A Semantic-based Concept Clustering Mechanism for Chinese News Ontology Construction

Chang-Shing Lee*, Chia-Hsin Liao** and Yau-Hwang Kuo**
*Department of Information Management, Chang Jung University, Tainan, Taiwan
**CREDIT Research Center, National Cheng Kung University, Tainan, Taiwan
E-mail: leecs@mail.cju.edu.tw / leecs@cad.csie.ncku.edu.tw

Abstract

In order to efficiently manage and use knowledge, the technologies of ontology are widely applied to various kinds of domain knowledge. This paper proposes an automatic concept construction approach that can help knowledge manager efficiently construct a Chinese News domain knowledge. The first feature of this paper is to utilize an object-oriented approach to represent the structure of ontology, called object-oriented ontology. Second, the automatic ontology construction approach for Chinese News domain is presented. We embed CKIP system to carry out the Chinese natural language processing including part-of-speech tagging, Chinese-Term analysis and Chinese-Term feature selection. Third, the concept clustering mechanisms for Chinese News domain ontology construction based on fuzzy compatibility relation approach are proposed. Furthermore, the parallel fuzzy inference mechanism is also adopted to infer the conceptual resonance strength of any two Chinese terms. By the simulation, the proposed mechanism can efficiently cluster the semantic concept for Chinese News ontology construction.

Keywords: ontology construction, feature selection, fuzzy inference, concept clustering.

1 Introduction

In the development of a knowledge-based system, the use of ontology is beneficial for two reasons. It allows for a more disciplined design of the knowledge base; and it facilitates sharing and reuse. Both advantages are particularly important when the knowledge base becomes large [1]. Research on ontology is becoming increasingly widespread in the computer science community. While this term has been rather confined to the philosophical sphere in the past, it is now gaining a specific role in Artificial Intelligence, Computational Linguistics, and Database Theory. In particular, its importance is being recognized in research fields as diverse as knowledge engineering, knowledge representation, qualitative modeling, language engineering, database design, information modeling, information integration, object-oriented analysis, information retrieval and extraction, knowledge management and organization, and agent-based systems design. Current applications areas of ontology are disparate, including enterprise integration, natural language translation, medicine, mechanical engineering, standardization of product knowledge, electronic commerce, geographic information systems, legal information systems, and biological information systems [2].

In this paper, we propose a concept clustering mechanism for Chinese News domain ontology construction. We embed CKIP system to carry out the Chinese natural language processing including part-of-speech tagging, Chinese-term analysis and Chinese-term feature selection. Furthermore, the parallel fuzzy inference mechanism is also adopted to infer the conceptual resonance strength of any two Chinese terms. The organization of this paper is as follows. The object-oriented structure of ontology representation will be presented in Section 2. A semantic-based concept clustering mechanism for domain ontology construction is proposed in Section 3. Section 4 shows the experimental results. Finally, some conclusion is presented in Section 5.

2 Object-Oriented Structure of Domain Ontology Representation

Ontology is a specification of an abstract, simplified view of the world that we wish to represent for some purpose. Therefore, ontology defines a set of representational terms that we call concepts. Inter-relationships among these concepts describe a target world. In this section, we propose an object-oriented structure of ontology and exhibit its architecture in figure 1.

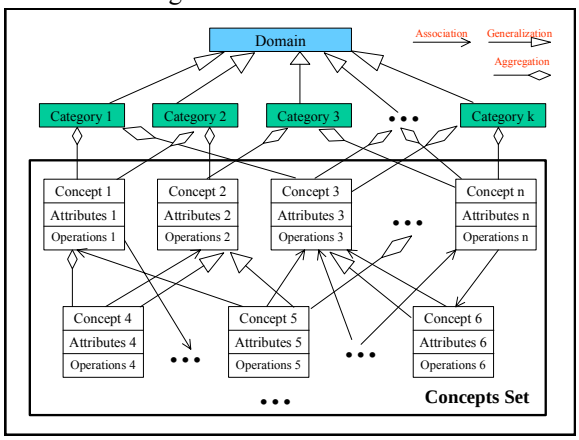

Figure 1. Object-Oriented Ontology Architecture.

The defined object-oriented ontology is consisted of several basic components, and we describe them as follows.

- Domain

The top layer of the ontology architecture is the name of domain knowledge. Because this ontology is constructed by domain knowledge, so it represents a domain ontology. In this paper, the ontology is

constructed for Chinese News document, so its domain name is Chinese News.

