A Domain Ontology Learning Approach Based on Soft-computing Techniques

基於軟式計算技術之領域實體論學習方法

Yuan-Fang Kao
CREDIT Research Center,
National Cheng Kung University,
Tainan, Taiwan
kaoyf@cad.csie.ncku.edu.tw

Chang-Shing Lee
Dept. of Information Management,
Chang Jung University, Tainan, Taiwan
leecs@mail.cju.edu.tw

Yau-Hwang Kuo
CREDIT Research Center,
National Cheng Kung University,
Tainan, Taiwan
kuoyh@cad.csie.ncku.edu.tw

I-Heng Meng
Advanced e-Commerce Technology Lab,
Institute for Information Industry, Taiwan
ihmeng@iii.org.tw

Abstract

Ontology is increasingly important in knowledge management and Semantic Web. The problem of it is that the construction of ontology is a time-consuming job and ontology engineers need to spend much time to maintain it. In this paper, we propose an incremental domain ontology learning method. This method can effectively extract new information from new domain documents to update the schema of domain ontology and make the knowledge base of domain ontology more complete based on a constructed domain ontology. First, we use schema of domain ontology to extract candidate instances. We also use genetic algorithm to learn the knowledge base of fuzzy inference. The three-layer parallel fuzzy inference mechanism is further applied to obtain new instances for ontology learning. In addition, new attributes, operations, and associations will be extracted based on episodes and morphological analysis to update the domain ontology.

Keywords: Ontology Learning, Chinese Natural Language Processing, Episode Mining, Soft-computing

摘要

實體論(Ontology)在知識管理及語意網(Semantic Web)中越來越重要，但建構實體論往往需要耗費大量的時間，且建構完成後維護實體論對知識管理者來說也是費時的工作。本論文中，我們提出一個漸進式領域實體論學習方法；此方法是以一個已建構出的領域實體論為基礎，來做領域實體論學習，此方法能有效的由新的領域文件中截取出新的資訊，以更新領域實體論的架構及領域實體論的知識庫，首先，我們利用原先領域實體論的架構來擷取出候選實體，再利用一個三層平行模糊推論機制來推論出新的實體，並且使用基因演算法來學習模糊推論機制所需的知識庫，此外，利用插曲探勘及自然語言處理來找出新的屬性(Attribute)、行為(Operation)與概念間的關連關係(Association)。

關鍵詞：實體論學習、中文自然語言處理、插曲探勘、軟式計算

1. Introduction

With the support of ontology, both user and system can communicate with each other by the shared and common understanding of a domain [13]. The problem of it is that the construction of ontology is a time-consuming job and ontology engineer needs to spend much time to maintain it. In recent years, there are many researchers have proposed various approaches for the ontology construction [9], [11], [15]. We also had proposed an automatic ontology construction method in [7]. Whether domain ontology is constructed by domain expert or automatically, we need a mechanism to maintain it. Since domain knowledge may change with times, domain ontology need to be updated timely. Therefore, in this paper, we will propose a method of incremental domain ontology learning.

Recently, there is no clear definition of on-

Corresponding Author: Prof. Chang-Shing Lee is with the Department of Information Management, Chang Jung University, Tainan, 711, Taiwan
Email: leecs@mail.cju.edu.tw / leecs@cad.csie.ncku.edu.tw

tology learning. Some researches of ontology learning just related to ontology construction not ontology maintenance. Weather ontology has been applied in Semantic Web or other information systems. Maintenance of ontology requires tremendous efforts that force future integration of many techniques to enable highly automated ontology learning, e.g., machine learning and knowledge acquisition. A. Maedche et al. [9] propose an ontology-learning framework that proceeds through ontology import, extraction, pruning, refinement, and evaluation, giving the ontology engineer coordinated tools for ontology modeling. They consider ontology learning as semiautomatic with human intervention, adopting the paradigm of balanced cooperative modeling for constructing ontology for the Semantic Web. Omelayenko [12] explains that ontology learning is an emerging field aimed at assisting a knowledge engineer in ontology construction and semantic page annotation with the help of machine learning techniques. They separate two main tasks in ontology learning. They are ontology construction task (ontology schema and instances extraction) and ontology maintenance task (ontology integration, navigation, update, enrichment). In this paper, we will propose an incremental domain ontology update method base on a constructed ontology. This method can effectively extract new information from new domain documents to update the schema of domain ontology and make the knowledge based of domain ontology more complete.

