科技部補助專題研究計畫成果報告
期末報告

提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式(I)

計 畫 類 別 ： 個別型計畫
計 畫 編 號 ： MOST 105-2221-E-343-003-
執 行 期 間 ： 105年08月01日至106年07月31日
執 行 單 位 ： 南華大學資訊管理學系

計 畫 主 持 人： 楊士霆

計畫參與人員： 大專生-兼任助理：廖偉哲
大專生-兼任助理：張育嘉
大專生-兼任助理：張婉瑄
大專生-兼任助理：蔡旻晉
大專生-兼任助理：徐櫻綺
大專生-兼任助理：廖瑜哲
大專生-兼任助理：謝函恩

報 告 附 件 ： 出席國際學術會議心得報告

中 華 民 國 106 年 10 月 30 日

中 文 摘 要 ： 現今虛擬論壇自由且便利之發言平台，透過發言規範與論壇管理員之審核，論壇使用者即可輕易地發表各項文章，然而，因使用者逐漸增加下，多數虛擬論壇開始設有管理者用以審核違規之文章，但因大量文章注入虛擬論壇中，導致論壇管理者難以逐一審核並回請修正所有之文章，且對於發文者而言，可能於無法意識中寫入之違規字詞，於此，文章撰寫者發文後將隨即觸犯論壇規範，並需再針對違規文章進行修改，使得發文者再發文之意願逐漸低下。有鑑於上述問題，本研究乃發展「提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式」，並劃分為「文章表達情緒判定模組」與「發文者文章語句重組模組」兩核心模組，以論壇文章為分析之基礎，發展一套適用於論壇之方法論。於前者中乃以文章代表事件分析、情緒詞彙隸屬係數建立、相似語句分析、情緒機率與穩定值解析等技術，以推論文章之情緒類別；於後者中，本研究則以整合語意相似分析、評閱分數分析、候選填充句建立、以及多重組合句建立等技術，於最終將可重組文章之語句架構。承接於本研究發展之模式，為確認本方法論於實務應用中之可行性，本研究乃建構一套以網際網路為基礎之提升論壇知識利用價值之論壇文章情感解析及語句結構重組系統。此外，為驗證本系統績效，本研究以論壇文章真實資料作為驗證資料之樣本，針對「文章表達情緒判定」與「發文者文章語句重組」進行獨立驗證，確保兩相議題間之驗證結果不相互影響。另一方面，本研究亦以案例為導向探討整合兩核心模組，所建構之模式與系統，於實務情境中之應用與管理意涵。

中 文 關 鍵 詞 ： 虛擬論壇、語句相似度分析、情緒解析、文章語句重組

英 文 摘 要 ： At present, the virtual forums are free and convenient speaking platforms, according to the speaking specifications and through the examination of forum administrators, the forum users can publish various articles easily. However, as the number of users increases, most of virtual forums have administrators to examine the violative articles, but there are so many articles imported into the virtual forums, it is difficult for the forum administrators to check them one by one and return all the articles to be revised. On the other hand, the publishers may write violative words unconsciously, so the article writers infringe the forum specifications with the publication, and shall revise the violative articles, so that the publishers' willingness to publish articles declines gradually. For the above problems, this paper develops "A Forum Article Sentiment Analysis and Statement Structure Reorganization Model for Increasing the Forum Knowledge Utilization Value", divided into two kernel modules, which are "Article Expressed Emotion Determination Module" and "Publisher's Article Statement Reorganization Module". A methodology applicable to forums is developed by using the forum articles as the basis of analysis. The former one deduces the emotion type of articles by

analyzing representative events of articles, creating emotion word membership coefficient, analyzing similar statements and analyzing emotion probability and stable value. In the latter one, this study uses integrated semantic similarity analysis, review score analysis, candidate filler sentence establishment and multiple combined sentences, the article's statement structure can be reorganized. In order to confirm the feasibility of this methodology in practical application, this paper builds a Web-based system. Furthermore, a real-world case is applied to evaluate the proposed model. To sum up, this paper analyzes the articles which may contain extreme words and reorganizes statements for this type of articles. For the forum administrators, the violative articles can be extracted rapidly from the articles with specific emotions and the violative statements will be reorganized. For the article writers, the statements of violative articles can be revised automatically, so as to save the time and manpower for revising violative articles, and the probability of violation can be known immediately.

英文關鍵詞：Virtual Forum, Sentences Similarity, Emotion Determination, Article Statement Reorganization

一、報告內容

1. 研究動機與目的

過去論壇尚未風行之時，使用者多數僅能與相關熟悉人士討論事物之看法，但現今論壇自由且便利之發言平台，已成為人們發表自身觀點與看法之主要媒介（Hsu 與 Lin，2007；Li 等人，2012），如「Mobile01」、「巴哈姆特」及「伊利討論區」等，透過發言規範與論壇管理員之審核，即可輕易地發表各項文章（Guido 與 Leendert，2006）。然而，因論壇使用者逐漸增加下，文章撰寫者所發表之文章篇幅與內容大多不盡相同，論壇管理者必須依據發言規範與內容之詞語用法審視所有之文章，方可從中過濾違規之文章，以維持論壇之文章品質（Alavi 與 Leidner，2001）。

針對上述，雖多數論壇設有管理者用以審核違規之文章，但因大量文章注入論壇中，導致論壇管理者難以逐一審核所有之文章（Gu 與 Grossman，2010）。此外，對於特定事物具有特別看法與觀點之文章撰寫者而言，可能因個人疏忽之關係，而於無意識中將帶有偏激或批評之詞語寫入文章內，因而違規而遭致論壇平台或管理者移除，如以圖 1 文章為例，文章標題為「*HTC Desire HD VS SAMSUNG GALAXY S I9000 選擇性*」，文章內容有「*此次楊 X 君事件，讓我看到媒體炒作及操控的功力不凡：閃躲、不敢去碰飛彈瞄準對著台灣，...，可以任意捏、踹？這個好用的韓國出氣包，雖然前科不少，但好像也只限於運動賽事...等*」，該篇文章撰寫者原先主要乃比較 HTC 與 Samsung 兩手機廠牌之選擇性（如圖 1 色框處），但因 Samsung 廠商所屬之國家為韓國，該篇文章撰寫者乃將此篇內容引導至韓國與大陸對台灣國際上行為之看法，其中包含「*一直想者終極統一的是誰*」、「*對我們台灣始終充滿著敵意*」等數個大陸欲統一台灣之獨特看法，且文章亦有提及「*雖然前科不少，但好像也只限於運動賽事*」等韓國於國際上對待台灣行為之觀點（如圖 2 色框處所示），然因文章中具有數個惡意稱呼代替或影射詞以及用字未正名等規定（如圖 3 色框處，包含「大流氓」、「大榴槤」乃影射中國、「軟柿子」影射南韓、「陳 x 扁」未正名等規定），因而違反論壇之發言規範而遭刪除。

圖 1、範例文章

圖 2、文章撰寫者所提之觀點

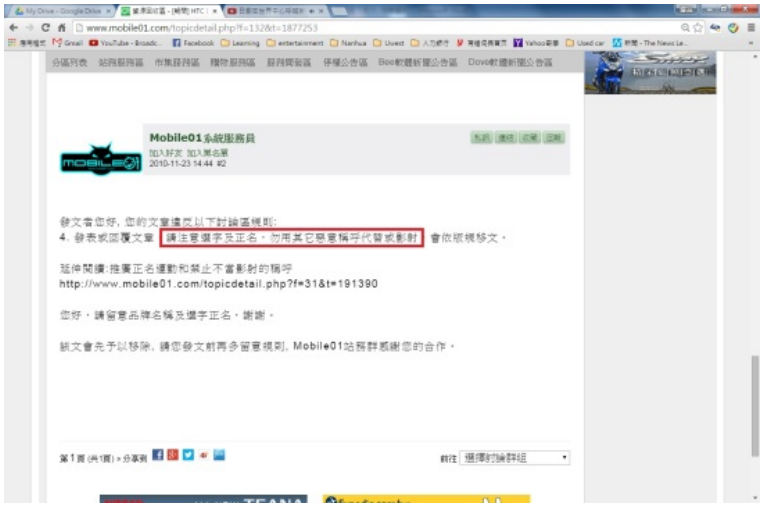

圖 1.3、文章所違反之規範

根據上述內容，論壇管理者僅能遵循論壇之規範將所有違規之文章移除，甚者，論壇之文章審核機制可以現有之文字比對方法，將出現違規字詞之文章直接移除（Ichifuji，2010），雖可藉由管理者

以自身觀點逐一閱讀所有違規文章，從中獲取具有獨特觀點之文章，並自行或再回請文章撰寫者修正，但此舉將花費大量人力與時間，以致於執行之效率不佳。綜合上述，其既有之運作模式如圖 4 之 AS-IS Model 所示。本研究之研究動機與目的可歸納為以下兩點：