- Category

The second layer is the categories of domain ontology. Each category inherits the domain properties from the domain knowledge, so the relationship between domain and category is a generalization. Furthermore, we also define the aggregation of relationship between each category and concepts. There are seven categories of Chinese News domain ontology in this paper. They are "Political(政治焦點)", "International(國際要聞)", "Finance(股市財經)", "Cross-Strait(兩岸風雲)", "Societal(社會地方)", "Entertainment(運動娛樂)" and "Life(生活新知)".

- Concept, Attribute and Operation

The following layers are the architecture of the concept hierarchy. We use the object-oriented approach to represent the ontology architecture. We treat each concept node in this ontology as a class, so this structure mode can be treated as a class diagram. Therefore, each concept node contains its concept name, attributes and operations. Furthermore, the relationships between the concepts may be association, generalization or aggregation.

3 A Semantic-based Concept Clustering Mechanism for Domain Ontology Construction

In this paper, we develop an automatic approach to construct the semantic concepts for Chinese News ontology. First, we use natural language processing technologies to deal with the Chinese News documents that we gathered from WWW. In documents pre-processing, we propose several technologies that contain part-of-speech tagging, refining tagging, stop word filter and term analyzer. We also use some tools in this task such as *CKIP* [3], *Academia Sinica Balanced Corpus* [4] and *Chinese Electronic Dictionary* [5] that help us to deal with documents. Then we will get the Chinese-Terms features to construct Chinese News domain ontology. In addition, the concepts clustering approach based on fuzzy compatibility relation is also proposed. Figure 2 shows the construction architecture for Chinese News ontology. There are several main modules and databases in this architecture.

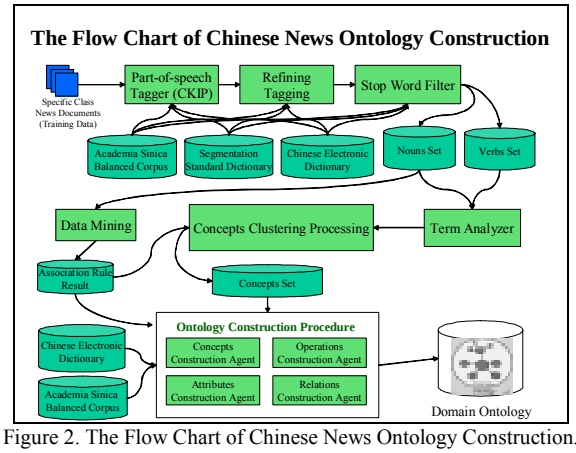

Figure 2. The Flow Chart of Chinese News Ontology Construction.

First, the specific class News documents will be previously tagged resulting in each word with its appropriate part-of-speech tag. The *CKIP* will be used in this step. Second, in refining tagging step, we refer to *Academia Sinica Balanced Corpus* and *Chinese Electronic Dictionary* to refine the part-of-speech tags. Therefore, we will have sufficient knowledge in Chinese part-of-speech to analyze what are the useful features in ontology construction. In third step, the stop word filer module will select the useful nouns and verbs terms as the candidate features after our analysis. The fourth step, the term analyzer will analyze the document frequency to help us to select important or representative words from a specific class documents. Concepts clustering play a major role in automatic ontology construction. So, in fifth step, we will evaluate the conceptual resonance of two Chinese terms based on parallel fuzzy inference mechanism [6].

In Chinese terms analysis, we select the more meaningful Chinese nouns in each document such as: Na (普通名詞), Nb (專有名詞) and Nc (地方名詞). Moreover, we filter the un-meaningful nouns such as: Nd (時間名詞), Ne (定詞), Nf (量詞), Ng (方位詞) and Nh (代名詞) for the Chinese News classification.

4 Evaluation of Conceptual Resonance of Two Chinese Terms for Concept Clustering

Conceptual resonance means the degree of the same concept between two different terms. Hence, two Chinese terms will have a higher possibility with the same concept if they have stronger conceptual resonance strength. In this paper, we proposed four fuzzy variables for conceptual resonance strength of any two Chinese terms, they are resonance in part-of-speech, resonance in term vocabulary, resonance in term association and resonance in common term association. Now we describe them as follows.

4.1 Resonance in Part-of-speech

The first fuzzy variable of conceptual resonance is resonance in part-of-speech (POS). First we define the tagging tree according the POS. Figure 3 shows the tagging tree structure that will be used to compute the resonance of POS for any two Chinese terms.

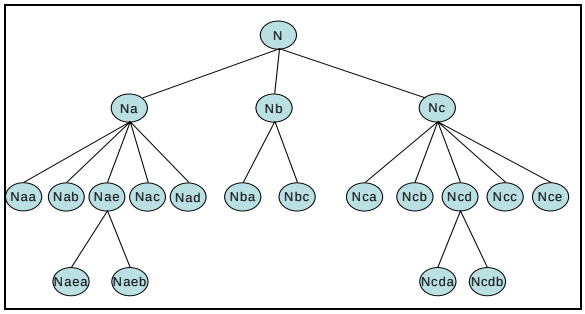

Figure 3. The Framework of Tagging Tree Diagram.