The organization of this paper is as follows. In Section 2, we describe the structure of four-layered object-oriented ontology. In Section 3, the method of incremental domain ontology learning is presented. The experimental results are shown in Section 4. Finally, we make conclusions in Section 5.

2. Structure of Four-Layered Object-Oriented Ontology

Ontology is an explicit specification of some topic, and represents knowledge based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them [4]. Therefore, ontology defines a set of representational terms that we call concepts. Inter-relationships among these concepts describe a target world [5]. The structure of four-layered object-oriented ontology is shown in Figure 1. There are four kinds of relations including *association*, *generalization*, *aggregation*, and *instance-of*, and four layers in this architecture. We describe them as follows.

The domain layer represents a name of domain ontology such as medical or news. The category layer is the categories of domain knowledge. The concept layer is the schema of domain ontology. It is composed of concept set. The instance layer is the knowledge base of domain ontology. It is composed of instance set. Each concept and each instance contain its name, attributes, and operations. Attributes describe various static features and properties. Operations describe various dynamic behaviors. A concept describes a group of instances with similar attributes, operations, and relations to other instances. An instance is an entity of its concept. Most instances drive their individuality from difference in their attribute values and relations to other instances.

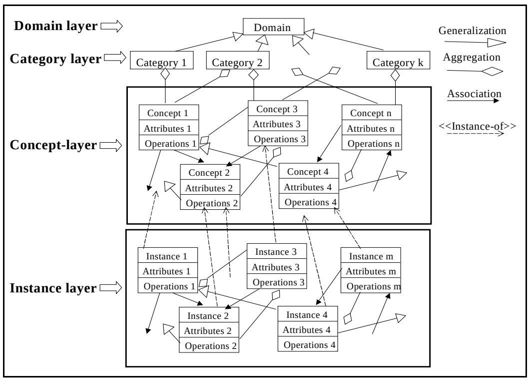

Figure. 1. The architecture of four-layered object-oriented ontology.

3. Incremental Domain Ontology Learning

This method can extract new information from new domain documents to update the schema of domain ontology and make the knowledge base of domain ontology more complete based on a constructed domain ontology. Figure. 2 shows the flowchart of incremental domain ontology learning. Next, we describe each process in it.

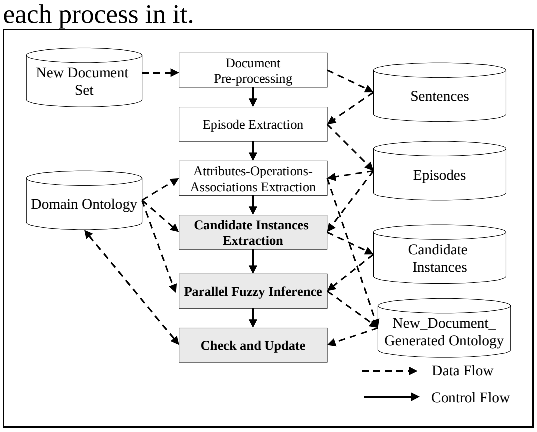

Figure. 2. Flowchart of incremental domain ontology learning.

3.1 Document Per-processing

We use the Chinese POS Tagger, CKIP [1], to parse Chinese documents and preserve the terms with partial noun tags or verb tags. In this paper, the preserved terms include Na (common noun), Nb (proper noun), Nc (location noun), Nd (time noun) and all kinds of verbs(VA, VB, VC, VD, VE, VF, VG, VH, VI, VJ, VK, VL). The filtered terms include Ne (stable noun), Nf (quantity noun), Ng (direction noun), Nh (pronoun), adjective, adverb, preposition, conjunction, particle, and interjection

3.2 Episode Mining

3.2.1 Concept of the Episode

The concept of the episode is proposed by Ahonen et al. [2] and Mannila et al. [10]. An episode is a partially ordered collection of events occurring together. The episode can be seen as partially ordered pattern. Consider, for instance, episodes α, β, and γ are extracted from the sequential data in Figure. 3, the episode α is a serial episode, and it means "*a*" always occurs before "*b*". The episode β is a parallel episode, and it means "*d*" and "*e*" always occurs together. The episode γ is an example of non-serial and non-parallel episode, it means "*c*" always occur before "*d*" and "*e*".

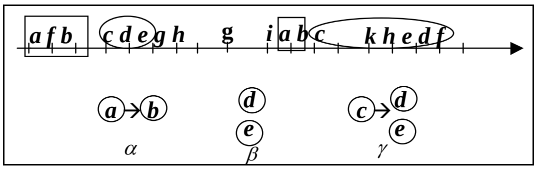

Figure. 3. An example of the episode.

