4**).

CONCLUSION

Marine organism ontology semi-automatic construction processing combines artificial intelligence, information retrieval, knowledge engineering and computational linguistics and other disciplines. Based on the existing Text2Onto system, we realized the ontology concept extraction, instance extraction and relation extraction terminology and other related technologies to facilitate access from the text. It largely simplified ontology building process. But the higher manual operation requirements is indispensable because the precision, recall and F-measure are not enough satisfied. In the next step, we will use the machine learning and text mining technology to perfect the method. Ontology automatic evaluation will also be considered.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 61401242), the Development Projects of Henan Province Science and Technology (No. 132102210264) and the Sponsored by Program for Science and Technology Innovation Talents in Universities of Henan Province (No. 15HASTIT023).

REFERENCES

[1] Uschold M and Gruninger M, “Ontologies: Principles, methods and applications,” *The knowledge engineering review*, vol. 11, pp. 93-136, February 1996.

[2] Xiong Jing, “Research and Application of Marine Ecology Ontology Modeling,” PhD thesis of Ocean University of China, 2010.

[3] Kozaki K, Kitamura Y, Ikeda M, *et al.*, “Hozo: an environment for building/using ontologies based on a fundamental consideration of ‘Role’ and ‘Relationship’,” *Lecture Notes in Computer Science*, vol. 2473, 2002, pp. 213-218.

[4] Guarino, N., “Some Ontological Principles for Designing Upper Level Lexical Resources,” In Proceedings of First International Conference on Language Resources and Evaluation, ELRA-European Language Resources Association, Granada, Spain, pp. 527-534, 1998.

[5] Loebe F., “Abstract vs. social roles-Towards a general theoretical account of roles,” *Applied Ontology*, vol. 2, pp. 127-158, February 2007.

[6] Masolo, C., Vieu, L., Bottazzi, E., *et al.*, “SocialRoles and their Descriptions,” In: Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning, Whistler, pp. 267-277, 2004.

[7] Gruber T R., “Toward principles for the design of ontologies used for knowledge sharing,” *International journal of human-computer studies*, vol. 43, pp. 907-928, May 1995.

[8] Steimann F., “On the representation of roles in object-oriented and conceptual modelling,” *Data & Knowledge Engineering*, vol. 35, pp. 83-103, January 2000.

[9] Xiao-Dong W., Xiao-Hong Z., Jing W., *et al.*, “Time Ontology Building Based on Role and Relation,” *Journal of Henan Normal University (Natural Science)*, vol. 36, pp. 29-31, January 2008.

[10] Mizoguchi R., Sunagawa E., Kozaki K., *et al.*, “The model of roles within an ontology development tool: Hozo,” *Applied Ontology*, vol. 2, pp. 159-179, February 2007.

[11] Lang J. and Lapata M., “Unsupervised semantic role induction *via* split-merge clustering,” in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, vol. 1, pp. 1117-1126, 2011.

[12] Wei M., Xu J., Yun H., *et al.*, “Ontology-based home service model,” *Computer Science and Information Systems*, vol. 9, pp. 813-838, February 2012.

[13] YUN Hongyan, XU Jianliang, GUO Zhenbo, *et al.*, “Modeling of marine ecology ontology,” *Journal of Computer Applications*, vol. 34, pp. 1105-1108, April 2014.

[14] Jia H., Xiong J., Xu J., *et al.*, “Research and application of role theory in ocean carbon cycle ontology construction,” *Journal of Ocean University of China*, vol. 13, pp. 979-984, December 2014.

[15] Xiong Jing, Guo Lei and Xu Jianliang, “Application of Domain Ontology in Marine Ecology Knowledge Management,” *New Technology of Library and Information Service*, vol. 28, pp. 15-22, March 2012.

[16] Du X., Man L. and Shan W., “Ontology Learning Survey,” *Journal of Software*, vol. 17, pp. 1837-1847, September 2006.

[17] Cimiano P. and Völker J., “Text2Onto,” *Lecture Notes in Computer Science*, vol. 3513, pp. 227-238, 2005.

[18] Navigli R. and Velardi P., “Learning domain ontologies from document warehouses and dedicated web sites,” *Computational Linguistics*, vol. 30, pp. 151-179, February 2004.

[19] LIU Kai-Peng and FANG Bin-Xing, “Ontology Induction Based on Social Annotations,” *Chinese Journal of Computers*, vol. 33, pp. 1823-1834, October 2010.

[20] MO Yuan-yuan, GUO Jian-yi, YU Zheng-tao, *et al.*, “Hyponymy Extraction of Domain Ontology Concept Based on CCRF,” *Computer Engineering*, vol. 40, pp.138-141, June 2014.

[21] Buitelaar Paul, Daniel Olejnik and Michael Sintek, "A protégé plug-in for ontology extraction from text based on linguistic analysis," *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, pp. 31-44, 2004.

[22] LIU Bai-song, “Web-based General Ontology Learning Framework,” *Computer Engineering*, vol. 34, pp. 229-231, April 2008.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014