

基於 Ontology 之中文文件自動摘要技術之研究

李健興 陳雅絹 郭雅琪 莊宏翊

長榮大學資訊管理學系

摘要

目前一般的文件自動摘要技術，多半以摘錄文章的某一段落或使用統計分析方法，計算句子的權重分數、位置以擷取重要的句子形成摘要，但其摘要內容之正確率、可讀性和整體連貫性卻有其不足之處。本論文建構一中文文件自動摘要系統，以 Ontology 為基礎架構，並且以「社會犯罪」領域之電子新聞為實驗主體，應用模糊推論 (Fuzzy Inference) 模擬人類之語意思考邏輯，以判斷新聞文件與 Domain Ontology 之間任一詞組 (Term-Pair) 之語意關聯強度。首先，資訊擷取代理人 (RA, Retrieval Agent) 每隔一段固定時間，即自動抓取網路電子報的社會新聞存放於資料庫，同時將文章內容透過文件前置處理代理人 (DPA, Document Pre-processing Agent) 加以斷詞和過濾後，模糊推論代理人 (FIA, Fuzzy Inference Agent) 將自動推論出重要的關鍵詞彙以形成 Dynamic sub-Ontology。再經由摘要代理人 (SA, Summarization Agent) 萃取出能夠表示原始文章內容主題且富含語意之摘要精簡版。除此之外，藉由 Web 操作介面，讓使用者在短時間內即能快速且有效地獲得符合需求的資訊。

關鍵詞彙：自動摘要技術，Ontology，模糊推論，代理人

壹 緒論

近年來，由於資訊科技與網際網路的蓬勃發展，無時無刻產生大量的資料，同時也產生了資訊氾濫的問題，使用者必須花更多的時間以人工方式篩選出有用或有意義的資訊。如何快速地取得有用的資訊，以及有效地應用這些資訊成為寶貴的知識，使得資訊擷取相關技術日益受到重視。大多數的網路線上文件都以全文格式呈現，無可避免的會產生資訊重複、繁雜或內容冗長等現象，使用者必須逐行逐字瀏覽觀看，無法迅速確實的掌握文件內容，不僅浪費網路瀏覽時間，而且缺乏效率和效果，由此可見，文件自動摘要技術是一個相當重要的趨勢。

人類的溝通與閱讀主要以文字、概念 (Concept) 為單位以傳達訊息，而文件背後也是在表達概念含意，摘要是指以較少或特定長度的文字來代表原始文件所欲傳達的意涵，其目的在於能產生一簡單明瞭的內容敘述，可讓使用者在很短的時間限制下，迅速、確實地掌握原始文件所包含的意思。為協助使用者能快速瀏覽線上文件，以判斷文件是否為所要尋找或感興趣的資訊，需要

運用文件擷取相關技術，萃取出文章中重要內容以形成文件之摘要。本論文應用具有提供語意概念的 Ontology 為基礎架構，運用資訊擷取、模糊推論等方法，建構出一個既能表達語意又符合個人化需求之文件自動摘要系統。如此一來，不但能提供使用者快速且有效地瀏覽文件，亦能避免冗長及重複資訊的產生，以減少時間和成本上的浪費。

本篇論文架構如下：第二節介紹相關之文獻探討，第三節提出文件自動摘要之系統架構，第四節說明本論文之研究方法，第五節列舉實例來說明文件摘要之流程，最後，第六節為本篇論文之結論。

貳 相關文獻探討

網際網路上大量資訊的容易取得，相對地衍生出資訊過量的困擾，使得使用者在面對如此龐大且多元的大型資料庫時，無法有效且快速地找到符合需求之資訊，因此，許多國內外學者乃相繼積極投入資訊擷取相關技術研究領域，而文件自動摘要技術亦即扮演了重要角色之一，且陸續有相當成效的研究成果發表出來。

一、文件自動摘要

文件摘要是指從原始文件中精鍊出最重要資訊的過程，其結果足以代表該原始文件之精簡化版本，且可作為人們判斷或其他資訊系統之決策依據 (Mani, 1999)。一般評估摘要句子的重要性，可考慮關鍵詞之出現頻率與位置以及上下文的關係，是以多年來學者的研究，多著重於向量統計或語意分析技巧，藉以摘錄文中具有代表性的句子 (黃純敏, 1999)；Kupiec et al. (1995) 利用字數統計方法，段落位置分析等特徵當作摘要選取的依據；Mani (2001) 介紹兩種摘要方法包括主題集中式和一般式摘要，認為在網際網路環境中，主題集中式摘要方法較適合於全文搜尋和瀏覽。Lam & Ho (2001) 提出 FIDS 用以作線上財經新聞文章之自動摘要，FIDS 可將不同來源出處的相關文件加以整合處理。Habn & Mani (2000) 定義文件摘要過程可分有三個階段：首先是「分析原始文件 (Analyze the source text)」，接著「決定重要的特徵值 (Determine it's salient points)」，最後是「選擇適當的表示法以形成摘要之輸出格式 (Synthesize an appropriate output)」。

文件摘要依原始文件數量的多寡可分為單文件摘要 (Singular Document Summarization) 與多文件摘要 (Multiple Document Summarization) (翁鴻加，

2001), 單文件摘要著重的是能否有效地刪減單篇文件內容中多餘且非必要性的資訊, 以求達到摘要之精簡化與重點化; 多文件摘要則是將多篇探討相同主題或事件的文件融合成單篇摘要, 除了注重刪減無用的資訊外, 還在於強調過濾重複的資訊; 從語言的角度來看, 摘要又可分為單語言摘要 (Monolingual Summarization) 及多語言摘要 (Multilingual Summarization), 所謂多語言摘要即指原始文件內容包含多國語言之意; 若依使用者需求的不同, 摘要結果可分為一般性摘要 (Generic Summary) 及特定使用者導向 (User-Oriented Summary) 摘要等; 而根據文件摘要所要達到的目的, 摘要產出結果主要可分為指示性摘要 (Indicative Summary)、資訊性摘要 (Informative Summary) 與評論性摘要 (Critical Summary) 三種 (Hovy & Lin, 1999):

- 指示性摘要: 指示性摘要提供足夠的資訊給使用者, 使其能夠依此判斷決定是否更深入地閱讀原始文件。
- 資訊性摘要: 資訊性摘要提供較豐富的資訊內容, 甚至於可以取代原始文件內容之意涵。
- 評論性摘要: 評論性摘要以摘要形式對原始文件作評論, 提供使用者從不同的角度評論。

綜上所述, 本論文之研究在於中文單一語言、一般性、指示性、單文件之自動摘要的產生。

二、Ontology

Ontology 最早是由哲學家所提出, 中文譯為「本體論」或「實體論」, 是指用來處理生命體或現實事物本質之存在理論, 後來被延伸應用在人工智慧方面, 乃用以描述知識或表現知識, 簡而言之, Ontology 可以定義說明某一特定領域知識或主題 (Zeng & Lee, 2001), 其內容包含了許多物件 (Object)、物件特質屬性 (Property) 和物件之間的關係 (Relation) 以表示某一特定領域的知識 (Chandrasekaran et al., 1999)。Ontology 主要包含下列這些元素: Classes (也稱為 Concepts、Objects)、Properties (也稱為 Attributes、Slots、Roles) 以及 Relations。Classes 為 Ontology 中最主要的部分, 是用來描述所要說明領域中的概念, Properties 是用來描述 Classes 或 Concepts 的特性或屬性, Relations 是用來說明 Class 與 Class 之間的關係。由於 Ontology 具有能提供語意的基礎, 使得它之於軟體代理人 (Software Agent)、電子商務 (E-commerce) 以及知識

管理 (Knowledge Management) 之發展有很大的幫助。要發展一個 Ontology 包含下列四個步驟，藉由此四個步驟，即可建構出一個 Domain Ontology 知識庫(Noy & McGuinness, 2001)。

1. 定義 Ontology 中的 Class。
2. 定義 Class 與 Class 之間的階層關係 (Subclass- Superclass)。
3. 定義 Class 中的屬性，並且說明對於屬性值的限制。
4. 將 Instance 的屬性值填入。

