Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 277-302

A Chinese Term Clustering Mechanism for Generating Semantic Concepts of a News Ontology

Chang-Shing Lee[*], Yau-Hwang Kuo[+], Chia-Hsin Liao[+] and Zhi-Wei Jian[+]

Abstract

In order to efficiently manage and use knowledge, ontology technologies are widely applied to various kinds of domain knowledge. This paper proposes a Chinese term clustering mechanism for generating semantic concepts of a news ontology. We utilize the parallel fuzzy inference mechanism to infer the *conceptual resonance strength* of a Chinese term pair. There are four input fuzzy variables, consisting of a *Part-of-Speech* (*POS*) fuzzy variable, *Term Vocabulary* (*TV*) fuzzy variable, *Term Association* (*TA*) fuzzy variable, and *Common Term Association* (*CTA*) fuzzy variable, and one output fuzzy variable, the *Conceptual Resonance Strength* (*CRS*), in the mechanism. In addition, the *CKIP* tool is used in Chinese natural language processing tasks, including POS tagging, refining tagging, and stop word filtering. The *fuzzy compatibility relation* approach to the semantic concept clustering is also proposed. Simulation results show that our approach can effectively cluster Chinese terms to generate the semantic concepts of a news ontology.

Keywords: Ontology, Chinese Natural Language Processing, Fuzzy Inference, Feature Selection, Concept Clustering

1. Introduction

An ontology is an explicit, machine-readable specification of a shared conceptualization [Studer *et al.* 1998]. It is an essential element in many applications, including agent systems, knowledge management systems, and e-commerce platforms. It can help generate natural language, integrate intelligent information, provide semantic-based access to the Internet, and extract information from texts [Gomez-Perez *et al.* 2002] [Fensel 2002] [Schreiber *et al.* 2001]. Soo *et al.* [2001] considered an ontology to be a collection of key concepts and their inter-relationships, collectively providing an abstract view of an application domain. With the

[*] Department of Information Management, Chang Jung Christian University, Tainan, Taiwan
E-Mail: leecs@mail.cju.edu.tw; leecs@cad.csie.ncku.edu.tw
[+] CREDIT Research Center, National Cheng Kung University, Tainan, Taiwan

[Received July 2, 2004; Revised January 25, 2005; Accepted January 26, 2005]

support of an ontology, a user and a system can communicate with each other through their shared and common understanding of a domain. M. MissiKoff *et al.* [2002] proposed an integrated approach to web ontology learning and engineering that can build and access a domain ontology for intelligent information integration within a virtual user community. The proposed approach involves automatic concept learning, machine-supported concept validation, and management. Embley *et al.* [1998] presented a method of extracting information from unstructured documents based on an ontology. Alani *et al.* [2003] proposed the Artequakt, which automatically extracts knowledge about artists from the Web based on a domain ontology. It can generate biographies that are tailored to a user's interests and requirements. Navigli *et al.* [2003] proposed OntoLearn with ontology learning capability to extract relevant domain terms from a corpus of text. OntoSeek [Guarino *et al.* 1999] is a system designed for content-based information retrieval. It combines an ontology-driven content-matching mechanism with moderately expressive representation formalism. Lee *et al.* [2004] proposed an ontology-based fuzzy event extraction agent for Chinese news summarization. The summarization agent can generate a sentence set for each piece of Chinese news.

In this paper, we propose a Chinese term clustering mechanism for generating the semantic concepts of a news ontology. The parallel fuzzy inference mechanism is adopted to infer the conceptual resonance strength for any two Chinese terms. The *CKIP* tool [Academia Sinica 1993] is used in Chinese natural language processing, including POS tagging, refining tagging, and stop word filtering. The remainder of this paper is structured as follows. Section 2 introduces the structure of the Chinese term clustering mechanism. Semantic concept analysis for Chinese term clustering is presented in Section 3. Section 4 introduces the parallel fuzzy inference mechanism for semantic concept generation. Section 5 presents experimental results. Finally, some conclusions are drawn in Section 6.

2. The Structure of the Chinese Term Clustering Mechanism

An ontology is defined as a set of representational terms called concepts. The inter-relationships among these concepts describe a target world. Here, we will briefly describe the structure of the object-oriented ontology [Lee *et al.* 2003]. An object-oriented ontology consists of several basic components: (1) *Domain*: The top layer of the ontology is the name of the domain knowledge. In this study, an ontology was constructed for Chinese news, so its domain name is Chinese news. (2) *Category*: The second layer contains the categories of the domain ontology. Each category is composed of some concepts with various inter-relationships. There are seven categories for our Chinese news ontology. 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). (3)

Concept Set: The *Concept Set* is composed of various concepts and relations. We treat each concept in the ontology as a class, so the structure of the *Concept Set* can be treated as a class diagram. Figure 1 shows an example for our Chinese Political news domain ontology [Lee *et al.* 2003].