The resonance will be strong when the path

distance of any two Chinese terms is near. For example, there are two terms with their POS are "電腦 computer (Nab)" and "軟體 software (Nac)", and the path distance of two terms is *2* (Nab -> Na -> Nac).

4.2 Resonance in Term Vocabulary

In the viewpoint of characteristic of Chinese language, any two terms with more common words, they will be more similar in semantic meaning. For example, each Chinese term in the terms set {民進黨, 民進黨團, 民主進步黨} has similar semantic meaning since they are composed with the common words "民", "進" and "黨". Besides, we also consider another characteristic of Chinese terms in side of term vocabulary. It assumes that because almost every Chinese word is a morpheme with its own meaning, very often terms having the same starting or ending word share some common linguistic properties and, thus, can form a term cluster [8][9]. The good examples of starting and ending word are following two terms sets: {星期一 (Monday), 星期六 (Saturday), 星期日 (Sunday)} and {昨天 (yesterday), 明天 (tomorrow), 今天 (today), 每天 (everyday)}. We propose the evaluative approach of common words and common starting/ending word to decide the similarity of resonance in term vocabulary as follows. We define strength of similarity be the count of common words number and increasing 0.5 extra strength if they have common starting or ending word. We give an example to evaluate the strength of two terms "民進黨團" and "民主進步黨" as follows. The evaluation method is divided into two parts "evaluation in common words" and "evaluation in common starting or ending word". They have three common words "民", "進" and "黨", and a common starting word "民". So the total strength is *3.5*.

4.3 Resonance in Term Association

A large amount of previous research has focused on how to best cluster similar terms together. The proposed methods can be roughly grouped into two categories: knowledge based clustering and data-driven clustering [9]. In Section 4.1 and 4.2, the information for concepts clustering we just focus on term's knowledge. However, we believe the information of terms knowledge themselves isn't still enough for concepts clustering.

Sometimes, the terms have the similar meaning and without any common properties of their knowledge, hence we must analyze their documents. We used the confidence value between two terms to decide the strength of term relation. The terms with large confidence mean that they have strong relationship and further could be combined into one concept. It will be supplement information for concepts clustering when terms have insufficient in their knowledge but they are similar. Here, we give an example of terms set {總統 (President) (Nab), 總統府(The Office of the President) (Nca), 陳水扁(President Chen) (Nb)}. In the human's viewpoint, these three terms represent similar concept so they will be clustered into together. But in the term's knowledge viewpoint, just only "總統(President)" and "總統府(The Office of the President)" two terms will be clustered in the same concept, and the term "陳水扁 (President Chen)" will be not clustered into the concept {總統(President), 總統府(The Office of the President)}. So we proposed the evaluative method to decide the strength of resonance in term association. The strength is decided by the confidence value between two terms, so we adopt the average of confidence to be the term association strength. For example, we adopt the two terms {總統(Nab), 陳水扁 (Nb)} from the News category "Political(政治焦點)". The confidence of (總統 -> 陳水扁) is *0.84*, and the confidence of (陳水扁 -> 總統) is *0.80*. Therefore, the strength of resonance in term association is *(0.84+0.80)/2 = 0.82*.

4.4 Resonance in Common Term Association

The common term association represents the strength of two terms according to common term numbers in their corresponding documents. For any two Chinese terms with the same common words or starting/ending words, they may not have the similar meaning. For example, consider the three Chinese terms "美國(U.S.A.)", "美方(U.S.A.)" and "警方 (police)", the Chinese term "美方(U.S.A.)" has common starting word "美" with "美國" and also has common ending word "方" with "警方(police)". The common document terms with a specific threshold of confidence for "美國", "美方" and "警方" are as follows:

- *美國(U.S.A.) -> {白宮(White House), 布希 (Bush), 紐約(New York)}*
- *美方(U.S.A.) -> {白宮(White House), 布希 (Bush), 五角大廈(Pentagon)}*
- *警方(police) -> {警員(policeman), 刑事組 (criminal investigation), 分局(police station)}*

Therefore, the common terms for {美國(U.S.A.), 美方 (U.S.A.)} and {警方(police), 美方(U.S.A.)} are as follows:

- *{美國(U.S.A.), 美方(U.S.A.)} -> {白宮(White House), 布希(Bush)}, the strength is 2.*
- *{警方(police), 美方(U.S.A.)} -> Null, the strength is 0.*

Therefore the term pair {美國(U.S.A.), 美方 (U.S.A.)} has stronger resonance strength than the term pair {警方(police), 美方(U.S.A.)} in common term association.