3.2.2 Episode Extraction

By the *Document Pre-processing* process, the text is separated into the nouns and sentences. Then the sentences will be fed into the *Episode Extraction* process to get the episodes. In this paper, a term is denoted as a triple *(term, POS, index)* where *index* is the position of this term in the sentence. An episode is extracted if it occurs within an interval of given window size and its occurrence frequency in the document set is larger than the defined minimal occurrence value. In order to get more accurate episodes, punctuation marks is filtered and the POS of terms with Na, Nb, Nc, Nd and verbs are retained in the sentence. The following shows a sentence example.

”德國門將卡恩贏得本屆世足賽代表最佳球員的金球獎。”

By the *Document Pre-processing* with CKIP process, the sentence with the terms and POS is generated as follows.

德國(Nc) 門將(Na) 卡恩(Nb) 贏得(VJ) 本(Nes) 屆(Nf) 世足賽(Nb) 代表(Na) 最佳(A) 球員(Na) 的(DE) 金球獎(Nb)。(PERIODCATEGORY)

By the *Stop Word Filter* process, the terms with triple *(term, POS, index)* representation are shown below.

(德國, Nc, 1) (門將, Na, 2) (卡恩, Nb, 3) (贏得, VJ, 4) (世足賽, Nb, 5) (代表, Na, 6) (球員, Na, 7) (金球獎, Nb, 8)

Finally, the *episode extraction* process will generate the episodes with window size 6 as follows.

德國(Nc)_門將(Na)_卡恩(Nb)
卡恩(Nb)_贏得(VJ)_金球獎(Nb)

3.3 Attributes-Operations-Associations Extraction

After getting the episodes, terms in episodes are mapped to the instances in original domain ontology to tag the concept name. An example is shown as follows:

南韓(Nca|球隊),擊敗(VC),義大利(Nca|球隊)

We can know that "南韓(Korea)" and "義大利(Italy)" are instances of concept "球隊(team)" by tagging the concept name. Here, we extract attributes, operations, and associations of these existed instances first. Then, when new instances are extract, their attributes, operations, and associations will also be extracted.

The extraction method is the same as our before proposed approach [7]. We use the morphological information of Chinese term and Chinese syntax. We attempt to extract patterns such as "object-attribute-value", "object-association-object", or "object-operation" from domain data. These patterns are considered as kinds of sentence patterns, e.g., "subject-verb-objective" or "subject-modifier". We analyze morphological features of Chinese terms to assist the extraction of attributes, operations, and associations from episodes.

Now we describe the morphological features of Chinese terms. The *Chinese Knowledge Information Processing Group* of Academia Sinica [1] classifies verbs into twelve categories. In our morphological analysis, these twelve categories of verbs are classified into five groups by their meaning and syntax. These five groups of verbs are treated as operations or associations by their morphological feature (see Table 1). Operations describe actions of a concept, so verbs, which only need a subject are selected as operations. Associations describe relationship between two concepts, so verbs, which need a subject and objectives are selected as associations. In the same way, nouns are treated as concept or property (include attributes and associations) by their morphological feature (see Table 2).

Table 1. Morphological analysis for Chinese verbs

Morphological feature	**POS of CKIP**	**Description**	**Example**	**Role in ontology**
Intransitive verb	VA	Only need subject	說話、進球	Operation
Transitive verb	VB、VC、VD、VE、VF	Need subject and objective	擊敗、預防	Association
Linking verb	VG	Link subject and objective and express a equivalent relation	等於、尊稱	Association
Status intransitive verb	VH	Only need subject and describe status of subject	動聽、炎熱	Attribute
Status transitive verb	VI 、VJ 、VK、VL	Only need subject	造成、遇到	Association

Table 2. Morphological analysis for Chinese nouns

Morphological feature	**POS of CKIP**	**Example**	**Role in ontology**
Substance noun (Uncountable concrete noun)	Naa	泥土、雨	Concept
Countable concrete noun	Nab	乘客、門將	Attribute, Association, Concept
Countable abstract noun	Nac	路徑、位置	Attribute, Association, Concept
Uncountable abstract noun	Nad	風度、香氣	Attribute, Association, Concept
Collective noun	Nae	車輛、獎金	Attribute, Association, Concept
Proper noun	Nb	雙魚座、世足賽	Concept
Proper local noun	Nca	西班牙、台北	Concept
Common local noun	Ncb	郵局、中心	Attribute, Association, Concept
A noun of locality	Ncc	海外、身上	Concept
Positional noun	Ncd	外頭、左	Concept
Named local noun	Nce	四海、當地	Concept