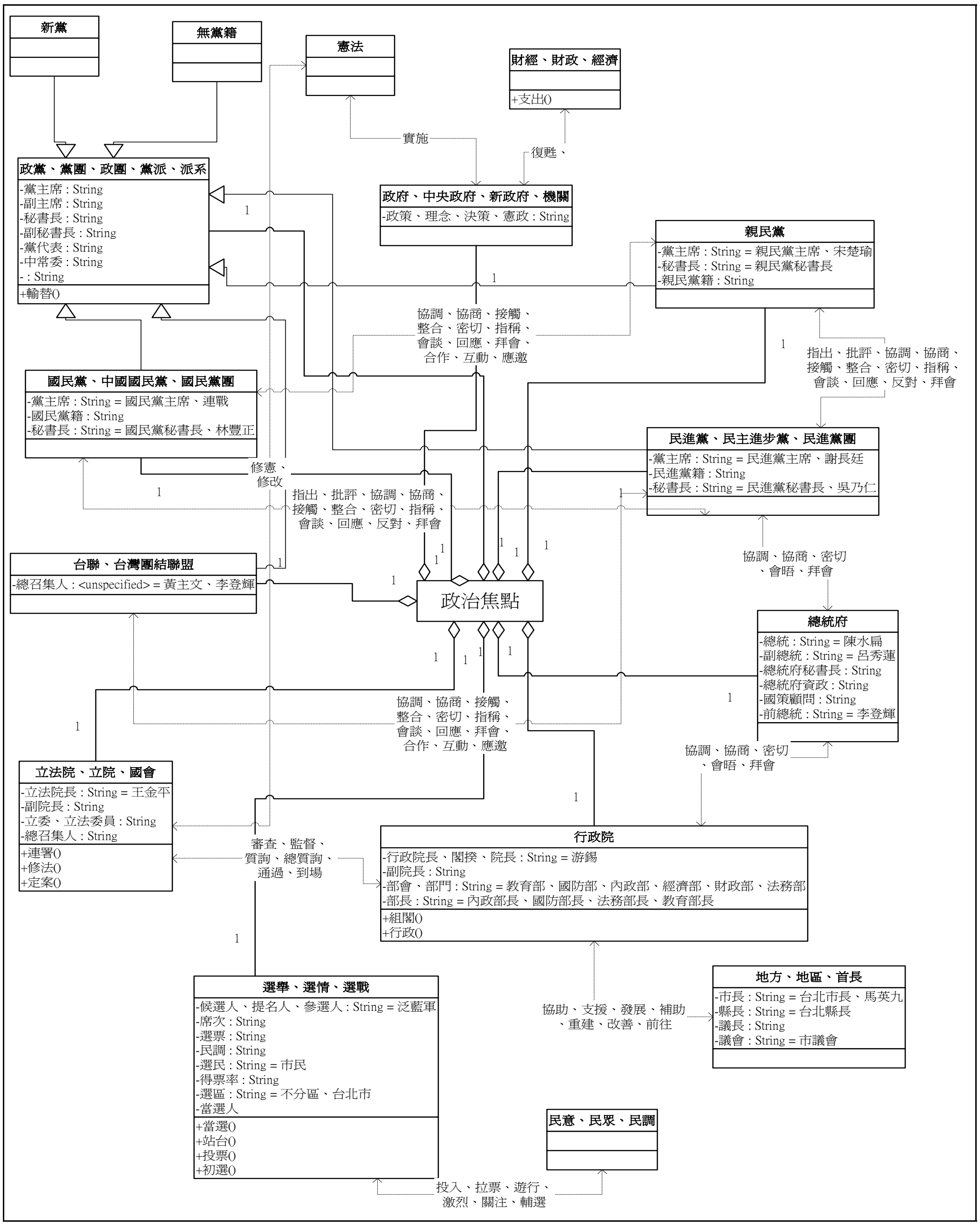

新黨
無黨籍
憲法
財經、財政、經濟
+支出()
實施
復甦、
政黨、黨團、政團、黨派、派系
-黨主席 : String
-副主席 : String
-秘書長 : String
-副秘書長 : String
-黨代表 : String
-中常委 : String
- : String
+輪替()
政府、中央政府、新政府、機關
-政策、理念、決策、憲政 : String
親民黨
-黨主席 : String = 親民黨主席、宋楚瑜
-秘書長 : String = 親民黨秘書長
-親民黨籍 : String
協調、協商、接觸、整合、密切、指稱、會談、回應、拜會、合作、互動、應邀
指出、批評、協調、協商、接觸、整合、密切、指稱、會談、回應、反對、拜會
國民黨、中國國民黨、國民黨團
-黨主席 : String = 國民黨主席、連戰
-國民黨籍 : String
-秘書長 : String = 國民黨秘書長、林豐正
民進黨、民主進步黨、民進黨團
-黨主席 : String = 民進黨主席、謝長廷
-民進黨籍 : String
-秘書長 : String = 民進黨秘書長、吳乃仁
修憲、修改
指出、批評、協調、協商、接觸、整合、密切、指稱、會談、回應、反對、拜會
協調、協商、密切、會晤、拜會
台聯、台灣團結聯盟
-總召集人 : <unspecified> = 黃主文、李登輝
政治焦點
總統府
-總統 : String = 陳水扁
-副總統 : String = 呂秀蓮
-總統府秘書長 : String
-總統府資政 : String
-國策顧問 : String
-前總統 : String = 李登輝
協調、協商、接觸、整合、密切、指稱、會談、回應、拜會、合作、互動、應邀
協調、協商、密切、會晤、拜會
立法院、立院、國會
-立法院長 : String = 王金平
-副院長 : String
-立委、立法委員 : String
-總召集人 : String
+連署()
+修法()
+定案()
審查、監督、質詢、總質詢、通過、到場
行政院
-行政院長、閣揆、院長 : String = 游錫
-副院長 : String
-部會、部門 : String = 教育部、國防部、內政部、經濟部、財政部、法務部
-部長 : String = 內政部長、國防部長、法務部長、教育部長
+組閣()
+行政()
選舉、選情、選戰
-候選人、提名人、參選人 : String = 泛藍軍
-席次 : String
-選票 : String
-民調 : String
-選民 : String = 市民
-得票率 : String
-選區 : String = 不分區、台北市
-當選人
+當選()
+站台()
+投票()
+初選()
協助、支援、發展、補助、重建、改善、前往
地方、地區、首長
-市長 : String = 台北市長、馬英九
-縣長 : String = 台北縣長
-議長 : String
-議會 : String = 市議會
民意、民眾、民調
投入、拉票、遊行、激烈、關注、輔選

Figure 1. The domain ontology for the news category "Political"

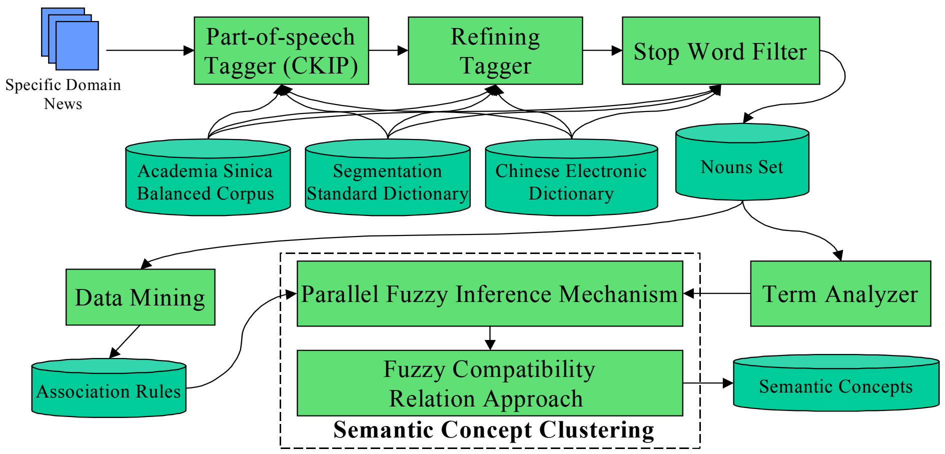

Specific Domain News
Part-of-speech Tagger (CKIP)
Refining Tagger
Stop Word Filter
Academia Sinica Balanced Corpus
Segmentation Standard Dictionary
Chinese Electronic Dictionary
Nouns Set
Data Mining
Parallel Fuzzy Inference Mechanism
Term Analyzer
Fuzzy Compatibility Relation Approach
Association Rules
Semantic Concepts
Semantic Concept Clustering

Figure 2. The structure of the Chinese term clustering mechanism

In this section, we will propose a Chinese term clustering mechanism for generating the semantic concepts of a news ontology. Figure 2 shows the structure of the Chinese term clustering mechanism. Natural language processing technologies were utilized to deal with the Chinese news that we gathered from the China Times website (http://www.chinatimes.com.tw). Several technologies, including a part-of-speech tagger, refining tagger, stop word filter, and term analyzer, were adopted for document pre-processing. Chinese language processing tools, such as *CKIP* [Academia Sinica 1993], the *Academia Sinica Balanced Corpus*, *Segmentation Standard Dictionary* [Academia Sinica 1998], and *Chinese Electronic Dictionary* [Academia Sinica 1993] provided by Academia Sinica, were used to deal with the Chinese news. In addition, the data mining technique and the concept clustering approach based on the *fuzzy compatibility relation* were employed. We will briefly describe these technologies in the following.

First, the *CKIP* is used to tag each word with its POS tag for the Chinese news. The refining tagger then refers to the *Academia Sinica Balanced Corpus* and *Chinese Electronic Dictionary* to refine the POS tags. With the aid of the corpus and the dictionary, we have sufficient Chinese POS knowledge to analyze the features of the terms for semantic concept clustering. The *stop word filter* is used to select terms with useful POS tags as candidate features. Table 1 shows unmeaning tags as stop words. Then, the *term analyzer* analyzes the term frequency of the news to select the important terms from a specific class of news. For example, the terms with the POS tags Na (普通名詞), Nb (專有名詞), Nc (地方名詞), and Nd (時間名詞) are preserved and sent to the *Parallel Fuzzy Inference Mechanism* for further processing. The *Data Mining* mechanism adopts the *Apriori Algorithm* to generate association rules, which are used in the *Parallel Fuzzy Inference Mechanism*. The *Apriori Algorithm* [Jacobes 1993] is described as follows.

Table 1. The stop word list used in the stop word filter [Academia Sinica 1993]