5 A Parallel Fuzzy Inference Network for Semantic-based Concept Clustering

A parallel fuzzy inference network for semantic-based concept clustering is proposed in this section. The fuzzy variables for computing the

conceptual resonance of any two Chinese terms will be discussed in the following subsections.

5.1 Aggregate Term Resonance with Parallel Fuzzy Inference Network

In this subsection, we describe how to aggregate four input fuzzy variables into one output fuzzy variable for computing the conceptual resonance strength for each term pair. Now we define the fuzzy variables and theirs linguistic terms as follows. There are four input fuzzy variables including *Part-of-speech Similarity (POS)*, *Term-Vocabulary Similarity (TV)*, *Term-Association Strength (TA)* and *Common Term-Association Strength (CTA)*, and one output fuzzy variable *Conceptual Resonance Strength (CRS)* used in this architecture. Figure 4 shows the fuzzy sets *{POS_Low, POS_High}* for fuzzy variable *POS similarity*. There are two linguistic terms *POS_Low* and *POS_High* defined in the fuzzy variables.

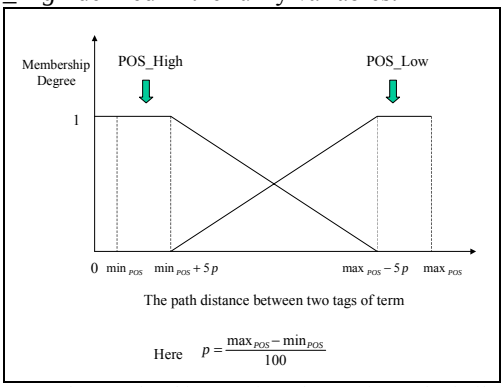

Figure 4. The Membership Function of Part-of-speech Similarity.

Figure 5 shows the membership functions of fuzzy sets *{TV_Low, TV_High}* for fuzzy variable *TV similarity*. There are two linguistic terms *TV_Low* and *TV_High* defined in the fuzzy variable.

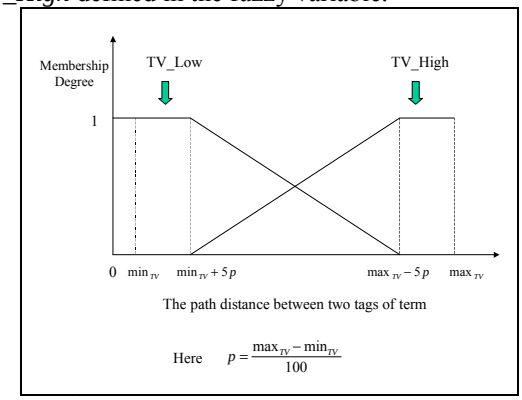

Figure 5. The Membership Function of Term Vocabulary Similarity.

The membership function of fuzzy sets *{TA_Low, TA_Medium, TA_High}* for fuzzy variable *TA strength* are show in figure 6. There are three linguistic terms including *TA_Low*, *TA_Medium* and *TA_High* defined in the fuzzy variables.

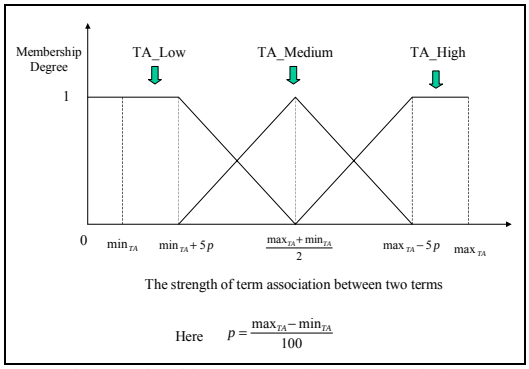

Figure 6. The Membership Function of Term Association Strength.

Figure 7 shows the membership functions of fuzzy sets *{CTA_Low, CTA_Medium, CTA_High}* for fuzzy variable *CTA strength*. There are three linguistic terms including *CTA_Low*, *CTA_Medium* and *CTA_High* defined in the fuzzy variables.