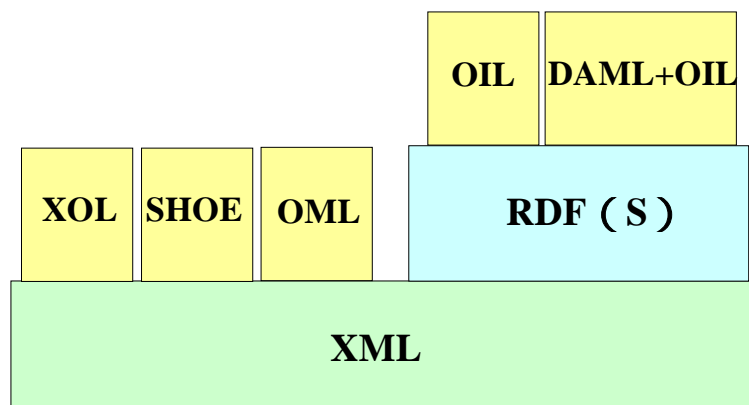
Ontology 是用來定義某個領域中的一些基本概念及它們之間的關連，其主要的用意是為了讓電腦更容易閱讀這些知識，OntoSeek (Guarino et al., 1999) 即是一個結合 Ontology-Driven 內容比對機制之資訊擷取系統；Embley et al. (1998) 提出了基於 Domain Ontology 的方法，應用於如何自非結構性的文件中擷取資訊。Ontology 的功用可由下列四點說明 (Noy & McGuinness, 2001)：

1. 一些相同領域的專家或機構可用相同的 Ontology 來定義 Class 或 Concept，如此軟體代理人或使用者可藉由 Ontology 而達到資訊共享之目的。
2. 當需要建立一個大範圍的 Ontology 時，可利用已存在的 General Ontology (ex. WordNet、UNSPSC Ontology) 或一些 Domain Ontology 來作為輔助。
3. Ontology 中定義的 Class 及 Relation 對於某個 Domain Knowledge 可做更明確的定義，當一個初學者想要瞭解某個 Domain Knowledge，可藉由 Ontology 而得到幫助。
4. 可利用已經存在的 Domain Ontology，來分析 Domain Knowledge 並重複使用 Ontology 或對 Ontology 做擴充。

三、Ontology Language

Ontology 可以被應用在各個不同的領域及應用層面，例如：智慧型代理人系統、知識管理、自然語言處理和網路資訊搜尋擷取等。然而，在此之前必須先讓電腦看得懂專家所建構出的 Ontology，因此需要一種電腦所能理解的語言來轉換描述 Ontology，以便電腦瞭解 Ontology 所欲表達之語意性概念。近年來，已經發展出許多 Ontology Languages，有些是基於 XML 語法，例如：XOL、SHOE 和 OML；還有由 W3C 協會 (World Wide Web Consortium) 所制

訂發展的 RDF 和 RDFS (RDF Schema)；另外，尚有建置於 RDFS 基礎上，即 OIL 與 DAML+OIL，其是為了改善加強 RDF(S) 功能特質之不足。如圖一顯示各語言之層級關係 (Asunción & Corcho, 2002)，以下將簡要說明各種 Ontology Languages。



圖一 Ontology Languages 之層級關係

(一)XOL

XOL (XML-based Ontology Exchange Language)，是在 1999 年由 Peter D. Karp, Vinay K. Chaudhri, and Jerome Thomere (Karp et al. 1999) 等專家所發展出來的。它是基於 XML (eXtensible Markup Language) 的 Ontology Language，可應用於不同異質系統之間的訊息交換。

(二)SHOE

SHOE (Simple HTML Ontology Extensions)，起源於馬里蘭大學 College Park 資訊科學系所的 Understanding Systems Group 所提出的 SHOE 計畫 (xml.coverpages.org)。SHOE 是一種 HTML 的延伸語言，它允許網頁作者在他們的網頁文件中註解機器可讀取的語意知識。SHOE 使得線上智慧代理人 (Intelligent Agent) 能更有效率地執行搜尋網頁和文件內有意義的資訊，這過程包含三個步驟：1. 定義 Ontology。2. 使用 HTML 來註解 Ontology 知識及描述它們和其他網頁的關連。3. 設計一個 Agent 使它可搜尋到語意網頁資訊 (Asunción & Corcho, 2002)。

(三) OML

OML (Ontology Markup Language), 是 1998 年 8 月由華盛頓大學所提出來的, 其嘗試以 XML 的語法為基礎來表達 Ontology, 且其有部分是建立於 SHOE (Simple HTML Ontology Extension) 的基礎上, 也就是說, OML 和 SHOE 有許多共有的特徵 (Asunción & Corcho, 2002)。

(四) RDF & RDF Schema

資源描述架構 (RDF, Resource Description Framework), 是 1999 年 2 月由全球資訊網協會 (W3C) 主導所發展而成的一個架構, 其提供一具有語意網路之機制, 可用來描述網頁資源, 允許資源描述機構訂定各自的控制詞彙, 提供結構化的相互相容機制, RDF 在語法上則是遵循 XML。

(五) OIL

OIL (Ontology Inference Layer or Ontology Interchange Language), 是從 On-To-Knowledge Project 中所發展出來的一種 Ontology 表達語言, 而這個 Project 是從 1999 年開始到現在, 其成員是由業界及學術界所組成的。然而 OIL 之所以會被發展出來是為了能正確地去表達在 Web 上電腦可以存取有語意的資訊, 而這些有語意的資訊必需能符合電腦可以存取的格式。而 OIL 的語法及語意是建構在已存在的一些標準語言上, 如 Open Knowledge Base Connectivity (OKBC)、XML-based Ontology Exchange Language (XOL)、Resource Description Framework (RDF) 等。目前可用以建構 OIL Ontology 的工具軟體有 OntoEdit、OILEd、WebODE 和 Protege2000 (Asunción & Corcho, 2002) (Fensel, 2000)。

(六) DAML+OIL

DAML (DARPA Agent Markup Language), 是由美國國防部高等研究計畫局研發。為比 RDF 能更佳表達 RDF Class 的定義, 在美國政府倡議的努力下, 於 2000 年 10 月發行了 DAML-ONT, 這是一種符合 RDFS 的簡單語言。不久 DAML 小組為提供更多的功能服務, 乃朝向結合 Ontology Inference Layer (OIL) 而努力, 使用人工智慧框架基礎 (Frame-based) 的架構。這些努力的結果即為後來發展的 DAML+OIL。DAML+OIL 是 Web Resources 中可以用來描述語意的 Ontology 標記語言。它是以 W3C 早期所制定的標準 (RDF 及 RDF Schema) 而建立的, 且擴充了許多 modelling primitives 於此語言。DAML+OIL

所提供的 Modelling Primitives，通常都是出自於 Frame-based Languages 裡。這樣的 DAML+OIL Language 是相當簡潔且容易來定義語意的。目前用以建構 OIL Ontology 的工具軟體如：OntoEdit, OIEd, WebODE 和 Protege2000，亦可以用來建構 DAML+OIL Ontology (Asunción & Corcho, 2002) (Fensel, 2000)。

本論文採用 DAML+OIL 標記語言來描述 Ontology，因 Daml+OIL Language 可引用 XML Schema 來訂定資料的型態，簡潔、容易定義語意且符合 W3C 所制訂的標準需求。

四、模糊推論 (Fuzzy Inference) 與代理人 (Agent)

(一)模糊推論 (Fuzzy Inference)

Fuzzy 理論是為解決真實世界中普遍存在的模糊現象而發展的一門學問，是由美國自動控制學家 Lotfi. A. Zadeh 於 1965 年首先提出的一種定量表達工具，用來表現某些無法明確定義的模糊性概念，尤其是在表現人類語言特有的模糊性現象方面有頗佳的成果 (孫宗瀛, 2001)。模糊理論發展至今，相關的研究成果也迅速增長，其應用發展延伸至許多科技學科層面，如：人工智慧、自動控制、圖像識別、決策支援等，在各個領域都有豐碩的應用成果。早在 1970 年代中期，Fuzzy 研究已有許多成功的實例，1980 年代 Fuzzy 應用開始實用化、商品化，到了 1990 年代無庸置疑地已經造成了一股研究熱潮。Lin & Lee (1991) 應用模糊邏輯控制於決策支援系統上；亦有將模糊推論應用在影像處理方面的 (Lee et al., 1997)；Lee et al. (2002) 將模糊推論應用在網路電子新聞之分類上，並透過智慧型代理人達到個人化的需求。而近代，Fuzzy 理論更是與知識工程、類神經網路、基因演算等多項技術相結合，進而有了更多的新突破與研究發展。