1. 文章撰寫者欲表達自身想法時，可能因個人之疏忽而於無意識中寫入不當或抨擊之詞語，因而違反論壇之發言規範。
2. 論壇管理者較難以針對文章內容逐一修改不當之詞語用法，僅能遵循論壇之規範移除違規之文章，以致於論壇失去大量有討論意義之文章。

有鑑於此，為能協助論壇管理者從違規之文章中，獲取具有討論價值之文章，以保留有意義之文章，本研究乃建構「提升論壇知識利用價值之論壇文章情感解析及語句結構重組」模式，從中分析帶有偏激詞語之文章，並針對此類型之文章重組語句內容，以避免違反論壇之發言規範而遭移除。本研究之期望運作模式如圖 5 所示，並將本研究之重點歸納為以下兩點：

圖 4、論壇管理機制審核文章之既有模式

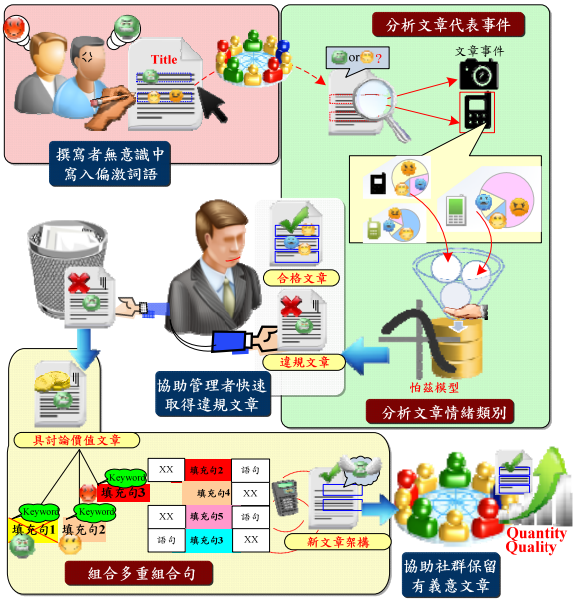

圖 5、論壇管理機制審核文章之期望模式

1. 分析文章撰寫者之情緒類別

文章撰寫者對於特定事物欲表達看法時，常會以不同之情緒字詞襯托以加強語句之語氣，但當文章撰寫者用以偏激之情緒字詞時，即容易觸犯論壇之規範，是故，為了取得文章撰寫者之情緒類別，本研究乃藉由文章之代表語句並分析語句之相似程度，再藉由情緒分析技術分析所有相似語句，以取得文章之情緒類別。

2. 重組具偏激情緒與字詞語句之文章

當得知文章撰寫者帶有惡意或偏激之情緒時，本研究針對此類型之文章分析內容之評級分數，對於評級分數較低之文章，即表示內容之順暢性與連結性不佳，因此，本研究乃分析文章內較為重要之語句，透過語句之多重組合句重組文章之內容，以避免文章內之用詞過於偏激而違反論壇規範，進而協助論壇管理者保留有意義且未違規之文章。

整體而言，為了協助論壇管理者保留具有討論意義且未違反論壇規範之文章，本研究乃分析文章之情緒類別，藉此分析文章帶有惡意或偏激情緒之可能性，再分析文章內容之評閱分數，以了解語句連貫程度，並以重要語句之組合句重組文章之內容，因此，本研究之目的即協助論壇管理者快速取得違規之文章，並針對違規但具有獨特看法之文章（可增進論壇使用者持續討論之文章）重組其內容，進而保留有意義之文章。

2. 文獻回顧

針對上述研究動機與目的，於探討相關文獻前先行釐清本研究之研究定位，以瞭解本研究與相關研究之差異性及本研究研究價值。

2.1 虛擬社群之管理機制探討

對於社群之管理機制探討議題而言，本研究乃針對「影響社群文章共享因素」及「社群文章審核機制」進行相關文獻探討，期望於其中觀察此議題應用於不同類型之不同角度與層面解析，以更深層

探討社群之管理機制之特性。

(I)影響社群文章共享因素

於影響社群文章共享因素中，本研究乃針對「管理技術層面」及「行為理論層面」等兩主題進行文獻探討，期望從中探討影響社群文章共享因素所涉及之範圍與領域。

(A) 管理技術層面

針對管理技術層面部分，使用者於虛擬社群中搜尋知識或服務時，其搜尋字串常與專業詞彙之語意無關聯，導致搜尋結果不彰，是故，Peng 等人（2008）乃提出一個以相似本體概念為基礎之語句解析方法，以分析搜尋語句與領域知識之關聯性。該研究乃結合 Web 服務之本體語言（Ontology Web Language for Services；OWL-S）與統一描述、發現語與集成（Universal Description, Discovery, and Integration；UDDI）之概念建立一個匹配分析法。甚者，隨著網路之發展使得虛擬社群知識庫之知識量逐漸增加，但大部分之知識乃無經鑑定與審核，導致知識庫之知識容易出現不真實、矛盾之情況，因此，為了解決上述之問題，Kaila 等人（2006）提供含三項功能之知識分享平台以過濾知識庫之知識，其中包含：(1)該平台可自動擷取外部知識庫之知識，將其轉化為內部知識庫真正所需之知識、(2)並利用內外知識語意之關係設計一個邏輯平台，將被錯誤知識所誤導之正確知識尋回、以及(3)提供知識擷取者正確之知識，亦可確保正確之知識不被淘汰。另一方面，知識共享之行為乃虛擬社群成功發展關鍵之一，為了提升使用者共享知識之意願，Wang 等人（2008）乃以 Gnutella 點對點網路為架構設計一套知識分享系統，將可協助知識擷取者取得所需之知識，並可同時藉由即時通訊功能與該知識之分享者進行直接溝通，亦即將同一領域之知識分享者歸於一類以促進知識之共享。

(B) 行為理論層面

針對行為理論層面之課題，虛擬社群之存在價值在於成員之知識共享，因此，為了瞭解社群成員共享知識意願之因素，Zhang（2009）乃將過去影響知識共享因素之相關研究進行彙整，以協助學者與社群管理者了解影響社群共享知識之因素。甚者，Fang 等人（2010）乃透過 143 位 IT 論壇之成員進行問卷調查，並於問卷中針對社群回饋成員之公平性（如激勵機制）、成員共享知識之方式（如知識分享平台之運作方式）及成員間、成員對社群之信任、知識可信度之分析，以瞭解成員使用社群之意圖及回饋機制對成員持續使用社群之成效，其分析結果可得知社群之回饋機制與社群成員對論壇之信任兩因素，主要乃影響成員持續與參與共享知識之因素。

此外，論壇文章所表達之情緒常反映使用者對於特定對象之評價，因此，若能即時於論壇之熱門主題中分析文章之情緒類別，將可提供決策者改變產品或服務之策略，有鑑於此，Li 及 Wu（2010）以新浪（Sina）論壇作為分析對象，先以人工標記主題之名稱、文章及網頁 URL，以形成主題之樹狀分類結構圖並作為訓練資料，之後將各主題之文章進行分詞處理，以利用文章之關鍵詞與知網（HowNet）所建構之情緒詞進行字詞比對，以取得文章關鍵詞之情緒值（正、負向情緒），再將所有關鍵詞之情緒值予以加總藉以得知文章之情緒類別。除此之外，Wang 與 Noe（2010）整合知識分享相關之文獻，統整並建立一個知識分享模式，該模式將知識分享分為五個主要重點領域，分別為組織環境、人際關係與團隊特徵、文化特徵、個人特徵以及激勵因素。

(II)社群文章審核機制

於社群文章審核機制中，過去研究乃針對「以管理者進行審核」、「以審核系統進行審核」等兩部分中，探討文章審核機制所涉及之範圍與領域。

(A) 以管理者進行審核

針對以管理者進行審核之課題，目前虛擬社群已逐漸有知識過量之問題，導致社群管理者難以監管與分類知識，是故，Gu 與 Grossman（2010）針對知識交流量較高之區域網路，架設一個知識分享平台，以協助管理者管理過量之知識，亦可協助知識擷取者於過量之知識中取得正確之知識量。此外，Matsubara 等人（2006）提出一套知識分享系統，以增設虛擬目錄並適當地控制其虛擬目錄之知識，可有效降低擷取知識檔案時所隱藏之危險性（盜版之檔案），亦可方便管理者審核知識安全性之問題。

另一方面，過去關鍵字擷取技術中，多數須以訓練資料之方式建構關鍵字庫，且較少研究以詞彙鏈之方式提取文章關鍵詞，故 Ercan 及 Cicekli（2007）建構一個以基線系統（Basedline System）為基之關鍵詞擷取模式。甚之，使用者於論壇中常需花費大量時間尋找知識，且當取得與需求相關之知識時，其延伸之話題常與原主題不符（即其他使用者對於主題之回覆訊息），導致使用者難以從中取得所需之知識，有鑑於上述問題，Li 等人（2012）提出一個基於馬爾可夫鏈之社交網路（Social