Part-of-Speech Tag	**Meaning**	**Examples**
Ca	並列連接詞	和、或者
Cb	關聯連接詞	雖然、不但
Da	數量副詞	一共、恰好
Dba	法相副詞	一定、也許
Dbb,Dbc	評價副詞	居然、果然
Dc	否定副詞	沒有、未
Dd	時間副詞	隨即、稍後
Df	程度副詞	非常、更
Dg	地方副詞	到處、遍地
Dh	方式副詞	如此、從中
Di	標誌副詞	過、起來
Dj	疑問副詞	爲何、何故
Dk	句副詞	據報、據了解
I	感歎詞	哦、哇
P	介詞	經過、遭受
T	語助詞	了、的

```
Apriori Algorithm:
        Find frequent itemsets using an iterative level-wise approach based on candidate generation.
Input:
        Database $D$ of transactions and minimum support threshold $min\_sup$.
Output:
        $L$, frequent itemsets in $D$.
Method:
Step 1: $L_1$=find_frequent_1-itemsets($D$,$min\_sup$);        //find_frequent_1-itemsets denotes to
                                                                   find frequent 1-itemsets in D
Step 2: For (k=2; $L_{k-1} \neq \phi$ ; k++) {
        Step 2.1: $C_k$=apriori_gen($L_{k-1}$,$min\_sup$);
        Step 2.2: For each transaction t$\in D$ {        //scan D for counts
                        $C_t$=subset($C_k$,t);    //get the subsets of t that are candidates
        Step 2.3: For each candidate c$\in$C$_t$
                    c.count++;
                  }
Step 3: $L_k$={c$\in C_k$|c.count$\geqq min\_sup$}
        }
Step 4: Return $L=\cup_k L_k$
Step 5: End.
Procedure find_frequent_1-itemsets($D$,$min\_sup$)
Step 1: Get $C_1$ from $D$            // $C_1$ denotes candidate 1-itemsets
Step 2: For each transaction $t \in D$ {        //scan D for counts
        $C_t$=subset($C_k$,t);   //get the subsets of t that are candidates
            }
Step 3: For each candidate $c \in C_1$
            c.count++;
        }
Step 4: $L_1=\{c \in C_k | c.count \geqq min\_sup\}$
Step 5: Return $L_1$
Step 6: End.
Procedure apriori_gen($L_{k-1}$; min_sup)
Step 1: For each itemset $l_1 \in L_{k-1}$
        Step 1.1: For each itemset $l_2 \in L_{k-1}$
        Step 1.2: If $(l_1[1]= l_2[1]) \wedge (l_1[2]= l_2[2]) \wedge \ldots \wedge (l_1[k-1]< l_2[k-1])$ then {
                    $c= l_1 \infty l_2$;     //join step: generate candidates
        Step 1.3: If Has_infrequent_subset(c, $L_{k-1}$) then
                      delete c;     //prune step: remove unfruitful candidate
                    Else add c to $C_k$
                    }
Step 2: Return $C_k$;
Step 3: End.
Procedure Has_infrequent_subset (c; $L_{k-1}$) //use priori knowledge
Step 1: For each (k-1)-subset s of c
        Step 1.1: If $s \notin L_{k-1}$ then return TRUE;
        Step 1.2: Else return FALSE;
Step 2: End.
```

3. Semantic Concept Analysis for Chinese Term Clustering

In this paper, we propose the C*onceptual Resonance Strength* (*CRS*) fuzzy variable for Chinese term clustering. The *CRS* is the similar degree for any term pair in the same concept. Hence, any Chinese term pair with a strong *CRS* will be classified as the same concept. We use four fuzzy variables, the *resonance strength in Part-of-Speech* (*POS*), *resonance strength in Term Vocabulary* (*TV*), *resonance strength in Term Association* (*TA*), *and resonance strength in Common Term Association* (*CTA*), to compute the *CRS* of the Chinese term pair. We will describe these variables in the following.

A. Resonance Strength in Part-of-Speech (POS)

The first fuzzy variable for *CRS* is the *resonance strength in Part-of-Speech* (*POS*). Figure 3 shows the structure of the tagging tree that is used to compute the *resonance strength of POS* for any Chinese term pair. Table 2 shows the refining POS tags of Chinese noun terms.

Table 2. The refining POS tags of Chinese noun terms

Part-of-Speech Tag		**Meaning**	**Examples**
粗詞類	細詞類		
Na		普通名詞	
	Naa	物質名詞	泥土、水
	Nab	個體名詞	桌子、杯子
	Nac	可數抽象名詞	夢、符號
	Nad	抽象名詞	風度、香氣
	Nae	集合名詞	車輛、船隻
Nb		專有名詞	
	Nba	正式專有名詞	雙魚座、余光中
	Nbc	姓氏	張、王
Nc		地方名詞	
	Nca	專有地方名詞	西班牙、台北
	Ncb	普通地方名詞	郵局、市場
	Ncc	名方式地方名詞	海外、身上
	Ncd	表事物相對位置的地方詞	上頭、中間
	Nce	定名式地方名詞	四海、當地
Nd		時間名詞	
	Nda	時間名詞(歷史性、循環重複)	唐朝、春、夏、秋、冬
	Ndc	名方式時間名詞	年底、週末
	Ndd	副詞性時間名詞	現在、當今
Ne		定詞	這、哪、少許
Nf		量詞	
	Nfa	個體量詞	一"張"桌子、一"個"杯子
	Nfb	跟述賓式合用的量詞	寫一"手"好字、下一"盤"棋
	Nfc	群體量詞	一"雙"筷子、一"副"耳環
	Nfd	部分量詞	一"節"甘蔗、一"段"文章
	Nfe	容器量詞	一"箱"書、一"碗"飯
	Nff	暫時量詞	一"頭"秀髮、一"地"落葉
	Nfg	標準量詞	公斤、法郎
	Nfh	準量詞	國、面
	Nfi	述詞用量詞	看一"遍"、摸一"下"
	Nfzz	零量詞	"三萬"人口
Ng		方位詞	接"上"、屋"後"、睡覺"之前"
Nh		代名詞	
	Nha	人稱代名詞	你、我、他、自己
	Nhb	疑問代名詞	誰、什麼
	Nhc	泛指代名詞	之、其

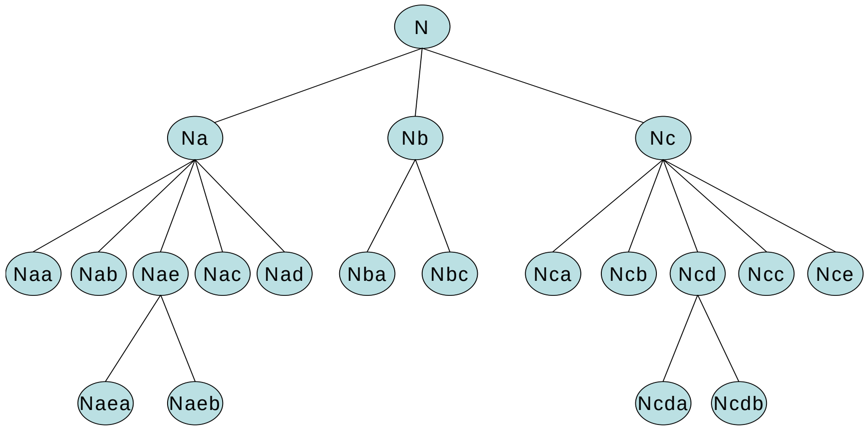


Figure 3. The structure of the tagging tree derived using CKIP

The resonance will be strong when the path distance of any Chinese term pair is short. For example, the two terms "電腦(computer)" and "軟體(software)" with their *POS* are "電腦(computer) (Nab)" and "軟體(software) (Nac)," respectively. Hence, the path distance of the term pair ("電腦(computer)", "軟體(software)") is *2* (Nab -> Na -> Nac).

B. Resonance Strength in Term Vocabulary (TV)

From the viewpoint of Chinese language characteristics, any term pair with common words will be similar in semantic meaning. For example, the Chinese terms in the term set {民進黨, 民進黨團, 民主進步黨} are similar in semantic meaning since they are composed of the common words "民", "進", and "黨". We also consider another characteristic of Chinese terms with respect to term vocabulary. This assumes that terms having the same starting or ending word will share some common linguistic properties [Yang *et al.* 1994][Gao *et al.* 2001]. Good examples of starting and ending words are as follows: {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)}. The first term set has the same starting word "星," and the second term set has the same ending word "天." The algorithm for computing resonance strength in *TV* [Lee *et al.* 2003] is shown below.

```
Algorithm for computing the resonance strength in TV
Input:    All terms $(t_1, t_2 ..., t_n)$ selected from Term Analyzer
Output:   Resonance strength in TV between any two Chinese terms
Method:
Step1: For all terms $(t_1, t_2 ..., t_n)$
    Step1.1: For all terms $(t_1, t_2 ..., t_n)$
        Step1.1.1: Generate a term pair $(t_a, t_b)$ $1 \le a < b \le n$
        Step1.1.2: $TV(t_a, t_b) = N$    /* N represents identical words between the Chinese
                                          term pair $(t_a, t_b)$ */
        Step1.1.3: If the starting word of two Chinese terms is the same
                   then $TV(t_a, t_b) = TV(t_a, t_b) + 0.5$.
        Step1.1.4: If the ending word of two Chinese terms is the same
                   then $TV(t_a, t_b) = TV(t_a, t_b) + 0.5$.
Step2: $Max_{TV}$ = maximum $TV(t_i, t_j)$ value of all term pairs.
Step3: $Min_{TV}$ = minimum $TV(t_i, t_j)$ value of all term pairs.
Step4: End.
```

For example, the two terms “民進黨團” and “民主進步黨” have three common words, “民,” “進,” and “黨,” and the same starting word, “民,” so the total strength is *3.5*.