3.4 Candidate Instances Extraction

In this process, we compare the schema of domain ontology and episodes to extract candidate instances. Candidate instances are defined as possible new instances. If the term in episode has similar attributes, operations and associations to some concepts, then it can be extracted as a candidate instance. For example, from the following episodes, We can extract a candidate instance “國際足總” and it has attributes “發言人”, “會長” and association domain “教練”, “裁判” (see Figure. 4(a)). According the domain ontology, the concept “比賽” has the association domain “裁判”, the concept “球隊” has the association domain “教練”, and the concept “組織” has the attributes “發言人” and “會長” (see Figure. 4(b)), so candidate instance possibly belong to concept “比賽”, “球隊”, or “組織”. Association domain means the arrowhead of associations is toward other concepts or instances, and association range means the arrowhead of associations is toward itself.

國際足總(Nb),發言人(Nab)
國際足總(Nb),布萊特(Nba),裁判(Nab)
黃武雄(Nb),國際足總(Nb),教練(Nab)
國際足總(Nb),會長(Nab),布萊特(Nba)

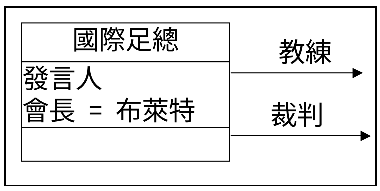

Figure. 4(a). An example of candidate instance.

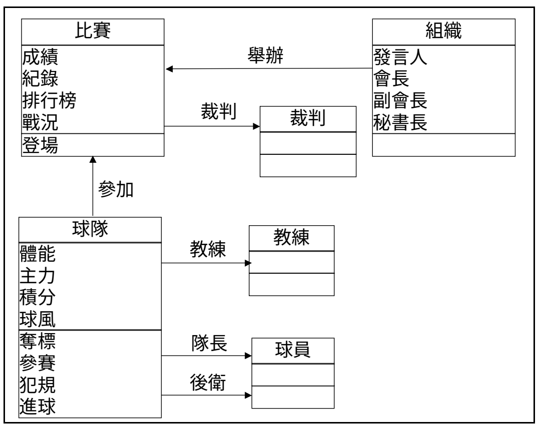

Figure. 4(b). Some concepts in the domain ontology for candidate instance extraction.

Figure. 4. An example of candidate instance extraction.

3.5 Parallel Fuzzy Inference

A parallel fuzzy inference model is utilized to infer candidate instances belong to which existed concept, and we use the genetic algorithm to learn the data base and rule base of this model. Next, we describe the conceptual resonance strength between a concept and an instance in section 3.5.1. In section 3.5.2, we describe how to generate the knowledge base by genetic learning. Moreover, a parallel fuzzy inference model for conceptual resonance strength computing is described in section 3.5.3.

3.5.1 Conceptual Resonance Strength Between a Concept and a New Instance

The conceptual resonance strength is defined as the belonging degree between a concept and an instance. Hence, a candidate instance may have a highly possibility belonging to the concept if their conceptual resonance strength is high. It determines a candidate instance belongs to an existed concept or to be a new concept. Concept describes a group of instances with identical attributes, operations and associations to other instances. Therefore, if a candidate instance is a new instance of some concept, their conceptual resonance strength is high and they must have some identical attributes, operations and associations. If all the conceptual resonance strength with all existed concepts is small, the candidate instance may possibly be a new concept.

There are four fuzzy input variables for computing conceptual resonance strength between a concept and a candidate instance. They are *conceptual resonance strength in attribute* x_A, *conceptual resonance strength in operation* x_O, *conceptual resonance strength in association domain* x_D and *conceptual resonance strength in association range* x_R. The fuzzy output variable is *conceptual resonance strength* y_{CRS}. Therefore

$$CRS = (x_A,\ x_O,\ x_D,\ x_R) \tag{1}$$

The methods of computing these four fuzzy variables are the same. Here, we only describe computing formula of x_A. The x_A means the similarity of attributes between an existed concept *C* of domain ontology and a candidate instance *I*. Eq. 2 illustrates the computing formula.