Network-Based Markov Chain；SNMC）系統，以推薦合適之知識專家並透過交流取得相關之延伸話題。

(B) 以審核系統進行審核

於以審核系統進行審核之課題中，甚者，為了解決問答系統推論答案時，常因詞彙具有同義或多義之情況而導致擷取錯誤答案或忽略正確答案，因此，Yu 等人（2006）提出一個以潛在語意分析（Latent Semantic Analysis；LSA）為基之問答相似句演算法。此外，「屬性詞」乃近代所衍伸之新興詞類，大多可用以描述事件之屬性或層級，甚者，可修飾動詞之表現方式與性質，因此，網路評論中所存在之屬性詞可視為評論者對於產品之觀點，從特定領域之網路評論中取得屬性詞（Li 等人（2011））。

知識管理乃一個能使知識更有效地被捕捉、儲存、使用與傳播之技術，但知識之變化、更新快速，甚者，陳述性知識與結構型知識等類型之不同，以致於知識管理者難以控管不同型態之知識，是故，為了能更進一步地提升知識管理之效率與效能，Lai（2007）提出一套以知識管理為主之知識工程理論模式（Knowledge Management through Knowledge Engineering；KMKE），以將所有知識轉換為相似型態並將知識分類與分層儲存於系統中，以協助知識管理者管理龐大且變動頻繁之知識。其中，使用者擷取資料之來源數眾多且常涉及龐大之資料紀錄，以致於尋找並轉換符合使用者需求之資料格式時須花費大量時間，因此，Pérez 等人（2007）以資料探勘網格結構（Data Mining Grid Architecture；DMGA）為基礎設計一套系統，並利用 Apriori 關聯規則演算法為基礎發展一個網格運算法。

2.2 社群文章資料探勘

針對社群文章資料探勘之課題，本研究乃針對「社群文章語句重組技術」以及「社群文章評級分析」等兩議題進行相關文獻探討，以更深層瞭解文章語句重組之方法，並從中探討評級文章之分類方法與技術。

(I)社群文章語句重組技術

於社群文章語句重組技術議題中，本研究乃針對「文章重點區塊分析」、「文章摘要建立技術」以及「文章語句合適度分析」等三主題進行相關文獻探討，期望從中探討語句重組與改寫之方法。

(A) 文章重點區塊分析

由於網路商品之評價更新快速且大多皆有正反兩面之評價，為協助網路使用者可快速得知商品最新之評價及情感傾向性，Pang（2010）提出一個基於主題規則之評價分析方法，以有效快速取得商品之最新評價，並同時得知商品之正反評價以供一個網路使用者購買商品之參考依據。此外，使用者對網路商品評論之情感傾向及特徵大多可視為商品及服務品質之依據，為協助網路服務商取得使用者對產品之評論重點，Li 等人（2008）乃先行標記評論之語句詞性，並以語句為單位利用 LingPipe 自然語言分析語句之情感傾向，再以 Aprori 關聯規則演算法建構商品之特徵（如外型、功能）屬性並排序商品特徵詞彙之頻率，以利用具有商品特徵之語句依頻率權重比例推論評論之情感傾向性，可幫助網路賣家取得使用者對商品之評價及關注特徵，亦可提供網路賣家改進產品或提高服務之決策依據。

另一方面，為了能從文字數較長之網頁文件中取得重要語句，並將語句組合為該文件之主旨，Chan（2006）提出一個擷取最具代表性語句之量化模式，以協助使用者有效率地閱讀並節省過濾所有文件之時間。甚之，為了能從非結構化之產品評論文章中取得評論之情感類別，以協助使用者了解產品之特性（Fan 等人（2009））。

(B) 文章摘要建立技術

虛擬社群常限制使用者須以制定格式發表文章，但考量非所有文章皆以特定格式（如標題與語句位置）撰寫於此影響多數文件摘要技術之形成，因此，Ko 與 Seo（2008）提出一個混合式統計連續虛擬語句之摘要形成模式，以協助多文章與非結構化文章中取得代表性語句以形成摘要。而針對自動摘要技術常忽略文章次主題之問題，其形成之摘要將可能不完整。是故，Xu 等人（2008）提出一個局部主題（段落）為基之關鍵句擷取方法以形成較完整之摘要，該研究藉由局部主題之關鍵句所形成之局部摘要，可有效擷取次主題之重要語句，進而形成較完整之文章摘要供使用者查閱。此外，有鑑於產品評論數過多，導致使用者須花費大量時間閱讀方可得知評論之重點，是故，Zhang 等人（2009）利用詞類標記器（Part-of-Speech Tagger）先標註訓練評論之詞性，取得各評論之重點語句與主旨，進而節省閱讀評論之時間。

(C) 文章語句合適度分析

醫療領域中具有大量之專有名詞、歧異詞及分歧句，導致知識需求者於網路中尋求資訊時，因搜尋字詞無法與該領域之特殊字詞串聯，以致搜尋之效果不彰， Win 及 Zhang（2006）提出一個以本體

語義為基之知識檢索方法，可協助知識尋求者於不熟悉之領域中，以簡短之搜尋字串中取得欲得知之領域知識。而維吾爾語具有延展性、時態變化及複雜之語法結構等特性，以致於過去研究甚少針對維吾爾語分析語句之情感傾向，因此，Huang 等人（2012）提出情感分析方法及調整規則，可有效協助使用者於語句結構複雜且存在轉折詞彙（否定詞）之維吾爾語中，分析語句之情感類別。

除此之外，問答系統所回應資訊之品質，多數與語料庫（詞彙、語句之組合）建置完整度成正向相關，然而，語料庫之建構多數須透過問與答之訓練資料串連語句間之關係，方可形成專業之語料庫，因此，Hao 等人（2007）乃提出一個基於問題類型導向之語料庫建立方法，以形成一個較完整之語料庫，進而協助使用者以語料庫為基礎給予需求者適當之回應內容。

(II)社群文章評級技術

於社群文章評級技術議題中，本研究乃針對「以前後文關係分析」、「以相異詞語分析法」以及「以語言文法特性分析」等三個主題進行文獻探討，期望從中瞭解文章評級議題所涉及之範圍與領域。

(A) 以前後文關係分析

為了比較自動文章輔助（Automatic Essay Assessor；AEA）及 K 鄰近分類法（K-nearest-neighbor based；K-NN）兩個文章語意分級方法之成效，Kakkonen 等人（2008）透過國文教師將兩方法之各項權重參數整合一致，並以教師文章作為訓練文章（即較無特殊字詞或模糊語意），以公平比較兩方法之分級成效。由於網路自由評論（影視、產品及新聞評論等）多樣之寫作風格、格式及詞彙語意定義，導致詞彙之情感傾向值僅適用於單一領域之評論中，為能同時建立自由評論中詞彙之情感傾向，Yin 及 Peng（2012）提出一個情感傾向性關係網路（Sentiment Orientation Relationship Network；SORN）概念，藉以分析詞彙於自由評論之情感傾向值，使用者可不須依靠情緒詞彙庫及人工標記情緒詞，即可自動於各領域中自動建構所有詞彙與情感詞彙之傾向值，以建立特定領域之情緒詞彙庫。

此外，由於網路訊息常以簡短、不分段及無標點符號之形成發佈，導致分析訊息常無法取得內文之重點與潛在之語意關係，使得分類之結果不準確，因此，Peng 等人（2007）建立一個語意內積空間分析模型（Semantic Inner Space Model），即可分類非結構化且簡短之網路訊息。另一方面，非英語系國家之使用者撰寫英文後，多數乃透過書本或人工批閱之方式，方可得知所有語句中文法之正確性，然而，此舉大多須花費大量時間及人力，因此，Lee 等人（2011）乃提出一個英文語句之文法分析機制，以幫助使用者立即判斷語句文法之正確性。

(B) 以相異詞語分析法

為了協助使用者於大量之作文中，自動取得與題目不符或離題之作文內容，Ge 及 Chen（2009）乃先將作文之數字、專有名詞等特徵詞歸類並過濾停用詞，以將作文內重要詞彙之詞根（不同時態或動作時詞彙之變化）還原，接著將所有作文之特徵詞及特徵詞之詞根以詞頻-逆向文件頻率法（Term Frequency-Inverse Document Frequency；TF-IDF）方法計算特徵之向量，以取得所有特徵之特徵權重值，進而利用向量空間模型之餘弦函數分析所有作文間之相似度，最後，該研究以階層式聚合分群法將相似度相近之作文分群於一類，並重新計算該群集內所有作文之相似度，以將同群集內之作文再分群直至剩餘一個群集為止，藉此形成一個聚類樹，以將未分類於任一群集內之作文內容視為離題之作文。透過該研究之方法，使用者可於不同領域之作文題目中，自動取得偏離題目之作文內容，並可有效節省人工審核及過濾各篇作文之時間。甚之，Zhang 等人（2011）提出一個核樹（Tree Kernel）函數為基之語句情感分類方法，以解決分類語句情感類別時較不精準之問題。另一方面，Jiang 等人（2011）乃以英文文法、型態等規則針對中國大學生撰寫之英文作文進行評級分類，以幫助審閱者節省人工批改之時間。