(二)代理人 (Agent)

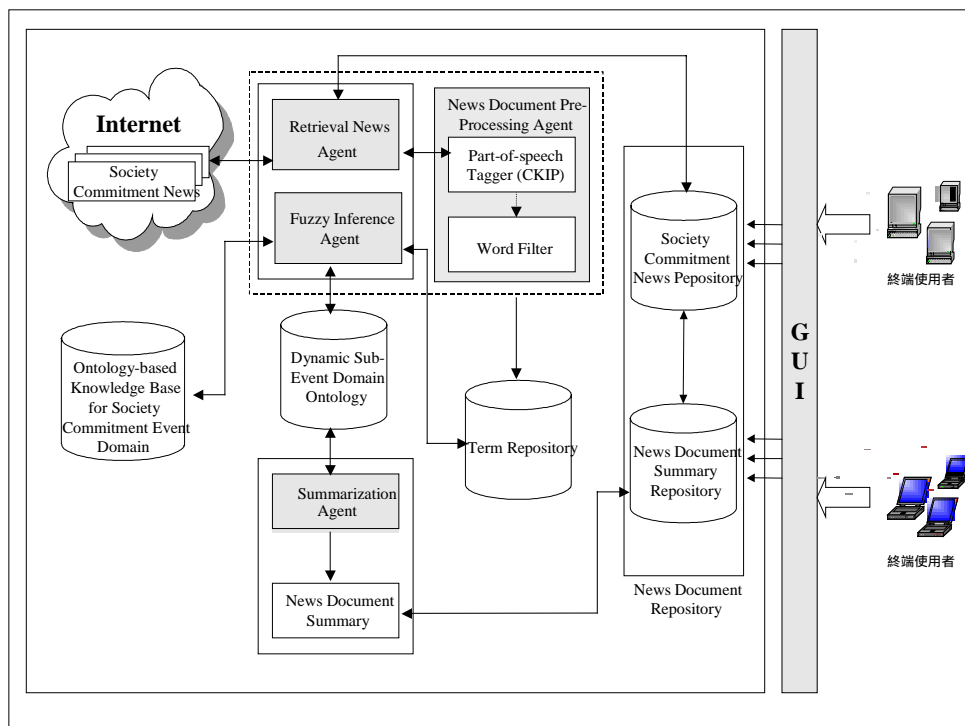
一個代理人 (Agent)，簡單來說就是由電腦語言所寫成的軟體程式，其可以協助使用者完成特定的任務工作。J. Ferber (1999) 則定義代理人為「一個代理人是能夠在一個環境中行動且可直接與其他的代理人互相溝通」；而所謂智慧型代理人 (Intelligent Agent) 則比一般代理人更有強大效益，因為其具有學習和推論的能力 (Chen & Yen, 1996)。由於網路上資訊的爆炸，使用者在處理資訊時往往無法事必躬親，促使智慧型代理人在網際網路技術發展中日益重要，使用者可委託軟體代理人執行繁複的工作，且各代理人程式之間也可達到

相互溝通、協調，甚至於個人化處理的能力。Huhns & Singh (1999) 提出代理人程式應用於電信服務訂單之工作流程處理上；Maes et al. (1999) 提出利用代理人程式，能夠協助使用者在電子商務環境中進行商務活動；亦有設計一具有模糊推論之代理人程式，應用在會議排程之決策支援系統上 (Lee et al., 2002)。智慧型代理人應用層面相當廣泛，如：電子商務、數位圖書館、遠距教學、知識管理等。Web 應用之資訊代理人技術是未來不可或缺的網路技術，其主要是要建立一個具有自主性、學習能力、行動能力、個人化、推理能力、以及協調溝通能力的軟體代理人，以協助人們處理資訊。

參 系統架構

一、摘要系統架構

本摘要系統為一個基於 Ontology 架構的模糊推論之事件萃取摘要系統，並以「東森電子報」網站 (<http://www.ettoday.com.tw>) 上之「社會犯罪」領域新聞文件為實驗主體來源。圖二為本系統之完整架構。



圖二 文件自動摘要系統之架構

依據專家知識以人工方式建置一有關「社會犯罪」之 Domain Ontology，提出包含三個軟體代理人，即資訊擷取代理人 (RA, Retrieval Agent)、文件前置處理代理人 (DPA, Document Pre-processing Agent) 以及模糊推論代理人 (FIA, Fuzzy Inference Agent)。首先，RA 會從網路上之電子報網站自動抓取社會犯罪領域之電子新聞，並將其轉換為 XML 格式檔案儲存至資料庫，同時將電子新聞文件傳送給 DPA 作前置處理，DPA 將利用由中央研究院所研發之 CKIP 斷詞斷句系統 (CKIP, <http://godel.iis.sinica.edu.tw/CKIP/>)，將文件內容之詞彙加以斷詞並標註詞性。由於文件中並非所有的關鍵詞都具有相同的重要性，一般說來，名詞和動詞的重要性要比冠詞、介系詞或感嘆詞等來得重要，且大多數語句都是「主詞-述詞-受詞」的結構 (葉鎮源, 2002)，而文件裡的主詞與受詞大部分都是名詞，述詞通常是動詞，因此，我們藉由保留重要的名詞和動詞以理解文件之語意，再透過詞字過濾器 (World Filter) 剔除不重要的詞 (如：的(DE)、了(T)、把(P)...等)。CKIP 的標記意義舉例如表一所示。

表一 CKIP 定義之標記意義

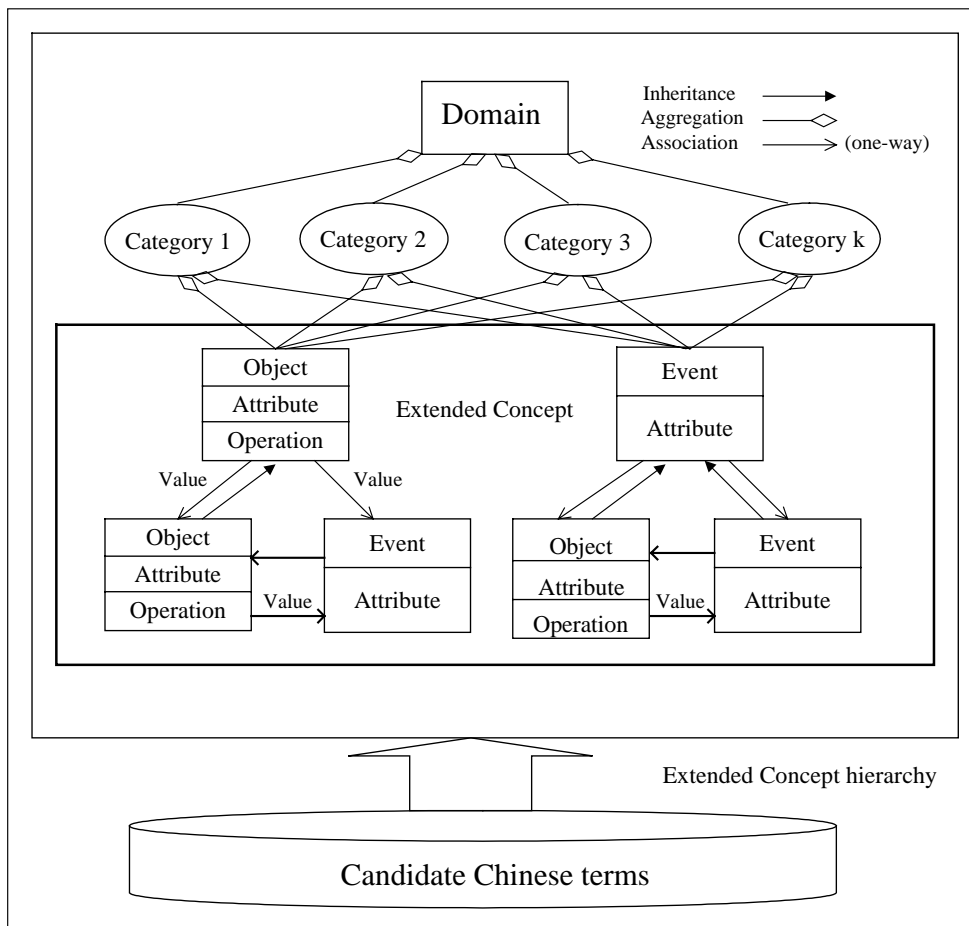
POS Tag	Meaning	Example
Na	普通名詞 (Common noun)	檢察官 (Na)
Nc	地方詞 (Place noun)	警察局 (Nc)
Ncd	位置詞 (Location noun)	東南東方 (Ncd)
VA	動作不及物動詞 (Intransitive verb)	降雨 (VA)
VAC	動作使動動詞 (transitive verb class)	移動 (VAC)
VB	動作類及物動詞 (Single verb class)	延期 (VB)
VC	動作及物動詞 (Single verb)	排除 (VC)
VCL	動作接地方賓語動詞 (Active location object verb)	遠離 (VCL)
VD	雙賓動詞 (Two words verb)	索取 (VD)
VE	動作句賓動詞 (Verb sentence)	表示 (VE)
VF	動作謂賓動詞 (Name verb)	打算 (VF)
VH	狀態不及物動詞 (Status Intransitive verb)	出現 (VH)
VHC	狀態使動動詞 (Status transitive verb class)	增強 (VHC)
VJ	狀態及物動詞 (Status single verb)	發生 (VJ)