$$x_A = \frac{n_{Attribute}(C \cap I)}{n_{Attribute}(C)} \times \frac{N_{Attribute}(C \cap I)}{N_{Attribute}(I)} \tag{2}$$

where $n_{Attribute}(C)$ is the number of attributes in concept *C*, $n_{Attribute}(C \cap I)$ is the number of identical attributes in concept *C* and candidate instance *I*, $N_{Attribute}(I)$ is the occurrence number of episodes in attributes of candidate instance *I*, $N_{Attribute}(C \cap I)$ is the occurrence number of episodes in identical attributes of concept *C* and candidate instance *I*. Moreover, $\frac{n_{Attribute}(C \cap I)}{n_{Attribute}(C)}$ means the ratio of identical attributes, and $\frac{N_{Attribute}(C \cap I)}{N_{Attribute}(I)}$ means the ratio of occurrence number of identical attributes. Figure 5 shows an example of membership function x_A, and we will introduce how to learn it based on genetic algorithm in section 3.5.2.

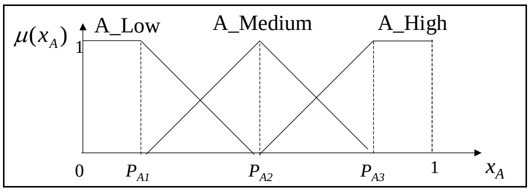

Figure. 5. An example of membership function of fuzzy variable x_A.

3.5.2 Generation of Knowledge Base by Genetic Learning

Our learning approach composed of two methods with different goals. We modify the genetic learning method proposed by O. Cordon et al. [3] and data-driven method proposed by Wang et al. [14] to generate the knowledge base. The knowledge base included the data base and rule base. The data base consists of the number of linguistic terms and the parameters of membership functions of each fuzzy variable x_A, x_O, x_D, x_R and y_{CRS}, and the rule base consists of fuzzy rules. Next, we describe our data base generated method in section 3.5.2.1 and rule base generated method in section 3.5.2.2, respectively.

3.5.2.1 Generation of Data Base by Genetic Learning

First, we encode the data base in the floating point implementation. Each chromosome is composed of two parts. One is the number of linguistic terms C_1, and the other is the parameters of membership functions C_2. The number of linguistic terms for each variable is stored into a vector C_1 of length 5.

$$C_1 = (L_1, L_2, L_3, L_4, L_5) \quad (3)$$

where (L_1, L_2, L_3, L_4, L_5) represents the number of linguistic terms of each fuzzy variable (x_A , x_O , x_D , x_R , y_{CRS}). The value of each L_i is restricted in the set $\{2, 3, 4\}$.

C_2 represents the parameters of each membership function. The shape of each member function is triangular, besides the first and the last have shoulder (see Figure. 6). The center vertex of the member function is used to represent the linguistic term in C_2 part. The parameters of membership function for each variable are stored into a vector C_2 of length 5*2 to 5*4. The length of C_2 changes with the number of linguistic terms.

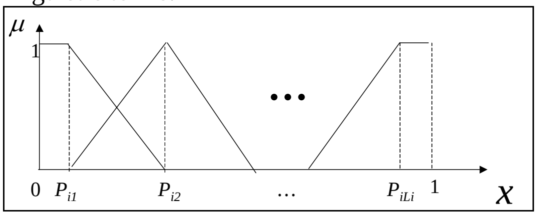

Figure. 6. The membership function in PFIS.

$$C_2 = (C_{21}, C_{22}, C_{23}, C_{24}, C_{25}) \quad (4)$$

$$C_{2i} = (P_{i1}, P_{i2}, \ldots, P_{iLi}) \text{ and } P_{i1} < P_{i2} < \ldots < P_{iLi}. \quad (5)$$

where C_{2i} represents the parameters of membership functions of ith fuzzy variable and P_{ij} represents the center vertex of jth linguistic term. Each chromosome $C = C_1\, C_2$ can generate a set of fuzzy rules.

The initial population is composed of three groups with the same number. The number of linguistic terms in C_1 part is random generated. In first group, the parameters of membership functions in C_2 part are random generated. In second group, the parameters of membership functions in C_2 part are uniformly distributed in the range [0,1]. In third group, the parameters of membership functions in C_2 part are defined by a domain expert.

New, we briefly describe the related genetic operators used in this paper as follows.