(C) 以語言文法特性分析

為了協助使用者於大量英文文章中，以客觀角度評閱文章之級分數，Huang 等人（2009）提出一個基於投票演算法之多分類器文章評分系統，以協助使用者針對中國人所撰寫之英文文章自動評閱分數。甚者，由於現有情緒詞彙庫所提供之情緒傾向值，可能隨著不同領域之語意及專有名詞之定義而有所改變，因而導致情緒詞彙庫不適用之問題，是故，Wang 等人（2011）乃以本體論為基提出一個情緒詞情緒傾向計算方法，以精確得知特定領域特徵詞之情緒傾向值。

除此之外，過去研究大多乃針對網路文章或大學生之寫作程度作為文章評級之標準，甚少研究根據國小學童之撰寫能力分析語句之評級分數，有鑑於此，Liao 等人（2012）乃提出一個語句評級方法以幫助使用者針對國小學生所撰寫之語句，建構一個合適之語句評級標準。

2.3 文章撰寫者行為探討

針對文章撰寫者行為探討之課題，本研究乃針對「文章撰寫者情感分類」、「文章撰寫者寫作習慣分析」以及「文章閱讀者閱讀感受分析」等三議題進行相關文獻探討，以更深層瞭解社群使用者撰寫與閱讀之習慣。

(I)文章撰寫者情感分類

針對文章撰寫者情感分類議題中，本研究乃針對「以資料探勘進行分類」、「以詞語相似度進行分類」以及「以語意關係進行分類」等三主題進行相關文獻探討，期望從中探討不同分類技術之成效與應用範疇。

(A) 以資料探勘進行分類

新聞評論所涉及之主題（人、事件及物等）甚為廣泛，以致於過去研究分類新聞評論之情緒類別時，會因主題之不同而有所誤差，有鑑於此，Yang 等人（2011）乃先將評論進行分詞處理以分析詞彙於新聞主題詞頻關係，進而取得語句與新聞主題之隸屬值。而為了從部落格、網路、論壇及新聞類型等新型文章中（即文句較簡短精簡之評論）取得使用者之情感傾向（針對積極及消極兩情感類別），Yang 等人（2010）利用網路最短覆蓋路徑演算法（Shortest Covering Path；SCP）針對多種類型之評論，可有效針對語句簡短且多種類型之評論進行情感分析，且該研究可隨著訓練評論量之增加而提高情感分類之準確率。此外，中日關係論壇之文章結構具有複雜且大量特徵詞詞性等特性，較無法以自然語言方法分析文章之情感類別，故 Wang 及 Li（2010）針對上述問題先以人工方式選取文章之指代（即文章之回應者乃承接他人之回應內容）、代字及縮寫詞三種結構，以依據文章特殊結構之重要區塊計算兩情感類別之傾向，以協助使用者於具有特殊詞彙且複雜結構之論壇文章中取得情感類別。且過去研究分析詞彙之褒義及貶義，多數乃藉由已標記褒貶義強度詞彙之語料庫（如 HowNet），直接分析詞彙與語料庫之詞彙關係以進行分類，甚少研究藉由詞彙之同義詞（語言學中最小粒度之層級）關係分析詞彙之情感傾向，是故，Wang 等人（2009）利用點式交互訊息法解析目標詞彙與所有詞彙之關係強度，藉以取得目標詞彙之同義詞集，進而於不同領域中建立詞彙之情感傾向表。

另一方面，過去研究多數以基於「語意規則」與「統計」兩方法分析文件之情緒類別，然而，基於規則方法可能因文件特徵之間相互矛盾（例如：文件主題為「天氣」，主要特徵可能為「冷」與「熱」兩矛盾特徵），使得特徵之結構錯誤，而基於統計方法則可能因情緒詞庫之不完整，以致於忽略比對文件中更為重要詞彙，有鑑於上述兩項分析方法之缺點可能導致情緒分類結果不精確，Zan 等人（2011）將此兩方法結合以互補優缺點，透過該研究之整合方法，可有效減少擷取錯誤語句特徵之發生，亦可取得文件中最為重要之語句特徵，進而提升分類文件情緒類別之準確率。

(B) 以詞語相似度進行分類

短句（或詞彙）之情緒傾向常隨著領域之變化而有所不同，亦即短句於不同領域中將有多種解釋及語意表達方式，導致短句可能於不同領域之情緒分類結果有所誤差，有鑑於此，Lin 等人（2011）提出一個以點式交互訊息法為基之情緒分類模式，可以非監督學習方式於多元領域中快速分析短句之情緒傾向。甚之，有鑑於情緒詞彙於不同領域之情緒傾向值（正負面之情緒）將有所差異，為協助使用者於多元領域中自動建立情緒詞彙之情緒傾向，Liu 等人（2009）藉由知網之相似演算法計算詞彙各種語意（即詞彙於知網中所延伸之語意）之語意距離，接著分析詞彙之語意距離與知網所含正負情緒詞之相關程度，並藉由正面與負面情緒之傾向值差異程度取得詞彙之情緒傾向值。

另一方面，情緒詞會隨著領域之不同而有其他解釋與定義，尤以虛擬社群所涉及之領域最為廣泛，使得使用者較無法於社群中同時分析文章語句之情緒類別，有鑑於上述問題，Zhan 等人（2011）首先乃標記文章之語氣轉折點（即文章內之轉折詞），以取得轉折點前後情緒詞彙之序列，並依據文章中程度副詞、否定詞對於情緒詞之影響程度，加權情緒詞之情緒傾向係數，以藉由轉折點前後程度副詞、否定詞，分析情緒詞對於語句之影響值，以累加計算語句之情緒傾向值，進而取得各語句之情緒類別。透過該研究之方法，使用者可根據各領域之特性自動加權情緒詞之情緒傾向值，藉此於多領域中同時分析語句之情緒類別。最後，由於現有孟加拉語之情感詞彙庫完整度不足問題，導致使用者難以取得網路文章之情感傾向，因此，Das 及 Bandyopadhyay（2010）利用部落格文章作為訓練資料，推論文章之情感類別，以幫助使用者於孟加拉語部落格中得知文章之情感類別。

(C) 以語意關係進行分類

語句之涵義常藉由上下文之串聯，方能得知語句欲表達之語意，但日本網路使用者常創造語意含

糊之語句或詞彙（日本將此些語句及詞彙稱為「Wakamono Kotoba」），以致無法藉由語句涵義分析語句之情感類別，為解決上述之問題，Matsumoto 等人（2011）將網路常使用且可形容他人情緒之詞彙或語句，先行歸屬為 Wakamono Kotoba 之語料庫，並以 IPA、UniDic 及 Naist 日語詞彙字典之語意架構過濾 Wakamono Kotoba 內之詞彙，以建立 Wakamono Kotoba 情緒詞彙庫，並透過貝氏及累積分類法進行情緒分類以獲取良好分類效果。甚者，為協助中文論壇管理者了解使用者表達意見時之寫作情緒，Zhang 等人（2009）乃分析不同領域之中文論壇（中國亞馬遜論壇、音樂論壇等）使用者之意見，於不同領域之論壇中自動建立語意規則，且具有良好之情緒分類成效。此外，Li 及 Liu（2012）提出一個基於聚類之情感分析方法，以改善過去須花費大量時間人工訓練文件資料，以及利用詞彙之情感偏向分數加總平均導致分類準確率不佳之問題。

(II)文章撰寫者寫作習慣分析

於文章撰寫者寫作習慣分析議題中，本研究乃針對「撰寫者寫作風格」以及「撰寫者用詞習慣」等兩主題進行相關文獻探討，期望從中探討文章撰寫者個人撰寫之風格與撰寫之習性。

(A) 撰寫者寫作風格

由於「微博」之評論具有多樣性、特有語意之風格等特性，常需以定義否定詞、連接詞等方式自訂語意規則，方能對評論計算情緒指數，然而，過去研究甚少以機器學習方法（Exploiting Machine Learning Methods）自動產生語意規則以針對評論進行情緒分類，是故，Liu 及 Liu（2012）乃以支援向量機、貝氏及 N 連詞三種分類方法，針對微博之電影評論進行情緒分類並比較三種分類方法之成效。而網路評論常以複雜（數字、字數不定）且非結構化之方式呈現，導致分類評論情緒之傾向時，難以建立語意規則以致須以人工方式建立語意關係，因此，Bai（2011）提出一個以馬可夫覆蓋（Markov Blanke Model）為基之情緒分類模式，以計算詞彙之情緒穩定值推論評論之情緒類別。

另一方面，由於網路中各虛擬社群皆具有獨特之文化風格，使得社群中各網頁文章之情緒表達方式不一，Fu 等人（2013）可同時分析不同文化社群中，主題欲表達之情緒類別。甚之，網路商品之多樣性使得評論之標準、格式不一，導致難以建立使用者對商品及評論之情緒特徵，為能協助使用者可於非結構化且大量之評論中自動建立評論之情緒特徵並分類情緒類別（Wang 及 Jiang（2012））。