之後，FIA 將接收 DPA 過濾處理後之詞字和詞性集合，透過本系統所提之模糊推論機制 (Fuzzy Inference Mechanism) 與 Domain Ontology 內的詞彙

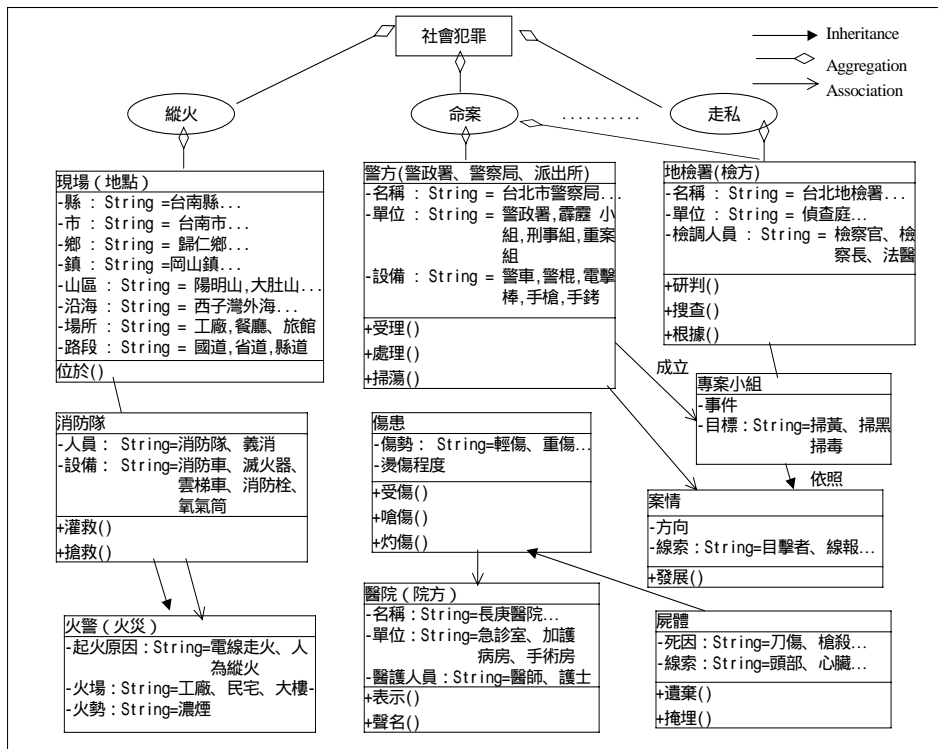
兩兩組合成一詞組 (Term-Pair) 進行模糊推論，推算並找出語意相關且重要的詞彙，以利後續的摘要代理人 (SA, Summarization Agent) 作摘要形成。

一、Domain Ontology架構

本論文提出一個 Domain Ontology 架構 (如圖三所示)，作為專家建構 Ontology 的基礎框架。在此架構中包含三個層級定義，由上而下依序是領域層級 (Domain Layer)、類別層級 (Category Layer) 以及延伸的概念層級 (Extended Concept Layer)，而在延伸的概念層級又包括兩種概念節點，即事件概念節點 (Event Concept Node) 和物件概念節點 (Object Concept Node)。概念節點之間定義存在有三種關係，分別是繼承關係 (Inheritance)、聚合關係 (Aggregation) 以及結合關係 (Association)。圖四顯示一個為社會犯罪領域一部份之 Ontology 架構例子，說明如下：



圖三 Domain Ontology 架構



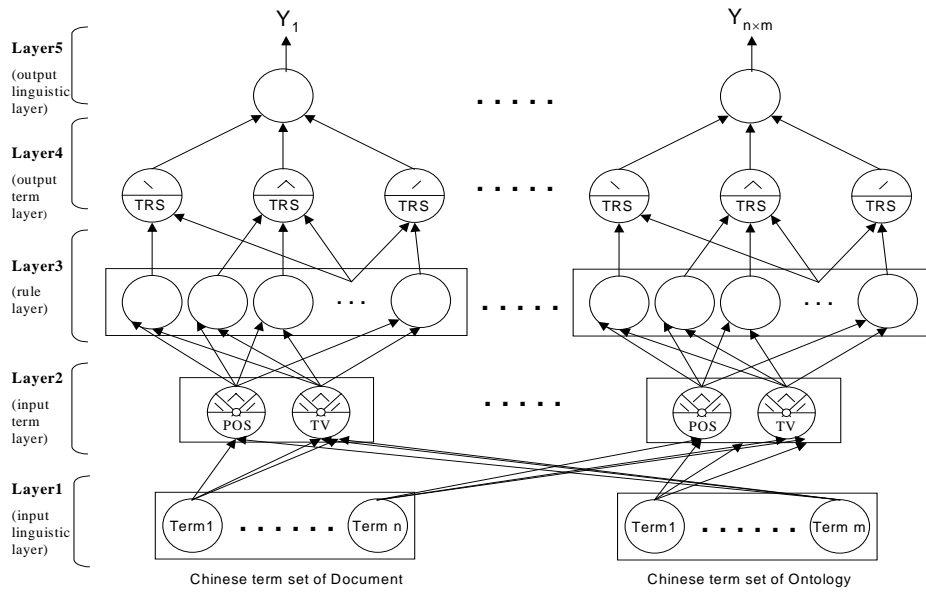
圖四 社會犯罪領域之部分 Ontology

在圖四中的領域名稱是”社會犯罪”，它可包含幾個類別如”縱火”、”命案”...等，另外，有一些概念如”警方”、”傷患”、”案情”...等，與上層各類別間分別存在不同的隸屬程度關係，如就”警方”這個概念節點來看，它又包含了所屬的屬性名稱是”名稱”、”單位”、”設備”和屬性值分別是”台北市警察局”、”警政署”、”警車”...等等，如此依據所提出之 Domain Ontology 架構及步驟，即可建構出一完整的社會犯罪領域之 Ontology。

肆 研究方法

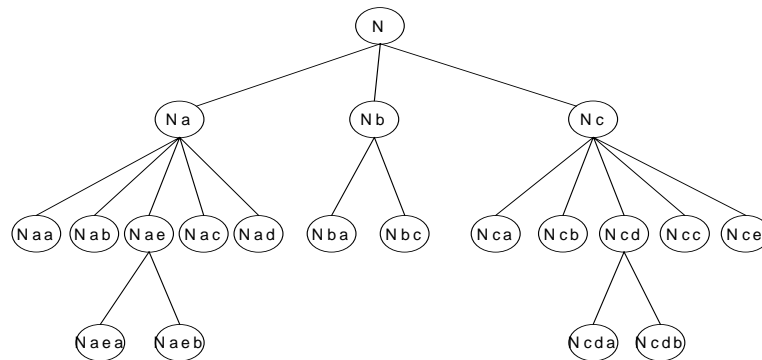
一、模糊推論機制 (Fuzzy Inference Mechanism)