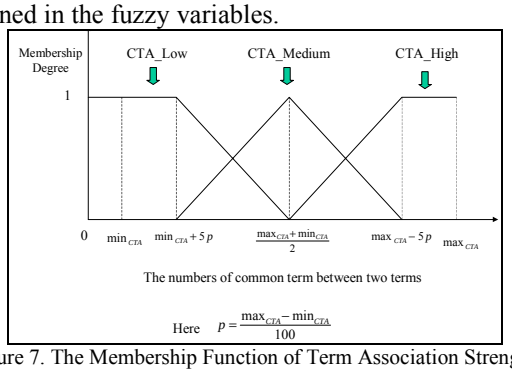

Figure 7. The Membership Function of Term Association Strength.

Figure 8 shows the membership functions of fuzzy sets *{CRS_Very Low, CRS_Low, CRS_Medium, CRS_High, CRS_Very High}* for fuzzy variable *CRS strength*. There are five linguistic terms including *CRS_Very Low*, *CRS_Low*, *CRS_Medium*, *CRS_High* and *CRS_Very High* defined in the fuzzy variables.

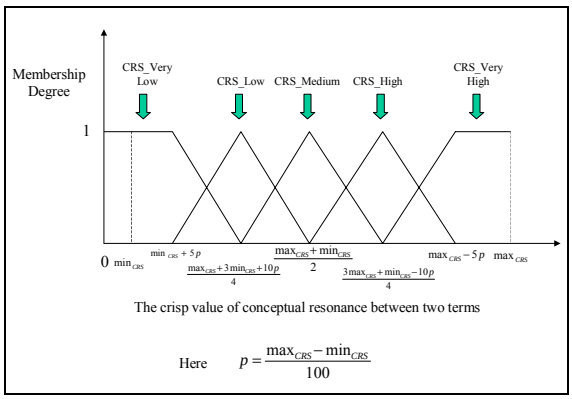

Figure 8. The Membership Function of Conceptual Resonance Strength.

After describing the fuzzy variables for computing the conceptual resonance of any Chinese term pair, we will propose the parallel fuzzy inference architecture for semantic concept clustering. Figure 9 shows the architecture.

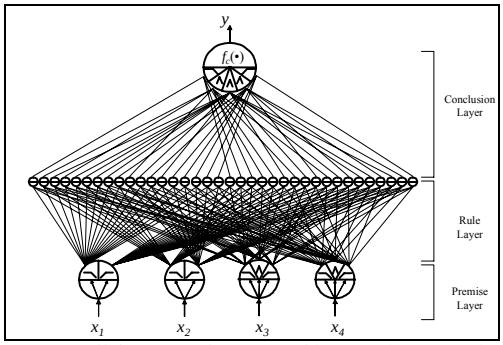

Figure 9. A Parallel Fuzzy Inference Architecture for Semantic-based Concept Clustering.

This structure is a three-layered network which can be constructed by directly mapping from a set of specific fuzzy rules, or learned incrementally from a set of training patterns. Here, the rules have been defined by expert's knowledge. The structure consists of premise layer, rule layer and conclusion layer. There are two kinds of nodes in this model: fuzzy linguistic nodes and rule nodes. A fuzzy linguistic node represents a fuzzy variable and manipulates the information related to that linguistic variable. A rule node represents a rule and decides the final firing strength of that rule during inferring. The premise layer performs the first inference step to compute matching degrees. The conclusion layer is responsible for making conclusion and defuzzification. Now, we will describe each layer in details.

- Premise layer:

The first layer is called as premise layer, which is used to represent the premise part of the fuzzy system we describe. Each fuzzy variable appearing in the premise part is represented with a condition node. Each of the outputs of the condition node is connected to some nodes in the second layer to constitute a condition specified in some rules. Note that the output links must be emitted from proper linguistic terms as specified in fuzzy rules. In other words, a linguistic node is a polymorphic object that can be viewed from different aspects by different fuzzy rules.

The premise layer performs the first inference step to compute matching degrees. The input vector is $x=(x_1,x_2,...,x_n)$, where x_i is denoted as the input value of *i*th linguistic node. Then, the output vector of the premise layer will be

$$\mu^1=((u^1_{11},u^1_{21},...,u^1_{N_11}),(u^1_{12},u^1_{22},...,u^1_{N_22}),...,(u^1_{1n},u^1_{2n},...,u^1_{N_nn}))$$

where u^1_{ij} is the matching degree of the *j*-th linguistic term in the *i*-th condition node. In this paper, triangular function and trapezoidal function are adopted as the membership functions of linguistic terms. For triangular and trapezoidal normal membership functions, Eq. 1 and Eq. 2 can be realized by the following formula:

$$f_{triangle}(x:a,b,c)=\begin{cases}0 & x<a\\ (x-a)/(b-a) & a\le x\le b\\ (c-x)/(c-b) & b\le x\le c\\ 0 & x>c\end{cases} \quad (1)$$

$$f_{trapezoidal}(x:a,b,c)=\begin{cases}0 & x<a\\ (x-a)/(b-a) & a\le x\le b\\ 1 & b\le x\le c\\ (d-x)/(d-c) & c\le x<d\\ 0 & x\ge d\end{cases} \quad (2)$$

$$f^1_{ij}=\begin{cases}f_{triangular} & j\ne 1 \text{ or } n\\ f_{trapezoidal} & j=1 \text{ or } n\end{cases} \quad (3)$$

where *n* is the numbers of linguistic term for *i*-th linguistic node. Therefore, for each element μ^1_{ij} of output vector μ^1 is

$$\mu^1_{ij}=f^1_{ij}(x) \quad (4)$$

- Rule layer:

The second layer is called as 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 logic rules. And the output of a rule node in rule layer will be linked with associated linguistic nodes in the third layer. In our model, the rules are defined by expert's knowledge previously. In rule node, f_r function provides the net input for this node like the Eq. 5.

$$f_r=\sum_{i=1}^{p}w_i\mu_i \quad (5)$$

- Conclusion layer:

The third layer is called conclusion layer. This layer is also composed of a set of fuzzy linguistic nodes. The fuzzy linguistic node can also operate in a reverse mode, called conclusion node. It should be noted that the reverse mode is only invoked in the rule inference phase. At this phase, the links representing linguistic term set become incoming links, while the link representing base variable becomes outgoing link. In the reverse mode, fuzzy linguistic nodes are responsible for making conclusion and defuzzification.

At the end of inference process, defuzzification may be necessary. In our model, the final output *y* is the crisp value that is produced by combining all inference results with their firing strength. Eq. 6 represents the formula of defuzzification.

$$CrispOutput=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}y^k_{ij}w^k_{ij}V_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c}y^k_{ij}w^k_{ij}} \quad (6)$$

Where $w^k=\frac{\sum_{i=1}^{n}\mu^1_i}{n}$,

V_{ij} is the center of gravity, *r* is the numbers of corresponding rule nodes, *c* is the numbers of linguistic terms of output node, *n* is the numbers of the fuzzy variable in premise layer and *k* represents in *k*-th layer. Therefore in our case, the value of *r*, *c*, *n* and *k* are *36*, *5*, *4* and *2*.

5.2 Concepts Clustering Based on Fuzzy Compatibility Relation Approach

After fuzzy inference and defuzzification processing, we will get the crisp values of conceptual resonance for all term pairs. We consider the conceptual resonance of terms as fuzzy compatibility relation, because they satisfied the properties of reflexive and symmetric. Therefore, the problem of concepts clustering is equal to find all classes of maximal α-compatibles with fuzzy compatibility relation. Here, α represents a specified membership degree of fuzzy compatibility relation. Then we propose the concepts clustering algorithm based on fuzzy compatibility relation approach as follows:

Concept Clustering Algorithm based on Fuzzy Compatibility Relation Approach

Input:

1. Fuzzy Compatibility Membership Degree α

2. The Term Set $X = \{Term[1], Term[2], ..., Term[n]\}$ with n Terms for the Specific Category News, and it's Corresponding Fuzzy Conceptual Resonance Matrix $A = [\alpha_{ij}]_{n \times n}$.

Output:

The *Final_Concepts_Set*, that is the set of Domain Ontology Concepts.

Method:

Step 1: *For* $i \leftarrow 1$ *to n*

Step 1.1: $Set_i \leftarrow \Phi$ /* Set_i denotes the Term Set regarding with $Term[i]$, and all the compatibility membership degree α_{ij} of the terms in Set_i are not less than α */

Step 1.2: $S_i \leftarrow 0$ /* S_i denotes the cardinality of Set_i */

Step 1.3: $Set_i \leftarrow Set_i \cup \{Term[i]\}$

Step 1.4: $Temp\_Set \leftarrow \Phi$, /* $Temp\_Set$ denotes the set of existing Concepts subsets*/

Step 1.5: *For* $j \leftarrow i$ *to n*

Step 1.5.1: If $\alpha_{ij} \geq \alpha$ Then

Step 1.5.1.1: $Set_i = Set_i \cup \{Term[j]\}$

Step 1.5.1.2: $S_i \leftarrow S_i + 1$

Step 1.6: Determine the power set p_k of Set_i.