在最終方面，因網路評論乃由使用者自由撰寫，使得評論之內容有不同格式及寫作風格，然而，並非所有評論內容皆在描述主題（包含新聞、商品等）之特性，Liu 等人（2010）為了能從評論中取得真實描述主題特性之內容，提出一個以文件物件模型樹（Document Object Model；DOM）為基之訊息擷取模式，協助使用者於不同網頁版面及寫作格式中，取得與主題較為相關之評論重點。

(B) 撰寫者用詞習慣

網路論壇中熱門主題所引起之熱烈討論不乏含有正負面之評論，其中，當評論含有「持續破壞」類型之評論時常引起網路筆戰（以情緒字眼攻擊他人）之發生，為從評論中取得此類型評論，Ichifuji 等人（2010）乃利用型態分析法（Morphological Analysis）先行分析蓄意破壞評論之文句主詞、動詞、受詞之組合字詞及雙字詞彙，以建構蓄意破壞詞庫，並先將評論分為一般評論、無意破壞及蓄意破壞評論等，再利用貝氏過濾器（Bayesian Filter）分析評論之雙字詞彙及詞性以取得蓄意破壞評論之評論，最後以一般評論之詞性、組合字詞及雙字詞等特徵比對蓄意破壞評論，藉此計算蓄意破壞評論可能為無意破壞評論之機率。除此之外，各領域之網路評論者皆有獨特之語意風格，導致分析評論之情感分類時常局限於單一領域中，有鑑於此，Li 等人（2008）利用超空間模擬語言（Hyperspace Analogue to Language；HAL）與訊息推理方法（Information Inference）整合一套多領域評論之情感分析模式，即可幫助使用者於各種領域之評論中推論情感傾向。甚者，各式虛擬社群所制訂之評論格式多數不盡相同，導致無法明確定義標題、主題及語句內容與情緒間之語意關係，以致無法同時分析各式評論之情緒傾向，為能幫助使用者針對不同格式之評論分析情緒傾向，Wang 等人（2010）提出一個以本體論（Ontology）為基之情緒分類方法，以取得該評論於特定領域之情緒傾向性，進而協助使用者於複雜且多樣之評論格式中取得情感傾向。

甚之，有鑑於中國新浪微博中訊息簡短且所發佈之格式不一致等特性，以致於分類訊息情緒傾向之成效不佳，是故，Xie等人（2012）乃提出三種文章情緒分類方法，使用者可有效於簡短且非結構化之訊息中取得情緒傾向，亦可依據訊息之格式於三種分類法中選擇分類成效最佳之方法。且目前烏爾都語言中尚無較完整且標註情緒傾向明確之詞彙庫，雖可利用英文情緒詞彙庫相互對照以對應情緒詞之情緒傾向值，但因烏爾都語中存有修飾詞與特殊形容詞之關係，導致對照後之成效不準確，為協

助使用者建立烏爾都語之情緒詞彙庫（Afraz等人（2010））。

最後，因網路產品評論所使用之語法及習慣用語多數不同，使得情緒詞彙庫多數僅適用於單一領域中，是故，為了協助使用者依據特定領域中評論用語之特性以建立情緒詞彙庫，Hamouda等人（2011）乃訂定收集訓練資料之規則以建構一套基於機器學習法之情緒詞彙庫（Machine Learning Based Senti-word Lexicon；MLBSL），進而幫助使用者取得特定領域之情緒詞彙庫，並可以此情緒詞彙庫為基分析評論之情緒類別。綜合上述，於此議題相關文獻之分析類型與過程分別彙整成表1.5。

(III)文章閱讀者閱讀感受分析

於文章閱讀者閱讀感受分析議題中，本研究乃針對「文章篇幅之於閱讀感受」以及「圖表之於閱讀感受」等兩主題進行相關文獻探討，期望從中探討文章閱讀者之閱讀習慣與感受。

(A) 文章篇幅之於閱讀感受

文章閱讀者閱讀網路文章時，常以首句詞語判斷主題與文章內容之相符程度，但文章閱讀者之生長環境與擅長領域不同，其解讀語句之涵義將與文章撰寫者原先表達之涵義有所差異，為了驗證此差異性之現象，David 等人（2011）針對 17 位測試者以實驗法方式測試使用者對於語句之認知。該研究結果顯示，不同受測者對於第一個語句之認知涵義將有所差異，但若制定規範或結構化之方式呈現文章時，受測者對於語句之認知差異性會有明顯之縮小。甚者，多文件摘要技術大多乃取目標文件之重要部分，再藉由其他文件之語句加以排序以形成摘要，然而，此方法常有語句不通順、不連貫或不符合事件（即兩連續語句所指之事件不同）之問題，是故，Danushka 等人（2012）乃針對新聞文件發佈時間定義語句之順序權重，接著比較新聞事件與相似事件之新聞文件，從中分析兩相似新聞中起頭、結論與過程等語句寫法，以給予目標新聞文件所有語句之順序機率值，其次，以動詞與專有名詞作為局部主題，將語句分類至對應之主題中，並根據主題與事件之關係及語句順序權重分析所有語句之優先順序，最後，利用語句排序演算法（Sentence Ordering Algorithm）計算所有語句與事件之關係函數，從中取得最合適之語句順序，以幫助使用者取得語句通順且與目標新聞相關之語句組合。

此外，目前電子新聞彙整機制乃以新聞標題與關鍵字方式呈現，但此方式缺乏具體主題以描述事件起始與重點等，待時間過去後閱讀者較難以追溯該新聞之事件。是故，Lin 及 Liang（2008）乃提出將事件主軸納入摘要建立條件之機制（Story-line based Topic Retrospection，SToRe），透過事件主軸能使讀者更了解事件之發展與概念。Zhou 等人（2010）提出一個機器學習之情感分類方法，可便於使用者於相異領域之新聞中，有效取得評論之情感類別，並提供使用者一個權重計算之參考建議。

(B) 圖表之於閱讀感受

使用者對於網路圖片之訊息來源，多數與圖像特徵（紋理、顏色）無相關之文字，亦即圖片之文字描述多數乃針對圖片中人、物之外觀或輪廓，是故，為了探討使用者以知覺描述圖片特徵之情況，Rorissa（2008）乃以圖像索引個人圖像及圖像檢索群體圖像為分析之主軸。藉由該研究之探討可了解使用者瀏覽介面圖片之描述習慣，進而提供決策者針對圖像之特性給予使用者不同之描述工具。甚之，Oh 等人（2008）乃探討網路店面呈現方式對於商品形象之影響，藉由該研究分析之結果，可提供網路賣家一個呈現網頁之參考方法，亦可透過圖片之呈現提升商品形象，以增加消費者購買意願。

且由於目前之圖像檢索技術仍未成熟，導致使用者對於欲搜尋圖片需求至今仍無法滿足，因此，Tractinsky 等人（2006）乃以兩項對比實驗並針對 40 位大學生（對於網頁設計較無概念之群體學生）進行實測，透過該研究之實驗結果，將可提供網站設計者一個設計概念，以針對特定群體之使用者呈現最合適之網站類型。最後，圖片之人、事、物等具體之意象將可提供知識需求者作為關鍵字搜尋之條件，然而現今大量圖片存於網路中，使得圖片之定義、註解及描述不完全，導致知識需求者利用圖片意象所建構之關鍵字進行搜索時之效率低下，有鑑於此，Liu 等人（2008）提出一項自動影像註解架構，以協助使用者欲檢索圖片且註記不清時，能提升檢索圖片之情況，並可有效取得所需之圖片。

2.4 小結

本研究之研究主題涉及「虛擬社群之管理機制探討」、「社群文章資料探勘」與「文章撰寫者行為探討」等三大研究方向。於2.1節「虛擬社群管理機制探討」議題中得知，過去研究主要著重於協助使用者快速取得正確之文章，同時透過文章審核機制過濾違規之文章，藉此管理及控管虛擬社群整體之品質。但過去之研究對於違規文章之審核多數僅能以文字比對技術，將出現違規字詞之文章移除，故本研究除針對違規字詞外，亦分析文章撰寫者撰寫時之情緒感受，藉此審核帶有偏激或惡意情緒之文

章，進而協助管理者於過量之文章取得違規之文章。於2.2節「社群文章資料探勘」之相關文獻中，可發現過去研究針對文章語句重組技術大多藉由語意規則及語句相似特徵重組語句之架構，而對於文章評級分析方面，過去多數研究乃透過相似詞彙比對並藉由文章分類法分析文章之評級類別，以分析文章語句之連貫程度。故本研究乃藉由相異詞語分析法及語句相似特徵之整合，建構「發文者文章語句重組模組」針對違規之文章重組文章之內容架構。最後於2.3節「文章撰寫者行為探討中」之相關文獻中，乃針對使用者撰寫習性與閱讀習慣進行探討，並從文章撰寫者情緒分類議題中，可得知以文章主題、語句、詞彙或語意關係可作為情緒之主要分析因素，故本研究乃以此為依據建構「文章表達情緒判定模組」以分析文章撰寫者之情緒類別，從中取得惡意情緒之文章。