本論文應用模糊理論，提出一個五階層式架構之模糊推論機制（如圖五所示），以進行每一詞組 (Term-Pair) 間語意相似度之推論。



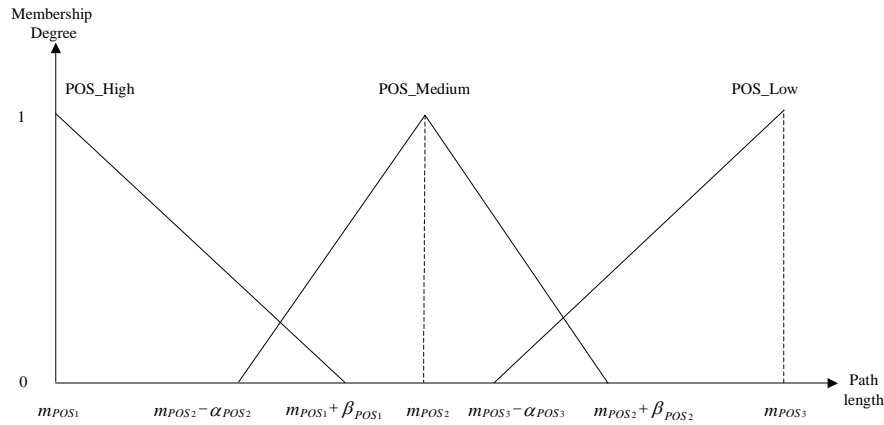
圖五 模糊推論機制

在模糊推論機制中包含五個層級分別是 Input Linguistic Layer、Input Term Layer、Rule Layer、Output Term Layer 以及 Output Linguistic Layer。Input Linguistic Layer 包含兩種輸入詞彙，一是 RA 抓取下來之電子新聞文件，經過 DPA 過濾處理後的詞彙，另一是 Domain Ontology 內的所有詞彙。此機制中定義有二個輸入模糊變數，分別是詞性相似度 (POS, Part-of-Speech Similarity) 和詞字相似度 (TV, Term Vocabulary Similarity)。圖六表示一個為 CKIP 所定義之 Tagging Tree (Lee et al., 2002)，作為計算任一詞組詞性距離之依據。例如：“警察 (Na)，警察局 (Nc)”此一詞組，其對照 Tagging Tree 之詞性路徑為 (Na N Nc)，故其詞性距離計算結果等於 2，即作為輸入模糊變數詞性相似度 (POS Similarity) 之輸入值。



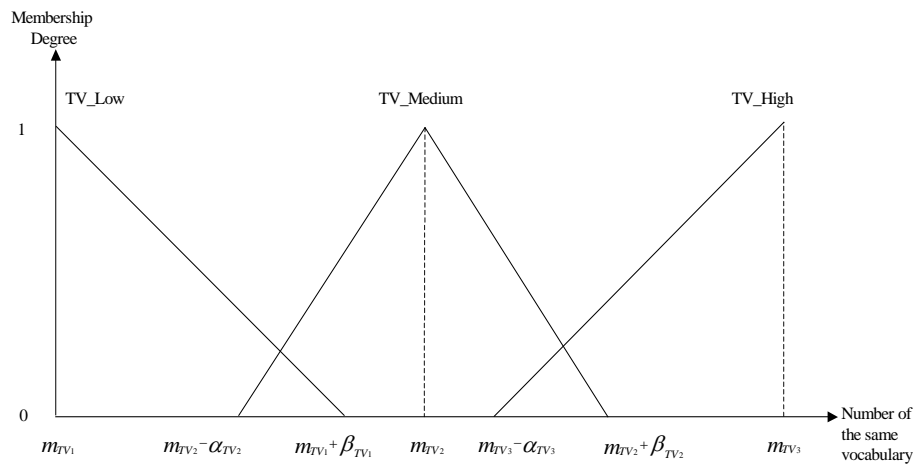
圖六 CKIP 定義之 Tagging Tree

我們利用任兩個詞性標記節點的路徑距離來推算兩個詞彙，即為詞組 (Term-Pair) 之間的 *POS Similarity*，詞性距離長度界定在 $[0, 6]$ 之間，並訂定三個語意項 (Linguistic Term) 為 *POS_High*、*POS_Median* 以及 *POS_Low*，如圖七所示。



圖七 任兩個詞彙之間 *POS Similarity* 之模糊集合

另一個定義之輸入模糊變數是詞字相似度 (*TV Similarity*)，我們利用任兩個詞彙之間相同字個數來計算兩個詞之間的詞字相似度，相同字個數界定在 $[0, 6]$ 之間，並訂定三個語意項為 *TV_High*、*TV_Median* 以及 *TV_Low*，如圖八所示。例如：“員警，警察”此一詞組，相同字為「警」字，故其相同字個數值為 1，又當詞組中的詞彙前後字亦相同時則分別再加上 0.5，例如：詞組為“台北縣，台南縣”中的「縣」字；“警察，警員”中的「警」字，運算後之結果用以當作輸入模糊變數詞字相似度 (*TV Similarity*) 之輸入值。

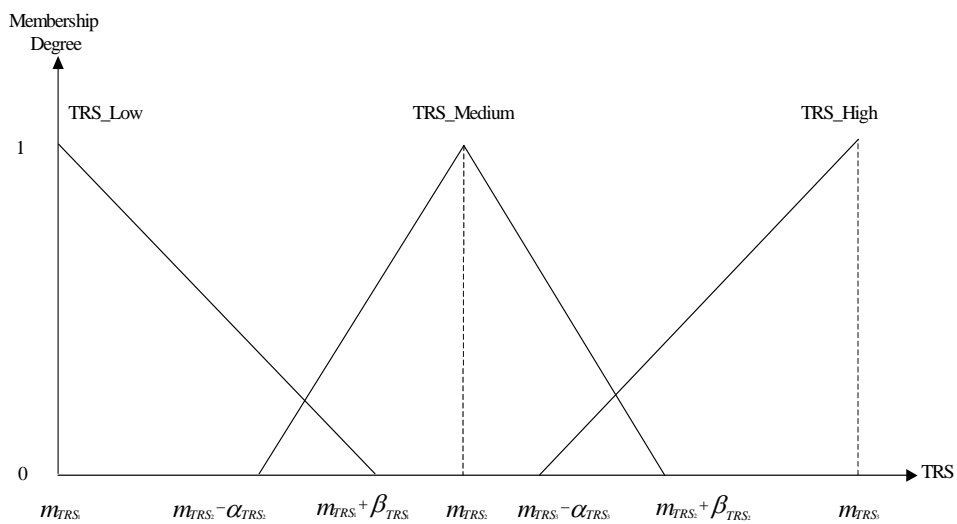


圖八 任兩個詞彙之間 *TV Similarity* 之模糊集合

本論文以專家經驗假設定九條推論法則，用以推算 *POS*、*TV* 二個輸入模糊變數之每一語意項所對應得到之 *Matching Degree*，利用正規化運算得到輸出模糊變數 *TRS* (*Term Relation Strength*) 值，並依據重心法求算得到每一條 Rule 推導之重心值，最後再作加權平均計算出每一詞組的語意相關程度值 (Lin & Lee, 1991)。表二顯示預先定義之九條推論法則，圖九為輸出模糊變數 *TRS* 之模糊集合。

表二 *FIA* 之推論法則

Rules	Fuzzy variables	<i>POS</i> Similarity	<i>TV</i> Similarity	<i>TRS</i>
1		L	L	L
2		L	M	L
3		L	H	M
4		M	L	L
5		M	M	M
6		M	H	H
7		H	L	M
8		H	M	H
9		H	H	H

圖九 輸出模糊變數 *TRS* 之模糊集合

二、實驗分析

在這一節，我們分別描述五階層式模糊推論機制各層級的輸入、輸出，並將部分實驗分析結果呈現如表三所示。

Layer1 (Input Layer)

Input: Domain Ontology 與新聞文件的詞彙。

Output: 任一詞組 (Term-Pair) 之詞性和詞字。

Layer2 (Input Linguistic Layer)

Input: 任一詞組 (Term-Pair) 的詞性和詞字。

Output: 輸入模糊變數 *POS* 與 *TV* 之每一個 Linguistic Term 之 Membership Degree。

Layer3 (Rule Layer)

Input: 輸入模糊變數 *POS* 與 *TV* 之每一個 Linguistic Term 之 Membership Degree。

Output: 每一條推論法則之 Premise Part 之 Matching Degree。

Layer4 (Output Linguistic Layer)

Input: 每一條推論法則之 Premise Part 之 Matching Degree。

Output: 每一條推論法則之 Premise Part 之 Matching Degree 對應到輸出模糊變數 TRS 計算得到之重心值。

Layer5 (Output Layer)