1. Fitness function: The mean square error (*MSE*) is used as fitness function.
2. Selection: The selection probability calculation follows linear ranking. The selection probability is computed by using the non-increasing assignment function.
3. Crossover: The standard crossover operator is applied over the two parts of the chromosomes. When C_1 is crossed at a random point, the corresponding values in C_2 are also crossed in the two parents.
4. Mutation: Two different operators are used in C_1 and C_2, respectively. In C_1 part, we random select the number of linguistic terms of the variable and change it to the immediately upper or lower value. In C_2 part, Michalewicz's nonuniform mutation operator is used.

3.5.2.2 Generation of Rule Base by Data-Driven Method

This method is utilized data-driven method to generate fuzzy rules from training data pairs and chromosomes. It consists of three steps described below.

1. *Generate fuzzy rules from given data pairs*

First, we prepare a set of desired input-output data pairs as training data.

$$(x_A^1, x_O^1, x_D^1, x_R^1, y_{CRS}^1), (x_A^2, x_O^2, x_D^2, x_R^2, y_{CRS}^2), \ldots, (x_A^n, x_O^n, x_D^n, x_R^n, y_{CRS}^n) \quad (6)$$

Second, determine the membership degree of given (x_A^i, x_O^i, x_D^i , x_R^i , y_{CRS}^i) in different fuzzy set. Third, assign (x_A^i, x_O^i, x_D^i , x_R^i , y_{CRS}^i) to the fuzzy set with maximum membership degree. Finally, obtain one rule from one desired input-output data pair.

2. *Assign a degree to each rule*Each data pair can generate a fuzzy rule, so we can get lots of rules after step 1. It is highly possibility that there will be some conflicting rules, which have the same *IF* part but a different *THEN* part. Therefore, we assign a degree to each rule and accept the rule with maximum degree to resolve conflicting rules. We use the following product strategy to calculate the degree of each rule.

$$D(Rule^i) = \mu_A(x_A^i)\mu_O(x_O^i)\mu_D(x_D^i)\mu_R(x_R^i)\mu_{CRS}(y_{CRS}^i) \quad (7)$$

3. *Create a combined fuzzy rule base*

In this step, we simplify the complex fuzzy rules by combining lots of fuzzy rules to a single fuzzy rule. The combined rules must satisfy following three conditions:

a. The linguistic terms in their *THEN* parts are the same.

b. The linguistic terms in their *IF* parts are the same except one variable.

c. The union set of all the different linguistic terms covers the domain of the variable.

Since the knowledge base of fuzzy inference is generated by learning, we don't need domain expert to design it. The learned knowledge base are different depend on different training data of different domain, so the inference results are more precision.

3.5.3 A Parallel Fuzzy Inference Model for Conceptual Resonance Strength Computing

In this section, we describe how to aggregate four input fuzzy variables (x_A, x_O, x_D, x_R) into one output fuzzy variable (y_{CRS}) for computing the conceptual resonance strength for each input data pair. The three-layered parallel fuzzy inference architecture proposed by Kuo et al. and Lin et al [6] [8] is used in this thesis. Figure. 7 shows the architecture

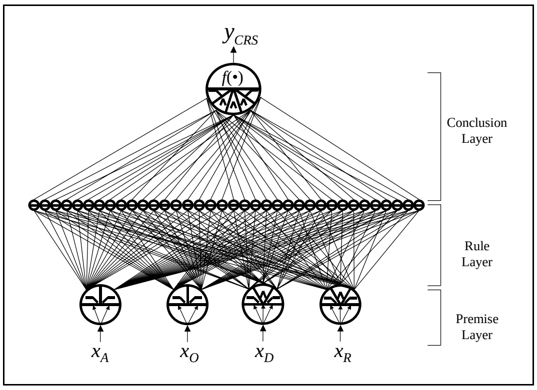

Figure 7. Three-layered parallel fuzzy inference mechanism for conceptual resonance strength computing.

The structure consists of premise layer, rule layer and conclusion layer. The premise layer performs the first inference step to compute matching degrees. The input vector is $x = (x_A, x_O, x_D, x_R)$, where x_i is the input value of fuzzy variable x_i. The output vector of the premise layer will be

$$\mu=((\mu_{A1},\mu_{A2}),(\mu_{O1},\mu_{O2}),(\mu_{D1},\mu_{D2},\mu_{D3},\mu_{D4}),(\mu_{D1},\mu_{D2},\mu_{D3},\mu_{D4})) \quad (8)$$

where μ_{ij} is the membership degree of the *j*-th linguistic term in the fuzzy variable x_i.