綜上所述，本研究所建立「提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式」乃透過相似語句分析文章內容所含之情緒特徵，並透過文章之情緒特徵區分文章所屬情緒，此外，本研究亦針對帶有惡意情緒之文章，先以相異詞語分析文章之語句連貫程度，藉此分析文章重組之合適性，再藉由相似語句延伸之多重組合句重組文章之架構，最終將可協助論壇管理者快速取得違規之文章並改寫其內容，以協助社群保留具有討論價值之文章。

3. 提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式

本研究所提之「提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式」乃以虛擬論壇之論壇文章為分析對象，先行經中文詞知識小組（Chinese Knowledge Information Processing Group；CKIP）之中文斷詞系統斷詞後，從中取得論壇文章之候選事件，以利用候選事件之詞性、前後文及標題相關性分數分析論壇文章之代表事件，並藉由代表事件之語句比對訓練文章之語句相似程度，分析論壇文章之情緒類別，進而使後續發文者文章語句重組模組重組內文之架構後，可與原先文章撰寫者欲表達之情感一致。透過論壇文章評閱分數取得合適改寫之文章，並去除語意相似及情緒與原文不一致之語句，分析論壇文章所需之候選填充句與組合句，並進行多重組合以重組論壇文章架構，最終形成一篇語意完整之論壇文章內容。本研究之主要流程可分為兩大部份，分別為如圖6之Part1「文章表達情緒判定模組」與Part2「發文者文章語句重組模組」所示。

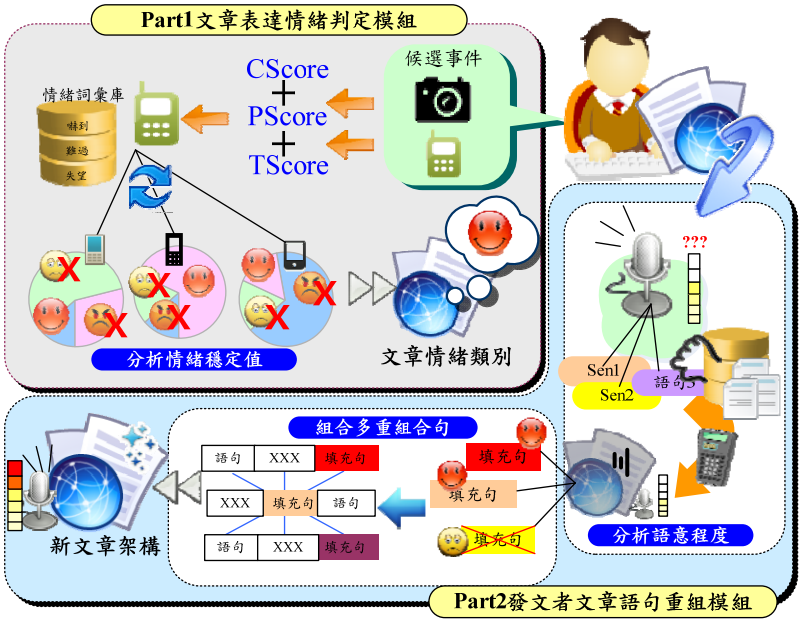

圖 6、提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式架構圖

3.1 文章表達情緒判定模組

根據本計畫中研究動機與目的之期望模式，於「文章表達情緒判定模組」之目的為自動推論文章之情緒類別，於最終推論情緒類別之步驟乃借鑒並改良 Zhao 等人（2010）之方法，該研究主要以非結構化之文章為分析對象，因此將可適用本模組中作為最終推論方法，而推論情緒類別方法時則須以文章之主題事件為基礎進行推論，因此本模組乃借鑑 Tung 與 Lu（2010）之方法規劃步驟(A1)至步驟(A3)中，以此三步驟取得文章之主題事件，此外，本研究情緒類別推論乃以情緒詞彙為基礎，並考量過去研究大多以人工方式定義情緒詞彙之隸屬係數，因此乃借鑒與改良 Miao 等人（2012）之方法規劃至步驟(A4)中，以於推論文章情緒類別前可先行自動推論情緒詞彙之隸屬係數，完成上述步驟後，即將 Zhao 等人（2010）改良後之方法規劃至步驟(A5)至步驟(A7)中。

本模組乃以論壇發文者文章為分析之基礎，以判定論壇文章所表達之情緒類別，進而作為後續發文者文章語句重組模組替換論壇文章情緒字詞之用。首先乃藉由「知網」(Hownet）所含 2090 個正負面情緒字詞以人工方式分類情緒類別，以建構情緒類別詞彙庫，但「知網」所提供之情緒詞字體為簡體中文，有鑑於簡體中文與正體（繁體）中文之方言差異，故依據教育部國語、成語字典篩選具方言之情緒字詞（如表 1 範例所示），以保持情緒類別詞彙庫之語意一致性；其次，利用中文詞知識小組（Chinese Knowledge Information Processing Group；CKIP）之中文斷詞系統挑選論壇文章候選事件，其中候選事件多數可視為論壇文章之代表詞彙，並依據 Tung 與 Lu（2010）之方法利用論壇文章候選事件之詞性、前後文配對及標題分數計算文章各候選事件之權重分數，以取得論壇文章代表事件，並參考 Zhao 等人（2010）以向量空間模型（Vector Space Model；VSM）之餘弦函數（Cosine）方式計算目標論壇文章代表語句與所有訓練文章之語句相似程度後，再以帕茲模型（Potts Model）取得目標論壇文章代表語句所表達之情緒類別，以作為目標論壇文章表達情緒類別之依據（如圖 7 所示）。

表 1、方言情緒字詞之比較

情緒詞類別	簡體字體	正體字體	是否為方言	教育部國語辭典之詞義	地方方言之詞義
正面情緒	留个 心眼儿	留個 心眼兒	是	-	「小心」之意
	不离儿	不離兒	是	-	「差不多」之意 形容行走歡樂的樣子
	敞开儿	敞開兒	否	恣意、儘量	-
	把稳	把穩	否	主意堅定，不可動搖	-
負面情緒	诧愕	詫愕	是	-	「驚奇」之意
	堵得慌	堵得慌	否	「悶得慌」之意	-
	发怔	發怔	否	因心神不貫注而眼睛呆視的樣子	-
	忿恚	忿恚	否	怨恨發怒	-

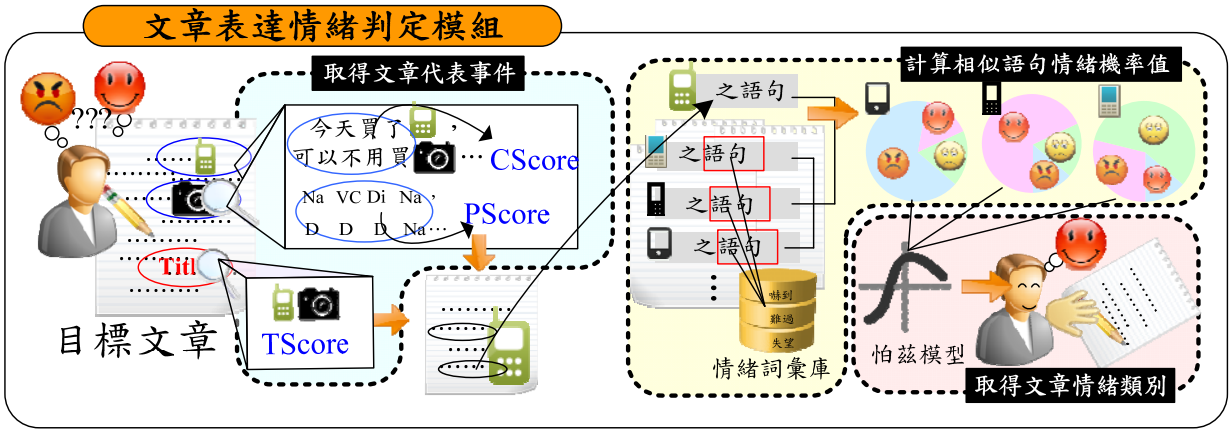

圖 7、文章表達情緒判定示意圖

步驟(A1)—計算文章事件之前後文配對分數及標題加權分數

根據候選事件 ACE_i 前後語句（即標點符號與標點符號間之語句）與事件觸發詞之關係，可得知觸發詞與候選事件 ACE_i 乃具有分佈規則，故本步驟乃計算各觸發詞於目標候選事件前後語句及論壇文章內之分佈比例並相加取平均值後，即可取得目標候選事件 ACE_i 前後文配對平均分數，其值愈大即表示該候選事件為文章事件之代表性愈高。此外，論壇文章之標題可能代表文章之主旨，因此，將論壇文章候選事件 ACE_i 與標題之關係計算標題加權分數，以提升取得論壇文章之代表事件精確度，最後將兩分數相加即為候選事件之前後文配對分數及標題加權 $TCScore(ACE_i)$，如公式(1)所示。

$$TCScore(ACE_i)=\left[\frac{\sum_{all\,j}\left(\frac{Fre(CES^{F}_{i,j})}{N(CES^{F}_{\bullet,\bullet})}+\frac{N(CES^{F}_{i,j})}{N(ACE_{\bullet})}\right)}{\sum_{all\,k}\left(\frac{Fre(CES^{B}_{i,k})}{N(CES^{B}_{\bullet,\bullet})}+\frac{N(CES^{B}_{i,k})}{N(ACE_{\bullet})}\right)}\cdot\frac{Fre(ACE_i,Title)}{}\right]\cdot\left[\frac{1}{N(CES^{F}_{i,\bullet})}\;\frac{1}{N(CES^{B}_{i,\bullet})}\;\frac{1}{Fre(ACE_i)}\right] \quad (1)$$