C. Resonance Strength in Term Association (TA)

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 [Gao *et al.* 2001]. However, the obtained term knowledge is not sufficient for concept clustering, because a term pair is sometimes similar in meaning but lacks common properties of knowledge. Therefore, the confidence value derived using the *Apriori Algorithm* for the term pair can be applied to decide the strength of term relation. A term pair with a high confidence value consists of two terms that have a strong relationship and can be classified as the same concept. For example, the term set {總統(President) (Nab), 總統府(The Office of the President) (Nca), 陳水扁(President Chen) (Nb)} represents similar concepts, so they will be clustered into the same concept. But from the viewpoint of term knowledge, only the two terms “總統(President)” and “總統府(The Office of the President)” will be clustered into the same concept. The term “陳水扁(President Chen)” will not be clustered into the concept {總統(President), 總統府(The Office of the President)}. Therefore, the *resonance strength in TA* is necessary for concept clustering. In addition, the resonance strength is decided by the confidence value of the two terms, so we adopt the average of the two confidence values as the resonance strength in *TA*. For example, the term pair {總統(Nab), 陳水扁(Nb)} for the “Political(政治焦點)” category (http://www.chinatimes.com.tw) with (總統 -> 陳水扁) has a confidence value of *0.84*, and the confidence value of (陳水扁 -> 總統) is *0.80,* so the resonance strength in *TA* is *0.82* (*(0.84+0.80)/2*).

D. Resonance Strength in Common Term Association (CTA)

Any two Chinese terms with the same common words or starting/ending words 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 terms “美國(U.S.A.)” and “美方(U.S.A.)” have the common starting word “美”; meanwhile, the Chinese terms “美方(U.S.A.)” and “警方(police)” have the common ending word “方”. But the common 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)}.*

Hence, the common term set for {美國(U.S.A.), 美方(U.S.A.)} is {白宮(White House)*, 布希(Bush)}*, and for {警方(police), 美方(U.S.A.)} is *Null*. Therefore, the term pair {美國(U.S.A.), 美方(U.S.A.)} has stronger resonance in *CTA* than the term pair {警方(police), 美方(U.S.A.)} does.

4. The Parallel Fuzzy Inference Mechanism for Semantic Concept Generating

We adopt the parallel fuzzy inference mechanism for semantic concept clustering. The fuzzy variables for computing the *CRS* of any Chinese term pair are adopted in the mechanism.

4.1 Aggregate Term Resonance with the Parallel Fuzzy Inference Mechanism

In this subsection, we will explain how four input fuzzy variables can be aggregated into one output fuzzy variable to compute the *CRS* of each Chinese term pair. There are four input fuzzy variables, consisting of *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)*, in the Parallel Fuzzy Inference Mechanism. Two linguistic terms, *POS_Low* and *POS_High*, are defined in the *POS* fuzzy variables. Figure 4 shows the membership functions of the fuzzy sets *{POS_Low, POS_High}* for the fuzzy variable *POS similarity*, where $p = \frac{\max_{POS} - \min_{POS}}{100}$.

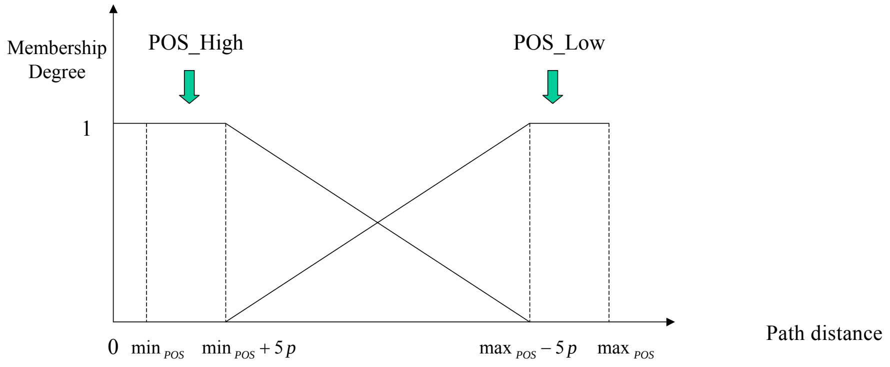


Figure 4. The membership functions of the POS fuzzy variable

Two linguistic terms, *TV_Low* and *TV_High*, are defined in the *TV* fuzzy variable. Figure 5 shows the membership functions of the fuzzy sets *{TV_Low, TV_High}* for the fuzzy variable *TV similarity*, where $p = \frac{\max_{TV} - \min_{TV}}{100}$.

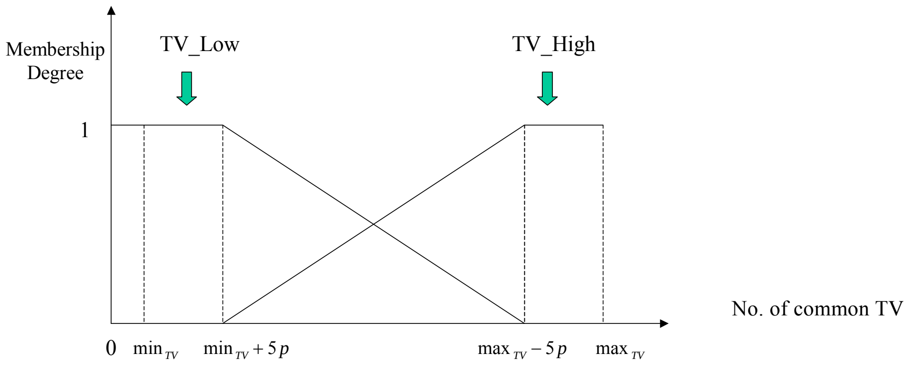


Figure 5. The membership functions of the TV fuzzy variable

Three linguistic terms, consisting of *TA_Low*, *TA_Medium*, and *TA_High*, are defined in the *TA* fuzzy variable. The membership functions of the fuzzy sets *{TA_Low, TA_Medium, TA_High}* for the fuzzy variable *TA strength* are shown in Figure 6, where $p = \frac{\max_{TA} - \min_{TA}}{100}$.

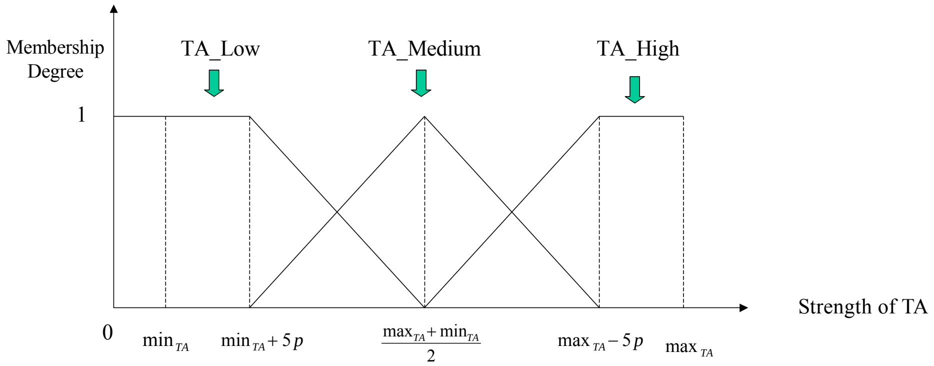


Figure 6. The membership functions of the TA fuzzy variable

Three linguistic terms, consisting of *CTA_Low*, *CTA_Medium*, and *CTA_High*, are defined in the *CTA* fuzzy variable. Figure 7 shows the membership functions of the fuzzy sets *{CTA_Low, CTA_Medium, CTA_High}* for the fuzzy variable *CTA strength*, where $p = \frac{\max_{CTA} - \min_{CTA}}{100}$.

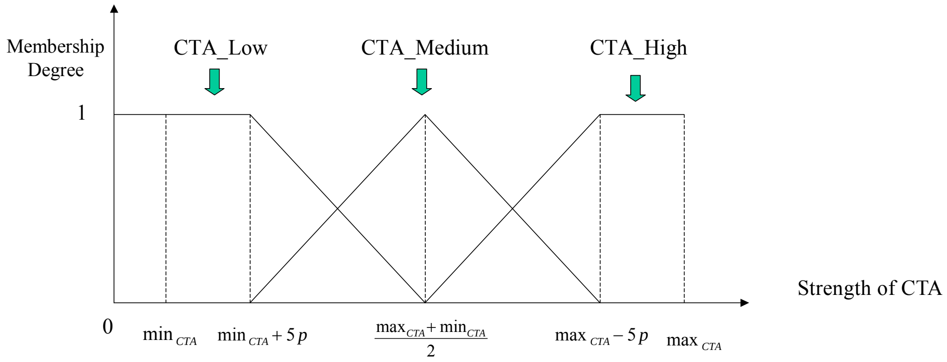


Figure 7. The membership functions of the CTA fuzzy variable

Five linguistic terms, consisting of *CRS_Very_Low*, *CRS_Low*, *CRS_Medium*, *CRS_High*, and *CRS_Very_High*, are defined in the *CRS* fuzzy variable. Figure 8 shows the membership functions of the fuzzy sets *{CRS_Very_Low, CRS_Low, CRS_Medium, CRS_High, CRS_Very_High}* for the fuzzy variable *CRS strength*, where $p = \frac{\max_{CRS} - \min_{CRS}}{100}$.