The second layer is called the rule layer where each node is a rule node to represent a fuzzy rule. The links in this layer are used to perform precondition matching of fuzzy rules, and the output of a rule node will be linked with associated linguistic term in the third layer. The third layer is the conclusion layer. In this layer, fuzzy linguistic nodes are responsible for making conclusion and defuzzification. The output fuzzy variable is y_{CRS}. Its linguistic terms and parameters of membership functions are also generated automatically. We use the Center of Area (COA) method to defuzzify.

Therefore, a candidate instance will become a new instance of an existed concept with the highest y_{CRS} value. If all y_{CRS} values are smaller than the threshold, the candidate instance is determined to generate a new concept or be discarded by domain expert.

3.6 Check and Update

A new-document-generated ontology is constructed from new document set by previous described process. In *Check and Update* process, the domain ontology is compared with the new-document-generated ontology to check which new information is in the new-document-generated ontology and use them to update the domain ontology. The new information includes new instances, new attributes, new operations and new associations. The new instance may contain different attributes, operations and associations from its concept, so the new instance and the different attributes, operations and associations will be added to the original concept to update the domain ontology. Figure. 8 illustrates the concept update process.

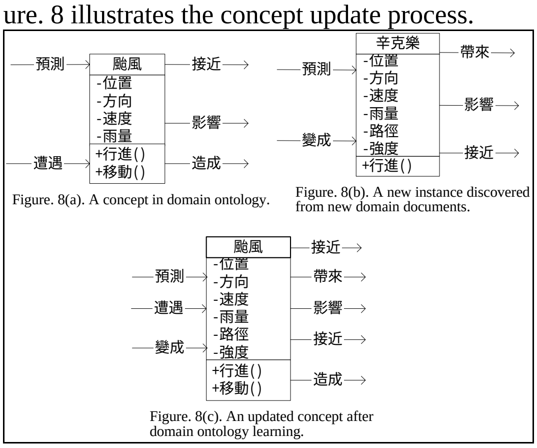

Figure. 8. The concept update process.

4. Experimental Results and analysis

In this section, some experiments are made to evaluate the performance of the proposed approach. We take the Chinese 2002 FIFA World Cup news (from udn.com) and typhoon news (from news.chinatimes.com) as domain data. In *2002 FIFA World Cup* domain, we use 440 documents to construct ontology and use another 439 documents to do domain ontology learning.

In *Typhoon* domain, we use 93 documents of 2001 news to construct ontology and use 279 documents of 2002 news to do domain ontology learning.

We set *Min* = 7 and *Win* = 10 to extract episodes from new document set of these two domains. Then, we take the revised and more detail domain ontology (*Min* = 3, *Win* = 10) to compare with episodes and extract candidate instances in order to extract more attributes, operations, and associations of candidate instances. In *2002 FIFA World Cup* domain, there are 72 candidate instances extracted. In *Typhoon* domain, there are 53 candidate instances extracted. We use the learned knowledge base of fuzzy inference to evaluate the conceptual resonance strength of candidate instances. Because the results are not good enough, the extracted new instances are revised by domain expert then add them to the domain ontology. After we add the new instances, the new–document–generated ontology is constructed. Next, we will compare the original domain ontology with new–document–generated ontology to check the new information. Finally, we add the new information into original domain ontology. When gather enough amount of new domain documents, the domain ontology can be updated by this incremental domain ontology learning method and do not need to reconstruct it again. We invite three domain examiners to judge the results. Table 3 and Table 4 show the correct number of new information we extract from new document set. This approach can extract new information to enrich the domain ontology.

In concept-layer, in *2002 FIFA World Cup* domain, little new information is extracted because the domain knowledge is static. In *typhoon* domain, more new information is extracted because the domain knowledge is dynamic (we use news in difference years). In instance-layer, this approach can make the knowledge base more complete in these two domains.

Table 3. Number of new information generated by *Domain Ontology Learning* in *2002 FIFA World Cup* domain (*Min* = 7, *Win* = 10).

Concept layer	Attribute	Operation	Association
Examiner A	17	5	1
Examiner B	20	4	2
Examiner C	19	3	1
Instance layer	**Attribute**	**Operation**	**Association**
Examiner A	108	17	7
Examiner B	94	16	7
Examiner C	86	11	11

Table 4. Number of new information generated by *Domain Ontology Learning* in *Typhoon* domain (*Min* = 7, *Win* = 10).