步驟(A2)—取得論壇文章事件之詞性相關性平均分數

由於事件之詞性與事件前後詞彙之詞性具有關聯規則，故根據中研院漢語平衡語料庫（Sinica Corpus）所制定之詞性，先行統計所有文章中所有詞彙之詞性所佔比率，並標記詞性距離參數值 $Sdis(POS_{\bullet})$（如表 2 所示），以訓練詞性距離之精準度。事件之詞性相關性平均分數計算公式如公式(2)、公式(3)所示，首先乃建立所有候選事件向前推算 $Sdis(POS_{\bullet})$個詞彙集合 Set(VEF)，並統計目標候選事件向前推算之各詞彙於詞彙集合 Set(VEF)及文章內出現次數，以取得各詞彙之詞性相關性機率，進而計算候選事件之詞性相關性平均分數 $PScore(ACE_i)$。

表 2、訓練文章所有候選事件詞性種類之統計

詞性種類	數量	所佔比率	詞彙距離
POS_1=Na（普通名詞）	36	34.6%	Sdis(POS)=1
POS_2=VE（動作句賓動詞）	16	15.4%	Sdis(POS)=2
POS_3=NH（代名詞）	16	15.4%	Sdis(POS)=3
POS_4=VC（動作及物動詞）	14	13.5%	Sdis(POS)=4
POS_5=ND（時間詞）	7	6.7%	Sdis(POS)=5
POS_6=VA（動作不及物動詞）	7	6.7%	Sdis(POS)=6
POS_7=VH（狀態不及物動詞）	5	4.8%	Sdis(POS)=7
POS_8=VB（動作類及物動詞）	2	1.9%	Sdis(POS)=8
POS_9=NC（地方詞）	1	1%	Sdis(POS)=9

$$Set(VEF)=\begin{bmatrix} VEF_{1,1} & VEF_{2,1}\cdots VEF_{i,1} \\ VEF_{1,2} & VEF_{2,2}\cdots VEF_{i,2} \\ \vdots & \vdots \quad\quad \vdots \\ VEF_{1,n} & VEF_{2,n}\cdots VEF_{i,n} \end{bmatrix} \tag{2}$$

$$PScore(ACE_i)=\frac{\sum_{all\,n}\frac{Fre[Set(VEF_{i,n})]}{N(VEF_{i,n})}}{\beta(POS_{\bullet})\cdot N(ACE_{\bullet})\cdot Fre[Set(VEF_i)]} \tag{3}$$

步驟(A3)—取得論壇文章代表性事件

因考量論壇文章類型之語意及標題重要性不同，故本步驟將透過步驟(A1)及步驟(A2)所取得候選事件之詞性相關性平均、前後文配對分數及標題加權分數加上比重分數並相加後，即可取得各候選事件之代表分數，其中代表分數最大之候選事件即為文章之代表事件，如公式(4)所示。

$$\begin{gathered} EScore(ACE_i)=PPScore(ACE_i)\cdot PScore(ACE_i)+PTCScore(ACE_i)\cdot TCScore(ACE_i) \\ Where\,PPScore(ACE_i)+PTCScore(ACE_i)=1 \end{gathered} \tag{4}$$

步驟(A4)—建立情緒詞彙與情緒類別之隸屬係數

本步驟藉由 Gregory 及 Daniel（2008）中所建構之 482 個常用英文情緒詞彙及所對應之情緒類別轉換為中文字詞，並將情緒類別 S_w 分為憤怒（Angry）、焦慮（Anxiety）、厭惡（Disgust）、恐懼（Fear）、快樂（Happiness）、悲傷（Sadness）、驚奇（Surprise）等七個類別，同時依據既有情緒詞彙所對應之情緒類別（如表 3 所示，各類別以 2 個詞彙表示）建立情緒語句集合 SS_Set_w（如公式(5)所示），之後參考 Miao 等人（2012）之方法計算目標情緒詞彙於該情緒語句集合 SS_Set_w 中出現頻率，並統計訓練文章中各情緒類別之情緒詞彙總數後，即以公式(6)計算目標情緒詞彙與各情緒類別之相關係數 $ReS'(DE_d,S_w)$，並將相關係數予以正規化如公式(7)所示，即可得知目標情緒詞彙與情緒類別之隸屬係數，其結果整理如表 4，此係數值愈大即代表情緒詞彙愈偏向該對應情緒類別。

$$SS\_Set_w=L_{p,b}\ \text{where}\ LS[S_w,DE_d]\ \text{exist in}\ L_{p,b}\ \forall d \tag{5}$$

$$ReS'(DE_d, S_w) = \frac{N(DE_d \cap SS\_Set_w)}{\sum_{alld} N[L_\bullet, LS(S_w, DE_d)]} \times \log_2(N(DE_d \cap SS\_Set_w) + 1) \quad (6)$$

$$ReS(DE_d, S_w) = \frac{ReS'(DE_d, S_w)}{\sum_{allw} ReS'(DE_d, S_w)} \quad (7)$$

表 3、情緒類別所對應之情緒詞彙

情緒類別	英文情緒詞	中文情緒詞
憤怒（Angry）	Contempt	鄙視、輕視、藐視
	Violent	激烈、暴力
焦慮（Anxiety）	Awkward	尷尬、笨拙、不熟練的
	Uneasy	不安、不自在、擔心
厭惡（Disgust）	Villain	壞人、惡棍
	Stinking	惡臭、非常討厭
恐懼（Fear）	Horror	恐怖、毛骨悚然
	Doom	死亡、毀滅、惡運
快樂（Happiness）	Friendly	友好、友善、親切
	Pleasant	愉快、爽快、舒適
悲傷（Sadness）	Despair	失望、絕望
	Grief	悲苦、悲痛、哀痛
驚奇（Surprise）	Amazed	吃驚、驚詫
	Shocked	震驚

表 4、情緒詞彙與各情緒類別之隸屬係數

情緒詞彙 / 情緒類別	DE_1	DE_2	…	DE_d	…
S_1	$ReS[DE_1, S_1]$	$ReS[DE_2, S_1]$	…	$ReS[DE_d, S_1]$	…
S_2	$ReS[DE_1, S_2]$	$ReS[DE_2, S_2]$	…	$ReS[DE_d, S_2]$	…
…	…	…	…	…	…
S_w	$ReS[DE_1, S_w]$	$ReS[DE_2, S_w]$	…	$ReS[DE_d, S_w]$	…

步驟(A5)—計算論壇文章代表語句與所有訓練文章內語句之相似值

透過前三步驟所取得論壇文章之代表事件後，本研究乃將訓練文章之語句向量集合 L_b^{ω} 及目標文章代表語句 ADS_q（即具有代表事件之語句）之集合向量 ADS^{ω}_q，以向量空間模型之餘弦函數計算目標論壇文章之代表語句與所有訓練文章內之語句相似度，並以公式(8)判斷代表語句 ADS_q 與所有語句之相似值 $Sim(ADS_q, L_b)$。

$$L_{p,b} = \{L_{p,1}, L_{p,2}, L_{p,3}, \cdots, L_{p,b}, \cdots\},\ L_b^{\omega} = [w_1, w_2, \cdots, w_b]^T,\quad ADS^{\omega}_q = [w_1, w_2, \cdots, w_q]^T$$

$$Sim(ADS_q, L_b) = \frac{ADS^{\omega}_q \cdot L^{\omega}_b}{|ADS^{\omega}_q| \cdot |L^{\omega}_b|} \quad (8)$$

步驟(A6)—計算相似語句之情緒類別機率值

當取得論壇文章代表語句 ADS_q 與所有語句之相似值 $Sim(ADS_q, L_b)$ 後，若相似值 $Sim(ADS_q, L_b)$ 大於門檻值 $\omega(ADS_q, L_b)$ 且愈趨近於 1，即表示該文章語句與代表語句 ADS_q 具有相似之語意，但語句常因否定詞之存在而導致語句與原意相反，故若語句否定詞頻率大於參數值 α(NW)，即視該語句與代表語句不為相似，此外，本步驟亦擷取相似語句中情緒詞彙與重要詞彙形成代表語句之極性項目組合，若相似語句中無名詞 Na 或地方詞 Nc 以及情緒詞彙之存在，將忽略該相似語句之極性項目組合，