Input: 每一條推論法則之 Premise Part 之 Matching Degree 對應到輸出模糊變數 TRS 計算得到之重心值。

Output: 所有重心值加權平均後得到任一詞組之語意相關強度值(Y)。

表三 部分實驗分析結果

Term pair	POS	TV	Y
A	0	3	1
B	0	4	1
C	0	2	0.92183813413992
D	0	0	0.39395171691022
E	5	0	0.00460997068934
F	6	0	0

由推論資料我們發現，本摘要系統之研究方法尚有其可能的缺點待改善，在模糊推論機制 (Fuzzy Inference Mechanism) 方面，由於我們的模糊推論機制中的模糊歸屬函數 (Membership Function) 及推論法則庫 (Fuzzy Rule Base)，均事先由領域專家憑藉經驗法則或直觀所決定的，相對較缺乏整體客觀性與正確性，其所引用的參數值是否適當，乃需要嘗試錯誤來調整才能獲得較理想的模糊模式 (Fuzzy Model)。不過，本摘要系統方法可實現平行模糊推論 (Parallel Fuzzy Inference) 之效益，是本論文之立足優點。

伍 摘要實例說明

在這一節，我們實際舉二篇「社會犯罪」領域之電子新聞為例，來說明摘要系統的實作流程，圖十為一篇由「東森電子報」網站上擷取下來之網路原始電子新聞，經由資訊擷取代理人 (RA) 所轉換後之 XML 格式檔。

例子一：

```
<?xml version="1.0" encoding="BIG5" ?>
<新聞>
<抬頭>夜半無名火基隆公寓騎樓疑遭縱火 1 死 28 傷</抬頭>
<新聞類別>社會萬象</新聞類別>
<日期>08/05</日期>
<記者>江宏龍</記者>
<地點>基隆報導</地點>
<內容>
基隆七堵崇智街某民宅，5 日凌晨傳出火警，雖然火勢很快就被撲滅，不過因為民眾都在熟睡當中，消防隊員一共救出 30 多人，有 28 人受傷，其中一人不幸喪生，警方不排除是人為縱火。消防隊員冒險進入火場，一一將受到濃煙嗆傷的民眾救出火場，這起火警是發生在今天凌晨 3 時 10 分。基隆市七堵崇智街 2 之 9 號的騎樓，不知道為何突然發出一聲巨響，瞬間將停放在現場的 20 幾部機車燒毀，而火舌也順著樓梯間開使漫延。由於當時正是民眾熟睡的時刻，許多人都來不及逃出，消防隊及基隆救難大隊隨即衝入火場，一一將受到嗆傷的 26 位居民救出送醫急救，其中一名翟姓中年男子因為傷勢嚴重，在送醫途中就不幸身亡。警方事後調查發現，造成這起意外，不排除有人為縱火的可能，不過也有鄰居看到，案發當時有人在燃燒金紙，到底是如何起火，警方將會同消防鑑識人員，調查詳細原因。
</內容>
</新聞>
```

圖十 原始電子新聞文件之 XML 格式

圖十一、圖十二分別為新聞文件經過 CKIP 斷詞系統斷詞後和 DPA 過濾後之結果。

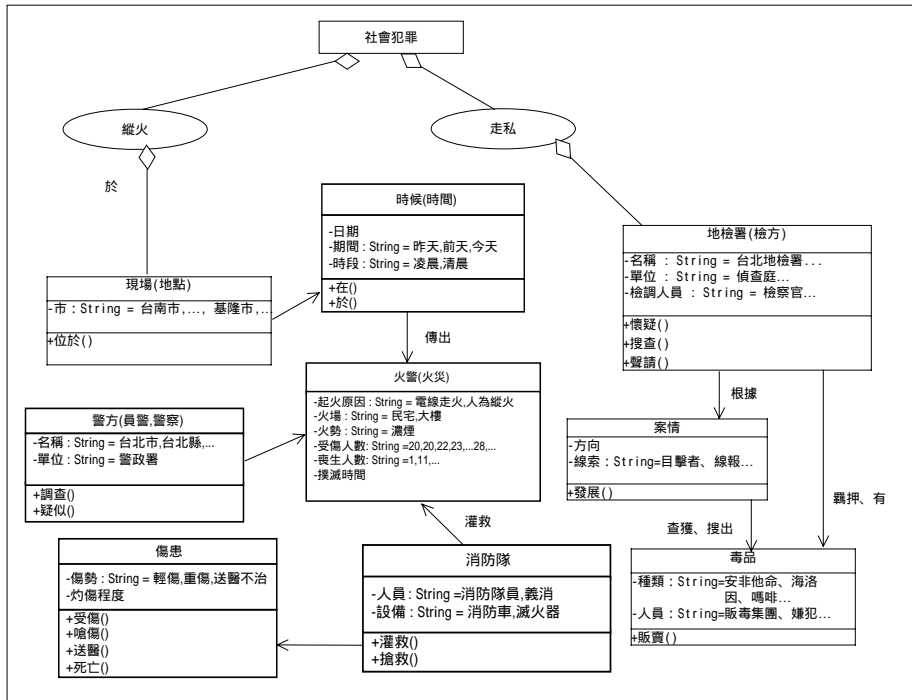
基隆(Nc) 七堵(Nc) 崇(VJ) 智(Na) 街(Na) 某(Nes) 民宅(Na) 5(FW) 日(Nd) 凌晨(Nd) 傳出(VC) 火警(Na) 雖然(Cbb) 火勢(Na) 很(Dfa) 快(VH) 就(D) 被(P) 撲滅(VC) 不過(Cbb) 因為(Cbb) 民眾(Na) 都(D) 在(P) 熟睡(VA) 當中(Ng) 消防隊員(Na) 一共(Da) 救出(VC) 30(FW) 多(Neqa) 人(Na) 有(V_2) 28(FW) 人(Na) 受傷(VH) 其中(Nep) 一(Neu) 人(Na) 不幸(VH) 喪生(VH) 警方(Na) 不(D) 排除(VC) 是(SHI) 人(Na) 為(VG) 縱火(VA) 消防隊員(Na) 冒險(VA) 進入(VCL) 火場(Nc) 一一(D) 將(D) 受到(VJ) 濃煙(Na) 嗆(VH) 傷(VH) 的(DE) 民眾(Na) 救出(VC) 火場(Nc) 這(Nep) 起(Nf) 火警(Na) 是(SHI) 發生(VJ) 在(P) 今天(Nd) 凌晨(Nd) 3(FW) 時(Ng) 10(FW) 分基隆市(Nc) 七堵(Nc) 崇(VJ) 智(Na) 街(Na) 2(FW) 之(DE) 9(FW) 號(Na) 的(DE) 騎樓(Na) 不(D) 知道(VK) 為何(D) 突然(D) 發出(VC) 一(Neu) 聲(Nf) 巨響(Na) 瞬間(Nd) 將(D) 停放(VC) 在(P) 現場(Nc) 的(DE) 20(FW) 幾(Neu) 部(Nf) 機車(Na) 燒毀(VC) 而(Cbb) 火舌(Na) 也(D) 順(VJ) 著(Di) 樓梯間(Nc) 開(VC) 使(VL) 漫延(VH) 由於(Cbb) 當時(Nd) 正(D) 是(SHI) 民眾(Na) 熟睡(VA) 的(DE) 時刻(Na) 許多(Neqa) 人(Na) 都(D) 來不及(D) 逃出(VCL) 消防隊(Na) 及(Caa) 基隆(Nc) 救難(VA) 大隊(Na) 隨即(D) 衝入(VCL) 火場(Nc) 一一(D) 將(D) 受到(VJ) 嗆(VH) 傷(VH) 的(DE) 26(FW) 位(Nf) 居民(Na) 救出(VC) 送醫(VB) 急救(VB) 其中(Nep) 一(Neu) 名(Nf) 翟(Na) 姓(VG) 中年(Na) 男子(Na) 因為(Cbb) 傷勢(Na) 嚴重(VH) 在(P) 送醫(VB) 途中(Nc) 就(D) 不幸(VH) 身亡(VH) 警方(Na) 事(Na) 後(Ng) 調查(VE) 發現(VE) 造成(VK) 這(Nep) 起意(VF) 外(Ng) 不(D) 排除(VC) 有(V_2) 人(Na) 為(VG) 縱火(VA) 的(DE) 可能(VH) 不過(Cbb) 也(D) 有(V_2) 鄰居(Na) 看到(VE) 案發(VH) 當時(Nd) 有(V_2) 人(Na) 在(P) 燃燒(VC) 金紙(Na) 到底(D) 是(SHI) 如何(D) 起火(VH) 警方(Na) 將(D) 會同(VC) 消防(A) 鑑識(VC) 人員(Na) 調查(VE) 詳細(VH) 原因(Na) 內容(Na)