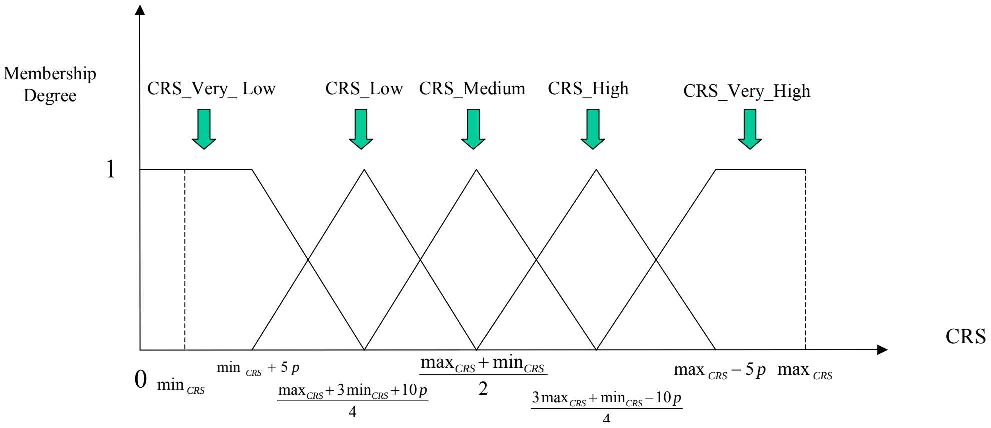
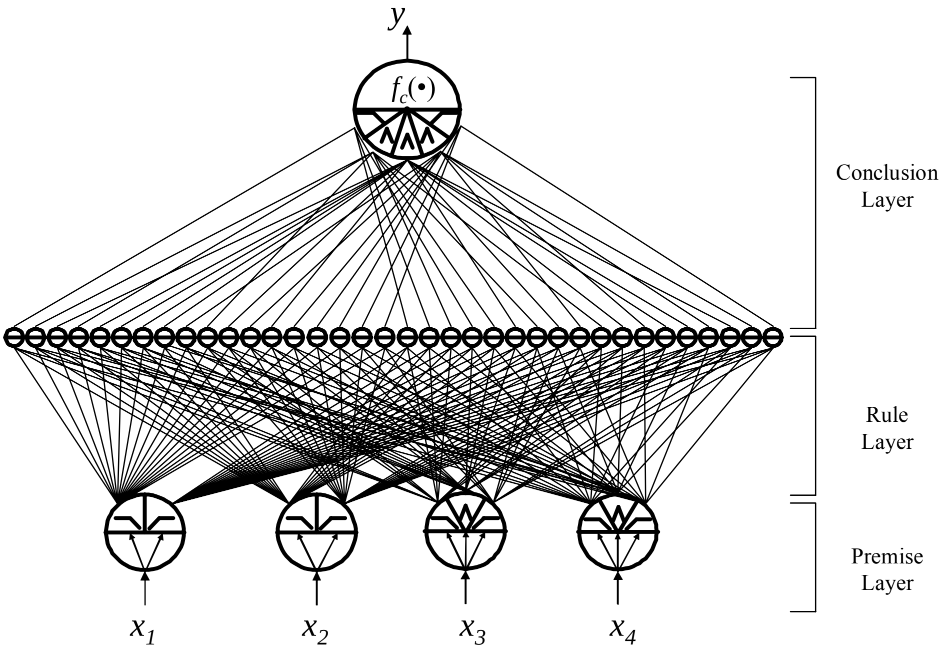


Figure 8. The membership functions of the CRS fuzzy variable

Having described the fuzzy variables used to compute the *CRS* of a Chinese term pair, we will next explain how the parallel fuzzy inference mechanism proposed by Kuo *et al.* [1998] and Lin [1991] is used to perform semantic concept clustering.



Figue 9. The structure of the parallel fuzzy inference mechanism for semantic concept clustering

Figure 9 shows the structure of the parallel fuzzy inference mechanism. It is a three-layered network which can be constructed by directly mapping from a set of specific fuzzy rules, or can be learned incrementally from a set of training patterns. In our approach, the rules are defined by the domain expert. The structure consists of a *premise layer*, *rule layer*, and *conclusion layer*. There are two kinds of nodes, *fuzzy linguistic nodes* and *rule nodes*, in this model. A *fuzzy linguistic node* represents a fuzzy variable and manipulates the information related to that linguistic variable. A *rule node* represents a rule and determines the final firing strength of that rule during the inferring process. The *premise layer* performs the first inference step to compute matching degrees. The *conclusion layer* is responsible for drawing conclusions and defuzzification. We will describe each layer in the following.

A. Premise layer:

As shown in Figure 9, the first layer is called the *premise layer* and is used to represent the premise part of the fuzzy system. 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 the fuzzy rules. In other words, a linguistic node is a polymorphic object that can be viewed differently by different fuzzy rules. Figure 10 shows the fuzzy linguistic node for the *TA* fuzzy variable.

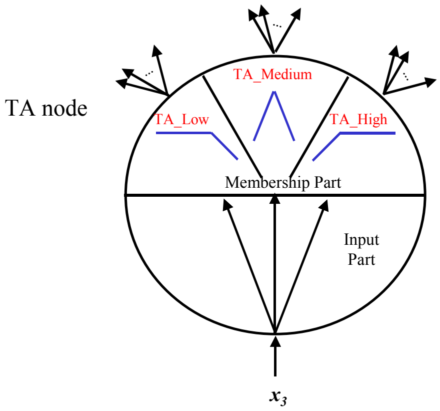


Figure 10. The structure of the fuzzy linguistic node for the TA fuzzy variable

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 the *i*th linguistic node. Thus, the output vector of the premise layer is $\mu^1 = ((u^1_{11}, u^1_{21}, ..., u^1_{N_1 1}), (u^1_{12}, u^1_{22}, ..., u^1_{N_2 2}), ..., (u^1_{1n}, u^1_{2n}, ..., u^1_{N_n n}))$, where u^1_{ij} is the matching degree of the *j*-th linguistic term in the *i*-th condition node. In our approach, the triangular function and trapezoidal function are adopted as the membership functions for the linguistic terms. Equation 1 and 2 show the triangular and trapezoidal membership functions, respectively:

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

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

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

where n is the number of linguistic terms for the i-th linguistic node. Therefore, $\mu_{ij}^{1}=f_{ij}^{1}(x)$.

B. Rule layer:

The second layer is called the *rule layer*. In it, each node is a rule node and is used to represent a fuzzy rule. The links in this layer are used to perform precondition matching of fuzzy logic rules. The output of a rule node in the rule layer is linked to associated linguistic nodes in the third layer. In our model, the rules are previously defined by domain experts. Table 3 shows the fuzzy inference rules for the parallel fuzzy inference mechanism. Figure 11 shows the structure of the rule node.