Step 1.6.1: $S_{p_k} \leftarrow |p_k|$, where $k = 1, ..., 2^{S_i}$

/* S_{p_k} Denotes the cardinality of p_k */

Step 1.7: For $k \leftarrow 1$ to 2^{S_i}

Step 1.7.1: If $p_k \in Temp\_Set$

Continue

Step 1.7.2: $flag \leftarrow 0$

Step 1.7.3: For $l \leftarrow 1$ to $S_{p_k} - 1$

Step 1.7.3.1: For $m \leftarrow l + 1$ to S_{p_k}

Step 1.7.3.1.1:

$n \leftarrow$ Index of $p_k[l]$ in X

$q \leftarrow$ Index of $p_k[m]$ in X

Step 1.7.3.1.2: If $\alpha_{nq} < \alpha$

Then $flag \leftarrow 1$ and Break

Step 1.7.3.2: If $flag = 1$

Then Break

Step 1.7.4: If $flag = 0$ Then

Step 1.7.4.1:

$Final\_Concept\_Set \leftarrow Final\_Concept\_Set \cup p_k$

Step 1.7.4.2:

$Temp\_Set \leftarrow Temp\_Set \cup \{P(p_k) - p_k - \Phi\}$

/* $P(p_k)$ Denotes the power set of p_k */

Step 2: End.

In our approach, how to decide α is very important. The α value will influence the number of concepts and the compatibility degree of terms for specific concept. The lower α value will form much number of concepts, and strengthen the compatibility degree of terms for specific concept. Therefore, we must decide the scope of number of concepts that are suitable for each News category first. Here, prune-and-search strategy will be used in the problem of α decision. In beginning, we adopt the median of conceptual resonance as the α value. Then concepts will be formed by α. According to the number of concepts, we will adjust α until it satisfies the scope we decide for each News category.

6 Experiment Results and Analysis

In this section, some experiment are made to test the performance of the proposed approach. There are seven categories of News including "Political" (政治焦點), "International" (國際要聞), "Finance" (股市財經), "Cross-Strait" (兩岸風雲), "Societal" (社會地方), "Entertainment" (運動娛樂) and "Life" (生活新知) used for the experiments. Those documents are gathered from China Times website and the period is between May 2001 from March 2002. After the tagging by *CKIP* and refining tagging processing, the stop word filter module will filter the stop words from documents. Table 1 lists the filter percent and remaining terms for each News category.

Table 1. The Experimental Results for the Proposed Filter.

News Category	政治焦點 (Political)	國際要聞 (International)	股市財經 (Finance)
Number of Doc.	11277	13542	22756
All Terms	25448	25484	18960
Remaining Terms	17091	15367	11346
Filter Percent	32.84%	39.70%	40.16%
兩岸風雲 (Cross-Strait)	社會地方 (Societal)	運動娛樂 (Entertainment)	生活新知 (Life)
6040	13441	5974	9279
22856	35846	24178	35932
15085	24813	16543	24287
34.00%	30.78%	31.58%	32.41%

After stop word filter, we can select the features from remaining terms, then cluster the semantic concept for ontology construction.

Next, we will analyze the results of conceptual resonance for any Chinese term pair. Figure 10(a) shows the partial result of conceptual resonance for the News category "Political" (政治焦點) with highest

values. Notice that each term pair not only has strong similarity in term knowledge (*POS* and *TV*) but also strong strength in term association and common term association.

	1	(民主進步黨,民進黨) 0.595481543
	2	(親民黨主席,親民黨) 0.594101448
	3	(國民黨主席,國民黨) 0.587348993
	4	(民進黨主席,民進黨) 0.581987023
	5	(親民黨主席,國民黨主席) 0.577182427
	6	(行政院長,行政院) 0.571720293
	7	(民進黨主席,民主進步黨) 0.571574542
	8	(立院,立法院) 0.56774317
	9	(立法院,立法院長) 0.558938037
	10	(民進黨團,民進黨) 0.558824035
	11	(台北市,台北市長) 0.557260479
	12	(民進黨籍,民進黨) 0.555527542
	13	(民進黨主席,國民黨主席) 0.554741061
	14	(高雄市,高雄) 0.553642597
	15	(台聯,台灣團結聯盟) 0.553602148
	16	(國民黨,民進黨) 0.551060166
	17	(國民黨,國民黨籍) 0.547553789
	18	(親民黨主席,民進黨主席) 0.54062093
	19	(民進黨籍,國民黨籍) 0.53928488
	20	(民進黨主席,黨主席) 0.538797148

(a)