圖十一 CKIP 斷詞結果

基隆(Nc) 七堵(Nc) 民宅(Na) 凌晨(Nd) 傳出(VC) 火警(Na) 火勢(Na) 撲滅(VC) 民眾(Na) 熟睡(VA) 消防隊員(Na) 救出(VC) 受傷(VH) 不幸(VH) 喪生(VH) 警方(Na) 不排除(VC) 人為(VG) 縱火(VA) 消防隊員(Na) 冒險(VA) 進入(VCL) 火場(Nc) 受到(VJ) 濃煙(Na) 嗆傷(VH) 民眾(Na) 救出(VC) 火場(Nc) 火警(Na) 發生(VJ) 今天(Nd) 凌晨(Nd) 基隆市(Nc) 七堵(Nc) 騎樓(Na) 不知道(VK) 為何(D) 突然(D) 發出(VC) 巨響(Na) 瞬間(Nd) 停放(VC) 現場(Nc) 機車(Na) 燒毀(VC) 火舌(Na) 順(VJ) 樓梯間(Nc) 開始(VL) 漫延(VH) 當時(Nd) 民眾(Na) 熟睡(VA) 時刻(Na) 來不及(D) 逃出(VCL) 消防隊(Na) 基隆(Nc) 救難大隊(Na) 隨即(D) 衝入(VCL) 火場(Nc) 受到(VJ) 嗆傷(VH) 居民(Na) 救出(VC) 送醫(VB) 急救(VB) 中年(Na) 男子(Na) 傷勢(Na) 嚴重(VH) 送醫(VB) 途中(Nc) 不幸(VH) 身亡(VH) 警方(Na) 調查(VE) 發現(VE) 造成(VK) 意外(Na) 不排除(VC) 人為(VG) 縱火(VA) 可能(VH) 鄰居(Na) 看到(VE) 案發(VH) 當時(Nd) 燃燒(VC) 金紙(Na) 到底(D) 如何(D) 起火(VH) 警方(Na) 會同(VC) 消防(A) 鑑識(VC) 人員(Na) 調查(VE) 詳細(VH) 原因(Na)

圖十二 DPA 過濾後的結果

經由 *FIA* 所推論得到的 Dynamic sub-Ontology，如圖十三所示。



圖十三 Dynamic sub-Ontology

圖十四為 Dynamic sub-Ontology 轉換為 DAML+OIL 語法後之部分結果。

```

<?xml version='1.0' encoding='Big5'?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
  xmlns:xsd="http://www.w3.org/2000/10/XMLSchema#"
  xmlns:a="http://protege.stanford.edu/system#">

<daml:Ontology rdf:about="">
</daml:Ontology>
<daml:Ontology rdf:about="">
<rdfs:comment>社會犯罪 Ontology</rdfs:comment>
<daml:imports rdf:resource="http://www.daml.org/2001/03/daml+oil" />
</daml:Ontology>

<!-- Classes -->
<daml:Class rdf:ID="社會犯罪">
</daml:Class>
<daml:Class rdf:ID="縱火">
<rdfs:comment>案件發生為社會犯罪的一種事件</rdfs:comment>
<rdfs:subClassOf rdf:resource="#社會犯罪" />
</daml:Class>

<daml:Class rdf:about="#現場">
  <rdfs:label>現場</rdfs:label>

```

```

</daml:Class>
<daml:Class rdf:ID="地點">
  <daml:sameClassAs rdf:resource="#現場"/>
</daml:Class>

<rdf:DatatypeProperty rdf:ID="市" a:range="symbol">
  <rdfs:domain rdf:resource="#現場"/>
  <rdfs:range rdf:resource="Literal"/>
  <a:allowedValues>台南市</a:allowedValues>
  <a:allowedValues>基隆市</a:allowedValues>
</rdf:DatatypeProperty>
<rdf:DatatypeProperty rdf:ID="operation" a:range="symbol">
  <rdfs:domain rdf:resource="#現場"/>
  <rdfs:range rdf:resource="Literal"/>
  <a:allowedValues>位於</a:allowedValues>
</rdf:DatatypeProperty>

<daml:Class rdf:about="#時候">
  <rdfs:label>時候</rdfs:label>
</daml:Class>
<daml:Class rdf:ID="時間">
  <daml:sameClassAs rdf:resource="#時候"/>
</daml:Class>

<rdf:DatatypeProperty rdf:ID="日期" a:range="symbol">
  <rdfs:domain rdf:resource="#時候"/>
  <rdfs:range rdf:resource="Literal"/>
</rdf:DatatypeProperty>

<rdf:DatatypeProperty rdf:ID="期間" a:range="symbol">
  <rdfs:domain rdf:resource="#時候"/>
  <rdfs:range rdf:resource="Literal"/>
  <a:allowedValues>昨天</a:allowedValues>
  <a:allowedValues>前天</a:allowedValues>
  <a:allowedValues>今天</a:allowedValues>
</rdf:DatatypeProperty>

<rdf:DatatypeProperty rdf:ID="時段" a:range="symbol">
  <rdfs:domain rdf:resource="#時候"/>
  <rdfs:range rdf:resource="Literal"/>
  <a:allowedValues>凌晨</a:allowedValues>
  <a:allowedValues>清晨</a:allowedValues>
</rdf:DatatypeProperty>

<rdf:DatatypeProperty rdf:ID="operation" a:range="symbol">
  <rdfs:domain rdf:resource="#時候"/>
  <rdfs:range rdf:resource="Literal"/>
  <a:allowedValues>在</a:allowedValues>
  <a:allowedValues>於</a:allowedValues>
</rdf:DatatypeProperty>
...
</rdf:RDF>

```

圖十四 Dynamic sub-Ontology 之 DAML+OIL 格式

最後，預期摘要結果如圖十五所示。

於基隆市今天凌晨傳出火警，消防隊員灌救火場民宅，受傷人數 28 喪生人數 1，警方調查疑似人為縱火。

圖十五 預期摘要結果

圖十六和圖十七分別是第二個例子的原始新聞 XML 檔案格式以及摘要結果。

```
<?xml version="1.0" encoding="BIG5" ?>
<新聞>
<抬頭>尾隨可疑小客車檢方查獲海洛因</抬頭>
<新聞類別>社會萬象</新聞類別>
<日期>04/15</日期>
<記者>曾恩仁</記者>
<地點>屏東報導</地點>
<內容>
屏東地檢署根據線報，在高雄市查獲 1 部小自客車涉嫌載運毒品，前後搜出 570 多公克海洛因磚。由於幕後疑有販毒集團，因而蘇姓嫌犯在接受訊問後，於 14 日下午被聲請羈押獲准。屏東地檢署林姓檢察官因接獲線報，指 36 歲的蘇姓男子可能藏有毒品，於是指揮保三總隊員警佈線追查。13 日下午尾隨蘇某至高雄市苓雅區後，見時機成熟，即上前攔檢蘇某。警方在蘇某的小自客車內，搜出半塊海洛因磚和 1 台電子秤，乃懷疑蘇某涉嫌販賣。隨後趕往大寮鄉 1 處倉庫搜索，又搜出 1 塊海洛因磚、電子秤、模具及分裝工具。檢警在估算後，發現海洛因總重量 570 多公克。檢察官在訊問蘇某時，由於蘇某不肯吐露內情，遂於 14 日中午 12 時聲請羈押禁見，並於下午 4 時經法官裁定。據悉，檢方懷疑幕後還有販毒集團，為了擴大偵辦，才以串證理由聲請羈押蘇某。
</內容>
</新聞>
```