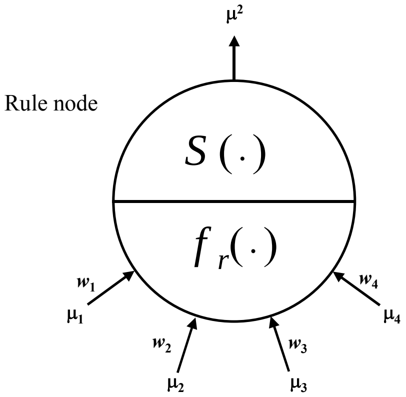


Figure 11. The structure of the rule node

Table 3. The fuzzy inference rule for the parallel fuzzy inference mechanism

Rule	***POS***	***TV***	***TA***	***CTA***	***CRS***		***Rule***	***POS***	***TV***	***TA***	***CTA***	***CRS***
1	L	L	L	L	VL		**19**	H	L	L	L	VL
2	L	L	L	M	VL		**20**	H	L	L	M	VL
3	L	L	L	H	L		**21**	H	L	L	H	L
4	L	L	M	L	L		**22**	H	L	M	L	L
5	L	L	M	M	L		**23**	H	L	M	M	M
6	L	L	M	H	M		**24**	H	L	M	H	M
7	L	L	H	L	M		**25**	H	L	H	L	H
8	L	L	H	M	H		**26**	H	L	H	M	H
9	L	L	H	H	H		**27**	H	L	H	H	H
10	L	H	L	L	M		**28**	H	H	L	L	M
11	L	H	L	M	M		**29**	H	H	L	M	H
12	L	H	L	H	H		**30**	H	H	L	H	H
13	L	H	M	L	H		**31**	H	H	M	L	H
14	L	H	M	M	H		**32**	H	H	M	M	H
15	L	H	M	H	H		**33**	H	H	M	H	VH
16	L	H	H	L	H		**34**	H	H	H	L	VH
17	L	H	H	M	VH		**35**	H	H	H	M	VH
18	L	H	H	H	VH		**36**	H	H	H	H	VH

The f_r function in Figure 11 provides the net input for this node and is defined as $f_r = \sum_{i=1}^{p} w_i \mu_i$.

The S function is used to normalize the f_r function and is defined in Eq. 4:

$$S(x:a,b) = \begin{cases} 0 \ ,x < a \\ 2(\frac{x-a}{b-a})^2 \ , \ a \le x \le \frac{a+b}{2} \\ 1 - 2(\frac{x-b}{b-a})^2 \ , \frac{a+b}{2} \le x < b \\ 1 \ ,x \ge b \end{cases} . \qquad (4)$$

In our case, the rule node has four inputs, and each input value is between *0* and *1*.

C. Conclusion layer:

The third layer is called the *conclusion layer*. This layer is also composed of a set of fuzzy linguistic nodes. A fuzzy linguistic node can also operate in a reverse mode, called a conclusion node. In the reverse mode, fuzzy linguistic nodes are responsible for drawing conclusions and defuzzification. Figure 12 shows the structure of a linguistic node in the reverse mode, and shows that it is also an output node.

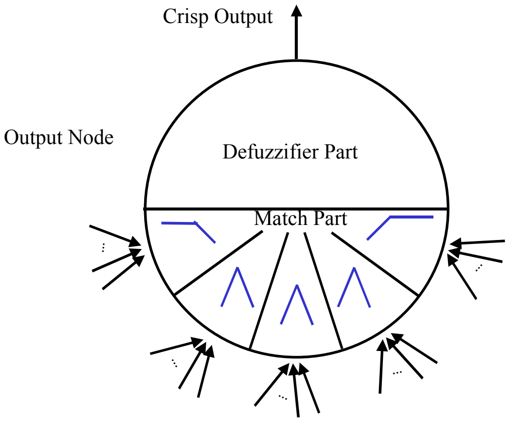


Figure 12. The structure of a fuzzy linguistic node for the conclusion layer

In our model, the final output *y* is the crisp value that is produced by combining all the inference results with their firing strengths. The defuzzification process is defined in Eq. 5:

$$CrispOutput = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}^{k} w_{ij}^{k} V_{ij}}{\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}^{k} w_{ij}^{k}}, \qquad (5)$$

where $w^k = \frac{\sum_{i=1}^{n} \mu_i^1}{n}$, V_{ij} is the center of gravity, *r* is the number of corresponding rule nodes, *c* is the number of linguistic terms of the output node, *n* is the number of fuzzy variables in the premise layer, and *k* represents the *k*-th layer. The values of *r*, *c*, *n*, and *k* adopted here are *36*, *5*, *4* and *2*, respectively.

4.2 Fuzzy Compatibility Relation Approach to Semantic Concept Clustering

The conceptual resonance of terms pair can be treated as a fuzzy compatibility relation, because it satisfies the properties of reflexivity and symmetricalness. Therefore, the problem of concept clustering is that of finding all the classes of maximal α-compatibles with fuzzy compatibility relations. In this model, the value of α represents a specified membership degree of the fuzzy compatibility relation. The semantic concept clustering algorithm based on the fuzzy compatibility relation approach is described as follows.

Semantic Concept Clustering Algorithm based on the 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 its corresponding fuzzy conceptual resonance matrix $A = [\alpha_{ij}]_{n \times n}$.

Output:

The *Final_Concept_Set*, which 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 $Term[i]$, and all the
              compatibility membership degrees $\alpha_{ij}$'s of the terms in $Set_i$ are not less
              than $\alpha$ */
    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 concept
              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.
```

The method used to determine α is very important for semantic concept clustering, because it will influence the number of concepts and the degree of compatibility of Chinese terms. The value of α may vary for different domain documents, because their properties may be different. Now, we use an example to explain the relation of concept clustering for different value of α. In Figure 13, the terms are clustered based on a specific value of α, and they point to the same concept if their *CRS* values are greater than α.

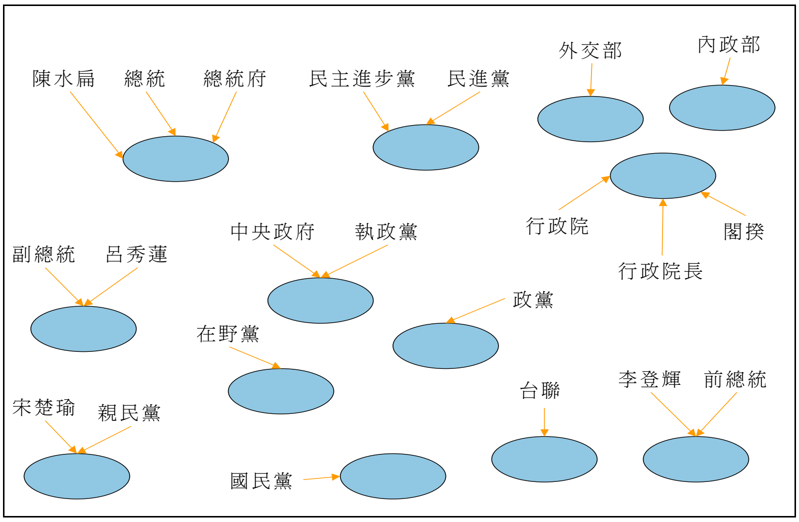


Figure 13. The concepts clustered based on a specific value of α

If we reduce the value of α, then the terms will be clustered with high compatibility. Figure 14 shows the concepts clustered based on a lower value of α corresponding to Figure 13.

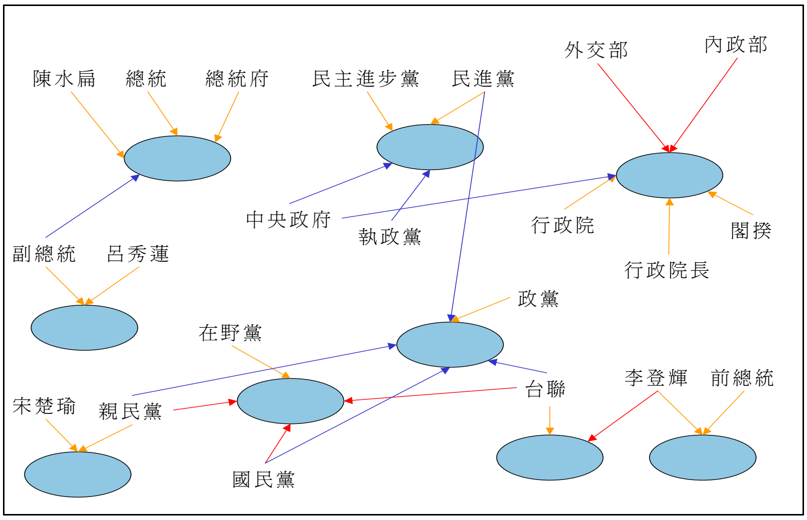


Figure 14. The concepts clustered based on a lower value of α

The lower α value will result in the formation of the more concepts and strengthen the compatibility degree of terms for a specific concept. The α decision algorithm for semantic concept clustering can be described as shown below. The prune-and-search strategy will be applied to solve this problem.