Concept layer	Attribute	Operation	Association
Examiner A	18	12	7
Examiner B	14	9	13
Examiner C	19	10	9
Instance layer	**Attribute**	**Operation**	**Association**
Examiner A	29	17	35
Examiner B	35	18	40
Examiner C	36	14	28

5. Conclusions and Future Work

In this paper, we propose an incremental domain ontology learning method. This method can effectively extract new information from new domain documents to update the schema of domain ontology and make the knowledge based of domain ontology more complete based on a constructed domain ontology. First, we use schema of domain ontology to extract candidate instances. We also use genetic algorithm to learn the knowledge of fuzzy inference. The three-layer parallel fuzzy inference mechanism is further applied to obtain new instances for ontology learning. In addition, new attributes, operations, and associations will be extracted based on episodes and morphological analysis to update the domain ontology.

Our approach has provided some automatic solutions for helping human to construct ontology. However, those approaches are not perfect. Therefore, there are two advanced technologies will be necessarily developed in the future. First, extend this proposed method to cross languages not only for Chinese. Second, make ontology learning dynamically, that is the concept-layer ontology can be updated with times. Finally, we hope the ontology will be appropriately introduced into many applications e.g., Semantic Web annotation or ontology inference in information systems.

Acknowledgment

This research was partially supported by the III Innovative and Prospective Technologies project of Institute for Information Industry and sponsored by MOEA ,R.O.C, 2003.

Reference

[1] Academia Sinica, Chinese Electronic Dictionary, in *Technical Report* (93-05), Taiwan, 1993.

[2] H. Ahonen, O. Heinonen, M. Klemettinen, A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections", in *Proc*. Advances in Digital Libraries Conference, pp. 2-11, 1998.

[3] O. Cordon, F. Herrera, and P. Villar, "Generating the Knowledge Base of a Fuzzy Rule-Based System by the Genetic Learning of the Data Base," *IEEE Trans. Fuzzy Systems*, vol. 9, no. 4, pp. 667-674, 2001.

[4] T. Gruber, "What is An Ontology?," URL Accessed on November 9, 2001, http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

[5] L. Khan and F. Luo, "Ontology Construction for Information Selection," in *Proc*. The 14th IEEE International Conference on Tools with Artificial Intelligence, pp. 122-127, 2002.

[6] Y.H. Kuo, J.P. Hsu and C.W. Wang, "A Parallel Fuzzy Inference Model with Distributed Prediction Scheme for Reinforcement Learning," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 28, no. 2, pp. 160-172, 1998.

[7] C.S. Lee, Y.F. Kao, Y.H. Kuo, and I.H. Meng "An Episode-based Fuzzy Inference Mechanism for Chinese News Ontology Construction," in *Proc*. The 7th World Multiconference on Systemics, Cybernetics and Informatics, pp. 453-458, 2003.

[8] C.T. Lin and C.S.G. Lee, "Neural-Network-Based Fuzzy Logic Control and Decision System," *IEEE Trans. Computers*, vol. 40, no. 12, pp. 1320-1336, 1991.

[9] A. Maedche and S. Staab, "Ontology Learning for the Semantic Web," *IEEE Trans. Intelligent Systems*, vol. 16, no. 2, pp. 72-79, 2001.

[10] H. Mannila, H. Toivonen, A.I. Verkamo, "Discovery of frequent episodes in event sequences", *International Journal of Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 259-289, 1997.

[11] M. Missikoff, R. Navigli, P. Velardi, "Integrated approach to web ontology learning and engineering", *IEEE Trans. Computer*, vol. 35, no. 11, pp. 60-63, 2002.

[12] B. Omelayenko, "Learning of Ontologies for the Web: the Analysis of Existent Approaches," in *Proc*. International Workshop on Web Dynamics held in conj. with the 8th International Conference on Database Theory (ICDT'01), Jan, 2001.

[13] V.W. Soo, C.Y. Lin, "Ontology-based information retrieval in a multi-agent system for digital library", in *Proc*. the 6th Conference on Artificial Intelligence and Applications, pp. 241-246, 2001.

[14] L.X. Wang and J. M. Mendel, "Generating Fuzzy Rules by Learning from Examples," *IEEE Trans. Systems Man, and Cybernetics*, vol. 22, no. 6, pp. 1414-1427, Nov/Dec 1992.

[15] L. Zhou, Q. E. Booker, D. Zhang, "ROD – Toward Rapid Ontology Development for Underdeveloped Domains", in *Proc*. the 35th Annual Hawaii International Conference on System Sciences, 2002.