圖十六 原始電子新聞文件之 XML 格式

地檢署根據線報，查獲毒品，搜出海洛因。檢方懷疑有販毒集團，聲請羈押嫌犯。

圖十七 預期摘要結果

陸 結論與建議

一、結論

在這篇論文，我們提出了四個代理人，包括有資訊擷取代理人 (RA,

Retrieval Agent)、文件前置處理代理人 (DPA, Document Pre-processing Agent)、模糊推論代理人 (FIA, Fuzzy Inference Agent) 以及摘要代理人 (SA, Summarization Agent) 來進行中文文件自動摘要。此外, 我們亦建置了一實驗性的中文文件自動摘要系統, 未來希望摘要方法能延伸適用於不同來源之多篇中文或英文摘要。最後, 在 Ontology 建構上, 期許發展出一個具有自我學習能力之機制。

二、後續研究之建議

1. 對於 Ontology 的建構往往需要大量的人力及時間去進行, 目前有些人工智慧 (Artificial Intelligence) 領域的專家致力於 Ontology 建構 (人工或自動方式) 之研究, 若能夠發展出 Ontology 自動建構機制, 便能節省下許多人力及時間的損耗, 以提昇 Ontology 的完整性及正確性。
2. 近年來有許多專家在探討 Ontology Learning 這個議題, 倘若能夠建立 Ontology 學習機制, 可使我們從大量的資訊中取得有用的知識, 並能加強其領域知識內容, 加強 Ontology 的完整性及實用性。
3. 本論文所提出的模糊推論機制中的模糊歸屬函數及推論法則庫, 均是預先由領域專家憑藉經驗法則或知識所決定, 建議未來可將學習機制加入模糊系統, 例如採用基因演算法 (Genetic Algorithm) 或是類神經網路 (Neural Network) 來調整模糊法則庫和模糊集合之形狀, 以求得這些參數的最佳解, 進而改善系統效率。

誌謝：

本論文為行政院國家科學委員會補助之專題研究計畫 NSC91-2815-C-309-006-E 之部分成果、以及財團法人資訊工業策進會九十一學年度分包學術機構研究計畫 (91) PEO-0616 之部分成果, 謹此誌謝。同時亦感謝參與本專案之其他成員：蕭文賢、李青芸、鄭淑蓮與余書明。

參考文獻

孫宗瀛、楊英魁, 「Fuzzy 控制：理論、時作與應用」, 台北：全華科技圖書股份有限公司, 2001 年。

- 翁鴻加, 「多文件自動摘要一些新技術及評估模型之建立」, 國立台灣大學資訊工程研究所碩士論文, 1999 年。
- 曾新穆、李健興, 「支援語意空間的 Ontology 擷取與建構技術研究」期中技術報告, 資策會, 2002 年。
- 黃純敏、吳郁瑩, 「網路中文文件自動摘要」, 台灣區網際網路研討會 TANET'99, 國立中山大學承辦, 1999 年。
- 葉鎮源, 「文件自動化摘要方法之研究及其在中文文件的應用」, 國立交通大學資訊科學研究所碩士論文, 2002 年。
- Asunción G. P. and Corcho O., "Ontology Languages for the Semantic Web", *IEEE Intelligence Systems*, Jan./Feb. 2002, pp.54-60.
- Chandrasekaran B., Josephson J. R. and Benjamins V.R., "What Are Ontologies, and Why Do We Need Them?", *IEEE Intelligent Systems*, Jan./Feb, 1999, pp.20-26.
- Lee C. S., Liao J. X. and Kuo Y. H., "A Semantic-based Concept Clustering Mechanism for Chinese News Ontology Construction", *International Computer Symposium*, Taiwan, 2002.
- Lin C. T. and Gerorge Lee C. S., "Neural-Network-Based fuzzy Logic Control and Decision system", *IEEE Transactions on Computers*, Vol.40, No.12, Dec. 1991, pp.1320-1336.
- CKIP AutoTag, <http://godel.iis.sinica.edu.tw/CKIP/>. Chinese Knowledge Information Processing Group, Academic sinica.
- Lee C. S. and Pan C. Y., "An Intelligent Fuzzy Meeting Agent for Decision Support System", *IEEE International Conference on Fuzzy Systems*, USA. 2003.
- Lee C. S., Chen C. P., Chen H. J. and Kuo Y. H., "A Fuzzy Classification Agent for Personal e-News Service", *International Journal of Fuzzy Systems*, Vol.4, No.4, Dec. 2002, pp.849-856.
- Lee C. S., Pan C. Y. and Chang M. J., "A Fuzzy Decision Agent for Meeting Scheduling Supported System", *International Conference on Fuzzy Systems and Knowledge Discovery*, Singapore, 2002.
- Lee C. S., Kuo Y. H. and Yu P. T., "Weighted Fuzzy Mean Filters for Image Processing", *Fuzzy Sets and Systems*, Vol.89, No.2, Jul. 1997, pp.157-180.
- Connolly D., Harmelen F. V., Horrocks I., McGuinness D. L., Patel-Schneider P. F., Stein L. A., <http://www.w3.org/TR/daml+oil-reference>, 2001.
- Fensel D., "The semantic Web and its languages", *IEEE Intelligence Systems*, Nov./Dec. 2000, pp.67-73.
- Hovy E. and Lin C.Y., "Automated Text Summarization in SUMMARIST", In I.Mani and M.Maybury(eds), *Advances in Automated Text Summarization*, MIT Press, 1999, pp.81-94,.
- Embley, D. W., Campbell, D. M., Smith, R. D., & Little, S. W., "Ontology-based Extraction and Structuring of Information from Data-rich Unstructured Documents", *Proc. of ACM Conference on Information and Knowledge Management*, USA, 1998, pp.52-59.

- Guarino, M., Masolo, C., & Vetere, G., "OntoSeek: Content-based Access to the Web", *IEEE Intelligence Systems*, 1999, pp.70-80.
- Chen H. and Yen J., "Toward Intelligent Meeting Agents", *IEEE Computer*, 1996, pp.62-70.
- Mani I., "Recent Developments in Text Summarization", Proc. of. CIKM'01, Georgia, Nov. 2001, pp.529-531.
- Ferber J., "Multi-Agent System", ADDISON-WESLEY, New York, 1999.
- Kupiec J., Pedersen J. and Chen F., "A Trainable Document Summarizer", In SIGIR, ACM, Seattle WA, USA, 1995.
- Mani and Maybury M., "Introduction" In I. Mani and M. Maybury(eds), *Advances in Automated Text Summarization*, MIT Press, 1999, pp.x-xv.
- Huhns M. N., Singh M. P., "Multiagent System for Workflow", *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol.8, John Wiley & Sons, Ltd., 1999, pp. 105-117.
- Noy N. F., McGuinness D. L., "Ontology Development 101: A guide to Creating Your First Ontology", Stanford University, 2001.
- Maes P., Guttman R. H. and Moukas A. G., "Agents That Buy and Sell", *Communications of the ACM*, Vol.42, 1999, pp.81-91.
- Karp P. D., Chaudhri V. K. and Thomere J., XOL Ontology Exchange Language <http://www.ai.sri.com/pkarp/xol/xol.html>, 1999.
- Karp P. D., Chaudhri V. K. and Thomere J., "XOL: An XML-based Ontology Exchange Language(version 0.4) ", Aug. 1999.
- Lam W. and Ho K. S., "FIDS: An Intelligent Financial Web News Articles Digest System", *IEEE Trans. on SMC-part A*, Vol.31, No.6, Nov. 2001.
- Zeng X. M. and Lee C. S., "A Study on Automatic Classification Technique for Chinese Document", *Technology Report of Institute of Information Industry*, Taiwan, Aug. 2001.

A Study on Automatic Abstraction Technology for Chinese News Documents based on Domain Ontology

**CHANG-SHING LEE, YEA-JUAN CHEN, YEA-CHI KUO, AND
HUNG-YI CHUANG**

Department of Information Management, Chang Jung University

ABSTRACT

Most automatic document abstraction technologies utilize statistic method to excerpt the document, they calculate the weight of the sentence to gather important sentence. In this paper, we propose a novel automatic abstraction technology for Chinese News documents based on domain ontology. The Retrieval Agent (RA) will retrieve the e-News from Internet periodically and send them to Document Processing Agent (DPA) for natural language processing. Furthermore, the Fuzzy Inference Agent (FIA) will compute the similarity of ontology concepts and retrieved News terms. Finally, the Summarization Agent (SA) will abstract the News documents based on the extracted event ontology.

Keywords: automatic abstraction technology, ontology, fuzzy inference, agent