*	26	(國民黨主席,連戰) 0.534303235
	27	(國民黨,親民黨) 0.534003082
	28	(民進黨籍,民進黨團) 0.5338768
	29	(副秘書長,秘書長) 0.531714804
*	30	(親民黨主席,宋楚瑜) 0.530113977
	31	(立委,立法委員) 0.52974897
	32	(國防部,國防) 0.529134478
*	33	(馬英九,台北市長) 0.522650369
	34	(親民黨,民進黨) 0.522529795
	35	(委員會,委員) 0.522146828
	36	(縣市長,縣市) 0.520918138
	37	(政府,中央政府) 0.518259497
*	38	(呂秀蓮,副總統) 0.51637767
	39	(民主進步黨,民進黨團) 0.516346569
*	40	(李登輝,前總統) 0.515308468
*	41	(陳水扁,總統) 0.513920884
	42	(副院長,立法院長) 0.512207578
	43	(民主進步黨,國民黨) 0.511493131
	44	(總統府,總統) 0.507264737
	45	(執政黨,政黨) 0.505084742
*	46	(王金平,立法院長) 0.504385767

(b)

Figure 10. The Partial Result of Conceptual Resonance for "Political" (政治焦點) with Highest Values.

Figure 10(b) shows the partial result of conceptual resonance for "Political(政治焦點)", and each term pair still has higher value. The term pairs marked by asterisks represent that they have stronger strength in *TA* or *CTA* but without strong similarity in *POS* or *TV*. Table 2 shows the results of concept clustering with various α for the News category "Life" (生活新知).

Table 2. The Analysis of Various α Values for Each News Category.

α =0.435	Concept 1	教育 教師 教授 教育部長
	Concept 2	學校 學生 學術 大學 台灣大學
	Concept 3	學者 學術 大學 教授
α =0.417	Concept 1	教育 教師 教授 教育部長 學生 學校 教育部
	Concept 2	學校 學生 學術 大學 台灣大學 校長
	Concept 3	學者 學術 大學 教授 研究
	Concept 4	學校 家長 老師 學生
	Concept 5	科學 學術 大學
	Concept 6	學者 科學家 專家
α =0.397	Concept 1	教育 教師 教授 教育部長 學生 學校 教育部 大學 課程 資源
	Concept 2	學校 學生 學術 大學 台灣大學 校長 院長 教授
	Concept 3	學者 學術 大學 教授 研究 成果 領域
	Concept 4	學校 家長 老師 學生 教育部長 教育部 高中 大學
	Concept 5	科學 學術 大學 研究所 研究
	Concept 6	學者 專家 科學家 科學
	Concept 7	學者 科學 學生 生物 領域 學術 大學
	Concept 8	成果 研究 科學 領域 學術
	Concept 9	技術 研究 領域 應用 產業

Notice that the concepts with higher α value are the subset of the concepts with lower α value. That is, the lower α value will generate the semantic concept with more Chinese terms.

7 Conclusions

A semantic-based concept clustering mechanism for Chinese News ontology construction is propoed in this paper. The structure of the object-oriented domain ontology is also proposed. In addition, the CKIP provided by Academia Sinica is embedded for Chinese natural language processing. Furthermore, a parallel fuzzy inference mechanism for computing the similarity of any Chinese term pair is presented. By the experimental results, the proposed approach can effectively cluster the semantic concept for the Chinese terms. In the future, we will extend our approach to help construct domain ontology more efficient. Moreover, the Chinese/English documents will also be considered to construct more complex domain ontology.

Reference

[1] P.E. van der Vet and N.J.I. Mars, "*Bottom-Up Construction Ontologies*", IEEE Trans. on Knowledge and data Engineering, Vol. 10, No. 4, pp.513-526, July/August, 1998.

[2] N. Guarino, "Formal Ontology and Information System," Proc. of the First International Conference (FOIS'98), Trento, Italy, June, 1998.

[3] "*Chinese Knowledge Information Processing Group* (*CKIP*)," Academia Sinica, Taiwan, 2001.

[4] "*Academia Sinica Balanced Corpus*," Technical Report, No. 95-02/98-04, Academia Sinica, Taiwan, 1998.

[5] "*Chinese Electronic Dictionary*," Technical Report, No. 93-05, Academia Sinica, Taiwan, 1993.

[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. on Systems, Man, and Cybernetics, Vol. 28, No. 2, pp.160-172, April, 1998.

[7] Y. J. Yang, et al., "An intelligent and efficient word-class-based Chinese language model for Mandarin speech recognition with very large vocabulary." Proc. of ICSLP-94, Yokohama, Japan, 1994, pp. 1371-1374.

[8] J. Gao, J. T. Goodman and J. Miao, "The Use of Clustering Techniques for Language Modeling – Application to Asian Language," Computational Linguistics and Chinese Language Processing , Vol. 6, No. 1, pp. 27-60, February, 2001.

[9] R. C. T. Lee, R. C. Chang, S. S. Tseng and Y. T. Tsai, *Introduction to the Design and Analysis of Algorithms*, Taipei, Unalis co., 1999.