```
The α Decision Algorithm for semantic concept clustering
Input:
    CRS of terms for a specific news category
    The interval $[a,b]$ for the scope of the concept number
Output:
    Fuzzy Membership Degree α
Method:
Step 1: Read all the values of conceptual resonance into a double array $R = double[n]$,
        where $n$ is the number of values.
Step 2: Sort all the elements of $R$, where $R[i] \le R[j]$ for $0 \le i < j \le n$.
Step 3: $p \leftarrow \frac{n}{2}$
Step 4: $count \leftarrow 1$
Step 5: $\alpha \leftarrow R[p]$, and let the number of classes of maximalα-compatibles be $c$.
    Step 5.1: $count \leftarrow count + 1$
    Step 5.2: If $c > b$ Then
        Step 5.2.1: $p \leftarrow p + \frac{n}{2^{count}}$
        Step 5.2.2: Go to Step 5.
    Step 5.3: If $c < a$ Then
        Step 5.3.1: $p \leftarrow p - \frac{n}{2^{count}}$
        Step 5.3.2: Go to Step 5.
    Step 5.4: If $a \le c \le b$
            Then go to Step 6.
Step 6: End.
```

5. Experimental Results

In this section, some experiments obtained using the proposed approach will be presented. The news corpus was gathered between May 2001 and March 2002 from the ChinaTimes website (http://www.chinatimes.com.tw). Seven categories of news, consisting of “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), were used in the experiments. Table 4 lists the number of documents for each news category, the Chinese terms produced by the refining tagger, the remaining terms produced by the stop word filter, and the filtering percentages for the Chinese terms and remaining terms.

Table 4. The experimental results obtained using the proposed filter

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

Next, we will analyze the results of *CRS* for any Chinese term pair. Table 5 shows the partial results of *CRS* for the "Political" (政治焦點) category with the highest values. Notice that each term pair not only exhibits strong similarity in term knowledge for the *POS* fuzzy variable and *TV* fuzzy variable but also high strength for the *TA* fuzzy variable and *CTA* fuzzy variable. The term pairs marked with asterisks (*) exhibited strong *TA* and *CTA* but weak *POS* and *TV*.

Table 5. Partial conceptual resonance results for the "Political" category with the highest values

Chinese Term Pair	*CRS*
(民主進步黨,民進黨)	0.595481543
(親民黨主席,親民黨)	0.594101448
(國民黨主席,國民黨)	0.587348993
(民進黨主席,民進黨)	0.581987023
(親民黨主席,國民黨主席)	0.577182427
(行政院長,行政院)	0.571720293
(民進黨主席,民主進步黨)	0.571574542
(立院,立法院)	0.56774317
(立法院,立法院長)	0.558938037
(民進黨團,民進黨)	0.558824035
(台北市,台北市長)	0.557260479
(民進黨籍,民進黨)	0.555527542
(民進黨主席,國民黨主席)	0.554741061
(高雄市,高雄)	0.553642597
(台聯,台灣團結聯盟)	0.553602148
(國民黨,民進黨)	0.551060166
(國民黨,國民黨籍)	0.547553789
(親民黨主席,民進黨主席)	0.54062093
(民進黨籍,國民黨籍)	0.53928488
(民進黨主席,黨主席)	0.538797148

(a)

	Chinese Term Pair	*CRS*
*	(國民黨主席,連戰)	0.534303235
	(國民黨,親民黨)	0.534003082
	(民進黨籍,民進黨團)	0.5338768
	(副秘書長,秘書長)	0.531714804
*	(親民黨主席,宋楚瑜)	0.530113977
	(立委,立法委員)	0.52974897
	(國防部,國防)	0.529134478
	(馬英九,台北市長)	0.522650369
	(親民黨,民進黨)	0.522529795
	(委員會,委員)	0.522146828
	(縣市長,縣市)	0.520918138
	(政府,中央政府)	0.518259497
*	(呂秀蓮,副總統)	0.51637767
	(民主進步黨,民進黨團)	0.516346569
*	(李登輝,前總統)	0.515308468
*	(陳水扁,總統)	0.513920884
	(副院長,立法院長)	0.512207578
	(民主進步黨,國民黨)	0.511493131
	(總統府,總統)	0.507264737
*	(王金平,立法院長)	0.504385767

(b)

The next experiment was conducted to obtain semantic concept clustering results under various α values. In this experiment, the number of concepts varied between 500 and 1,000 for each news category. Table 6 shows that the different values of α produced various numbers of concepts containing different terms. The experimental results show that the semantic concept clustering results were influenced by the values of α.

Table 6. Analysis of various α values for each news category

News Category	政治焦點 (Political)	國際要聞 (International)	股市財經 (Finance)	兩岸風雲 (Cross-Strait)	社會地方 (Societal)	運動娛樂 (Entertainment)	生活新知 (Life)
α	0.40	0.40	0.42	0.41	0.40	0.39	0.40
Number of Concepts	971	948	543	791	783	640	880
Number of Average Terms per Concept	3.45	3.56	3.13	3.39	3.08	3.65	3.64

Table 7 shows the concept clustering results under various values of α for the “Life” (生活新知) category. Table 8 shows a partial listing of the concepts, including concept names, attributes and operations, in the golden standard ontology of the “Life” (生活新知) category.

Table 7. Analysis of various α values for the “Life” (生活新知) category

α =0.44	Concept No. 1	教育 教師 教授 教育部長
	Concept No. 2	學校 學生 學術 大學 台灣大學
	Concept No. 3	學者 學術 大學 教授
α =0.42	Concept No. 1	教育 教師 教授 教育部長 學生 學校 教育部
	Concept No. 2	學校 學生 學術 大學 台灣大學 校長
	Concept No. 3	學者 學術 大學 教授 研究
	Concept No. 4	學校 家長 老師 學生
	Concept No. 5	科學 學術 大學
	Concept No. 6	學者 科學家 專家
α =0.40	Concept No. 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 No. 2	學校 學生 學術 大學 台灣大學 校長 院長 教授
	Concept No. 3	學者 學術 大學 教授 研究 成果 領域
	Concept No. 4	學校 家長 老師 學生 教育部長 教育部 高中 大學
	Concept No. 5	科學 學術 大學 研究所 研究
	Concept No. 6	學者 專家 科學家 科學
	Concept No. 7	學者 科學 學生 生物 領域 學術 大學
	Concept No. 8	成果 研究 科學 領域 學術
	Concept No. 9	技術 研究 領域 應用 產業

Table 8. An example of the gold-standard concepts for the "Life" (生活新知) category

Concept name	Attribute	Operation
教育、教育部	教育部長：String 教師、教授、教師：String 學生：String	Null
學校、校園、校方	小學：String 中學：String=國中、高中 大學、學院：String=台灣大學 研究所：String 課程：String=考試、暑假、寒假	Null
學術、研究	學者：String=教授、科學家、專家 領域：String=生物、醫學、電腦科學 研討會：String	發展、研發、研究、實驗
科技、科學	領域：String=電機、電子、資訊、通訊、半導體、光電、網路 產品：String=手機、硬體、軟體、電腦、處理器、液晶螢幕	Null
醫院、醫界	醫師、醫生：String	手術、檢驗
衛生局、衛生署	健保：String=全民健保、健保卡	Null
氣象、天氣、氣候	雨量、雨勢、雨：String=大雨、豪雨、雷雨、陣雨、雷陣雨 鋒面：String 氣溫：String 季風：String=東北季風	變化

Notice that the concepts with higher values of α are subsets of the concepts with lower values of α. That is, a lower value of α generated a concept with more Chinese terms. In the final experiment, we tested the performance measures Precision and Recall. We choose four students who were currently working toward the M.S. degree in Computer Science and Information Management, and let them to evaluate the values obtained using Eq.(6) and Eq.(7) for precision and recall. Figure 15-20 show the average precision and recall curves based on the evaluations performed by the four experts. Table 8 shows an example of the gold-standard concepts for the "Life" (生活新知) category. The Precision and Recall measure formulas used in this study are as follows:

$$Precision = \frac{\begin{array}{c}\text{The number of relevant common terms in gold-standard concept}\\ \text{and the automatically generated semantic concept}\end{array}}{\text{The number of terms in the automatically generated semantic concept}}, \quad (6)$$

$$Recall = \frac{\begin{array}{c}\text{The number of relevant common terms in gold - standard concept}\\ \text{and the automatically generated semantic concept}\end{array}}{\text{The number of terms in the gold - standard concept}}. \quad (7)$$

Figure 15-17 show the average precision results obtained based on the evaluations performed by the four domain experts for various α values. Figure 18-20 show the average recall results obtained based on the evaluations performed by the four domain experts for various α values.

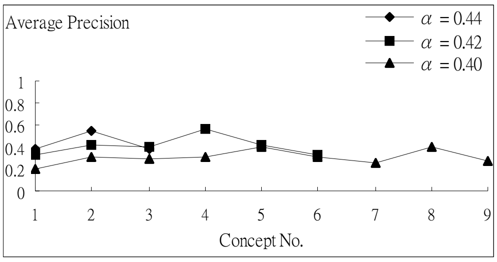


Figure 15. The average precision results for differentα values (Concept name)

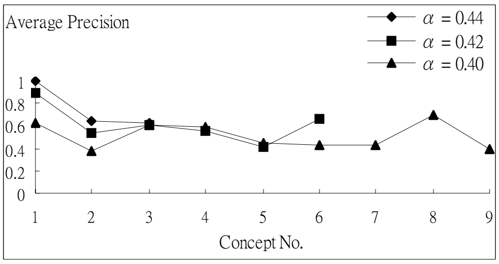


Figure 16. The average precision results for different α values (Concept name + Attribute)

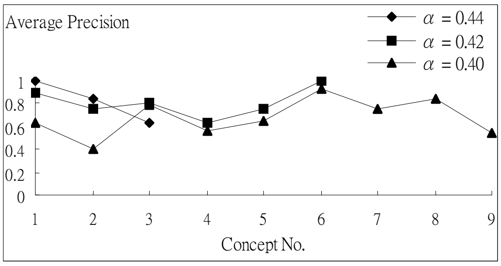


Figure 17. The average precision results for different α values (Concept name + Attribute + Attribute value)

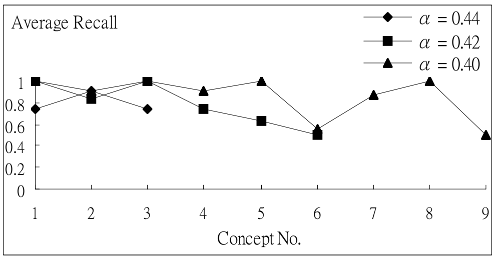


Figure 18. The average recall results for different α values (Concept name)

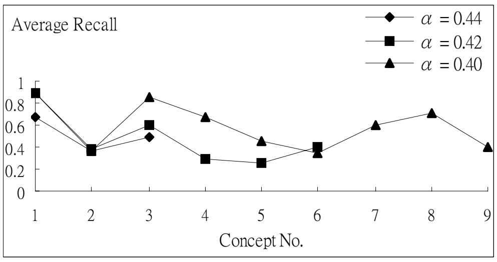


Figure 19. The average recall results for different α values (Concept name + Attribute)

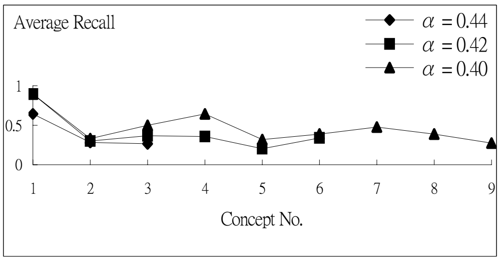


Figure 20. The average recall results for different α values (Concept name + Attribute + Attribute value)

6. Conclusions

This paper has presented a Chinese term clustering mechanism for generating semantic concepts of a news ontology. We utilize the parallel fuzzy inference mechanism to infer the conceptual resonance strength of any two Chinese terms. In addition, the *CKIP* tool is used in Chinese natural language processing, including part-of-speech tagging, Chinese term analysis, and Chinese term feature selection. A fuzzy compatibility relation approach to semantic concept clustering has also been proposed. Simulation results show that our approach can effectively cluster Chinese terms to generate the semantic concepts of a news ontology. In the future, we will extend use our approach to help construct a domain ontology more efficiently. Moreover, we will adopt the genetic learning mechanism to learn the membership functions of fuzzy inference rules for the parallel fuzzy inference mechanism. Finally, mixed Chinese/English documents will also be employed to construct more a complex domain ontology.

Acknowledge

The authors would like to express their gratitude to the anonymous reviewers for their comments, which improved the quality of this paper. This work was partially supported by the Ministry of Economic Affairs in Taiwan under grant 93-EC-17-A-02-S1-029 and partially sponsored by the National Science Council of Taiwan (R. O. C.) under grant NSC-93-2213-E-309-003.

Reference

Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. R. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents," *IEEE Intelligent Systems*, 18 (1) 2003, pp. 14-21.

CKIP, "Academia Sinica Balanced Corpus," Technical Report, No. 95-02/98-04, Academia Sinica, Taiwan, 1998.

CKIP, "Chinese Electronic Dictionary," Technical Report, No. 93-05, Academia Sinica, Taiwan, 1993.

Embley, D. W., D. M. Campbell, R. D. Smith and S. W. Liddles, "Ontology-based extraction and structuring of information from data-rich unstructured documents," *Proceeding Of ACM Conference on Information and Knowledge Management*, USA, 1998, pp. 52-59.

Fensel, D., "Ontology-based Knowledge Management," *IEEE Computer*, 35 (11) 2002, pp. 56-59.

Gao, J., J. T. Goodman and J. Miao, "The Use of Clustering Techniques for Language Modeling – Application to Asian Language," *Computational Linguistics and Chinese Language Processing*, 6 (1) 2001, pp. 27-60.

Gomez-Perez, A and O. Corcho, “Ontology languages for the semantic web,” *IEEE Intelligent Systems*, 17 (1), 2002, pp. 54-60.

Guarino, N., C. Masolo and G. Vetere, “OntoSeek: Content-based access to the web,” *IEEE Intelligent Systems*, 14 (3) 1999, pp. 70-80.

Jacobes, P. S., “Using Statistical Methods to Improve Knowledge-Based News Categorization,” *IEEE Expert*, 8 (2) 1993, pp. 13-23.

Kuo, Y. H., 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*, 28 (2) 1998, pp. 160-172.

Lammari, N. and E. Metais, “Building and maintaining ontologies: a set of algorithm,” *Data & Knowledge Engineering*, 48 (2) 2004, pp. 155-176.

Lee, C. S., Y. J. Chen and Z. W. Jian, “Ontology-based Fuzzy Event Extraction Agent for Chinese e-news Summarization,” *Expert Systems with Applications*, 25 (3) 2003, pp. 431-447

Lee, C. S., S. M. Guo and Z. W. Jian, “Weighted Fuzzy Ontology for Chinese e-News Summarization,” *2004 IEEE International Conference on Fuzzy Systems*, USA, 2004.

Lee, R. C. T., R. C. Chang, S. S. Tseng and Y. T. Tsai, “Introduction to the Design and Analysis of Algorithms,” Unalis co., Taipei, 1999.

Lin, C. T. and C. S. G. Lee, “Neural-Network-Based Fuzzy Logic Control and Decision System,” *IEEE Trans. Computer*s, 40 (12) 1991, pp. 1320-1336.

Missikoff, M., R. Navigli and P. Velardi, “Integrated approach to web ontology learning and engineering,” *IEEE Computer*, 35 (11) 2002, pp. 60-63.

Navigli, R. and P. Velardi, “Ontology learning and its application to automated terminology translation,” *IEEE Intelligent Systems*, 18 (1) 2003, pp. 22-31.

Schreiber, A.T., B. Dubbeldam, J. Wielemaker and B. Wielinga, “Ontology-based photo annotation,” *IEEE Intelligent Systems*, 16 (3) 2001, pp. 66-74.

Soo, V. W. and C. Y. Lin, “Ontology-based information retrieval in a multi-agent system for digital library,” *Proceeding Of the sixth conference on artificial intelligence and applications*, Taiwan, 2001, pp. 241-246.

Studer, R., R. Benjamins and D. Fensel, “Knowledge engineering: principles and methods,” *Data and Knowledge Engineering*, 25 (1) 1998, pp. 161-197.

Van Der Vet, P.E. and N.J.I. Mars, “Bottom-Up Construction Ontologies,” *IEEE Trans. on Knowledge and data Engineering*, 10 (4) 1998, pp. 513-526.

Yang, Y. J., S. C. Lin, L. F. Chien, K. J. Chen and L. S. Lee, “An intelligent and efficient word-class-based Chinese language model for Mandarin speech recognition with very large vocabulary,” *Proceeding of ICSLP-94*, Yokohama, Japan, 1994, pp. 1371-1374.