以保持代表語句之情緒詞彙與重要詞彙關聯性，如公式(9)所示，並將極性項目所含情緒詞彙與情緒類別之隸屬係數 $ReS(DE_d,S_w)$，作為該極性項目與情緒類別之隸屬係數 $R[ADS\_DI_{q,y},S_w]$，如公式(10)所示，最後即可以公式(11)計算所有相似語句偏向各情緒類別之機率值 $SP[Fi(ADS_q,S_u),S_w]$，其計算結果整理如表 5 所示。

$$\begin{aligned}&IF\ Sim(ADS_q,L_b)\geq\omega(ADS_q,L_b)\ and\ N(L_b\cap Set(NW))<\alpha(NW)\ and\\&NANC\ exist\ in\ L_b\ and\ DE_d\ exist\ in\ L_b\ Then\ Sim(ADS_q,L_b)\in Sim\_Fi(ADS_q,S_u)\\&and\ NANC,DE_d\in ADS\_DI_q\ \forall_d\end{aligned}\tag{9}$$

$$R[ADS\_DI_{q,y},S_w]=ReS(DE_d,S_w)\ \text{ where } DE_d \text{ exist in } ADS\_DI_{q,y}\ \forall d\tag{10}$$

$$SP[Fi(ADS_q,S_u),S_w]=\frac{\exp\left(I[Fi(ADS_q,S_u),TW)]+Sim\_Fi(ADS_q,S_u)\times\frac{\sum_{all\ y}R[ADS\_DI_{q,y},S_w]}{N(ADS\_DI_{q,y}\cap Fi(ADS_q,S_u))}\right)}{\sum_{all\ w}\exp\left(I[Fi(ADS_q,S_u),TW)]+Sim\_Fi(ADS_q,S_u)\times\frac{\sum_{all\ y}R[ADS\_DI_{q,y},S_w]}{N(ADS\_DI_{q,y}\cap Fi(ADS_q,S_u))}\right)}\tag{11}$$

$$\text{where } I[Fi(ADS_q,S_u),TW)]=\begin{Bmatrix}0,\ N(Fi(ADS_q,S_u)\cap TW)>0\\1,\ otherwise\end{Bmatrix}\ \text{and}\ ADS\_DI_{q,y}\ \text{exist in}\ Fi(ADS_q,S_u)\ \forall\ y$$

表 5、相似語句與情緒類別之機率值

情緒類別 \ 相似語句	$Fi(ADS_q，S_1)$	$Fi(ADS_q，S_2)$	…	$Fi(ADS_q，S_u)$	…
S_1	$SP[Fi(ADS_q，S_1)，S_1]$	$SP[Fi(ADS_q，S_2)，S_1]$	…	$SP[Fi(ADS_q，S_u)，S_1]$	…
S_2	$SP[Fi(ADS_q，S_1)，S_2]$	$SP[Fi(ADS_q，S_2)，S_2]$	…	$SP[Fi(ADS_q，S_u)，S_2]$	…
…	…	…	…	…	…
S_w	$SP[Fi(ADS_q，S_1)，S_w]$	$SP[Fi(ADS_q，S_2)，S_W]$	…	$SP[Fi(ADS_q，S_u)，S_w]$	…

步驟(A7)—取得論壇文章之情緒類別

透過相似語句與各類別情緒之機率值 $SP[Fi(ADS_q,S_u),S_w]$ 即可以怕茲模型（如公式(12)所示）判斷論壇文章代表語句對於情緒類別之穩定值 $ST[AL\_DI_q,S_w]$（整理如表 6），若該值愈趨近於 0 即表示文章代表語句對於該情緒類別偏向性愈高，但由於文章中大多包含數個代表語句，且論壇文章中常隱含數種情緒類別，故本研究乃將各代表語句之情緒類別視為目標論壇文章所表達之情緒類別。

$$\begin{aligned}ST[ADS_q,S_w]=&-\sum_{allu}SP[Fi(ADS_q,S_u),S_w]\times I[Fi(ADS_q,S_u),TW)]\\&-\sum_{allu}SP[Fi(ADS_q,S_u),S_w]\times Sim\_Fi(ADS_q,S_u)\times I[Fi(ADS_q,S_u),TW)]\\&-\sum_{allu}-SP[Fi(ADS_q,S_u),S_w]\times\log(SP[Fi(ADS_q,S_u),S_w])\end{aligned}\tag{12}$$

表 6、代表語句與情緒類別之穩定值

情緒類別 \ 代表語句	ADS_1	ADS_2	…	ADS_q	…
S_1	$ST[ADS_1，S_1]$	$ST[ADS_2，S_1]$	…	$ST[ADS_q，S_1]$	…
S_2	$ST[ADS_1，S_2]$	$ST[ADS_2，S_2]$	…	$ST[ADS_q，S_2]$	…
…	…	…	…	…	…
S_w	$ST[ADS_1，S_w]$	$ST[ADS_2，S_w]$	…	$ST[ADS_q，S_w]$	…

於「文章達情緒判定」模組中乃將文章中具有代表事件之語句作為論壇文章代表語句，並分析代表語句與所有語句之關聯性，從中篩選與代表語句相似之語句，以藉由相似語句中情緒詞彙與情緒類別之關係，推論相似語句之情緒類別偏向性，再以怕茲模型推論文章代表語句之情緒類別偏向，以視為論壇文章所表達之情緒類別，亦即透過本模組可幫助使用者了解論壇文章所隱含不同之情緒類別。

3.2 發文者文章語句重組模組

本研究乃將「發文者文章語句重組模組」規劃為六大步驟，並於最終可重組文章之語句內容。於步驟(B1)與步驟(B2)中乃考量重組文章語句之內容後，期望可與原文章之語句流暢程度相似，並可以於最終推論之數個語句架構中，取得語句流暢程度最高之文章作為最終結果，故乃參考Chen等人（2010）判定語句流暢程度之方法規劃至步驟(B1)與步驟(B2)中，待取得文章之語句流暢程度後，考量重複分析語意相似之語句時，將影響後續推論之結果，因此於步驟(B3)中乃參考Kuo（2007）之方法規劃去除相似語句推論；此外，由於過去研究大多以相似語句或詞彙替換作為重組語句之方法，因此，規劃步驟(B4)至步驟(B5)時乃參考同為分析論壇與非結構化文章Liu等人（2011）之方法，將相似語句、填充詞與候選填充句取得方法規劃於此兩步驟，完成上述步驟後，即可滿足語句重組之基礎條件，最後於步驟(B6)規劃中，本研究乃提出一方法依據語句順序重組語句之內容。

過去研究針對文章語句重組部份，改寫後大多無法與原先欲傳達之情緒、情感一致，而無法滿足使用者之需求，且論壇發文者發表文章時，內文完整性與流暢性乃吸引讀者注目關鍵之一。針對發文者語意表達程度部分，本研究先行參考Chen等人（2010）之方法針對目標論壇文章之間接特徵（文章所使用之詞語種類數）給予初始分數，再以投票演算法（Voting Algorithm）將文章間之相似程度以累加式方式，持續修正目標論壇文章評閱分數值直至收斂穩定狀態，以判斷該論壇文章是否須進行後續改寫內容之程序。本研究針對文章語句重組部分乃替換文章撰寫者所描述之情緒語句及重要語句，故本研究參考Kuo（2007）之方法先行針對各語句詞義消歧部分去除語意相近之句子，以取得文章之重要語句，接著本研究乃將重要語句及代表語句（具代表事件之語句）與訓練文章庫之所有語句進行相似分析，以取得論壇文章重組時所需之候選填充句，但因候選填充句、重要語句、代表語句組合後可能產生句子不連貫，是故，本研究乃藉由訓練文章庫中所有文章之語句彙集候選填充句之非關鍵詞(即填充詞)，以形成候選填充句之多重組合句，最後，透過多重組合句之組合形成數個論壇文章之語句架構，並依據最高評閱分數之語句架構，作為目標論壇文章重組後之內容。本模組之語句重組流程如圖8所示。

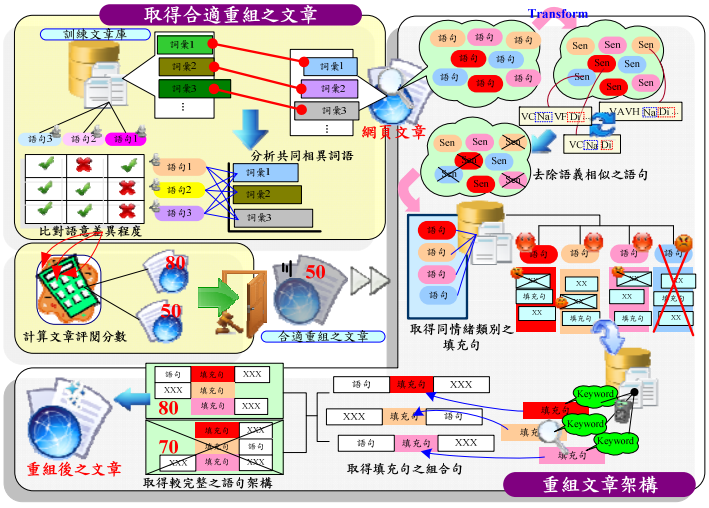

圖 8、發文者文章語句重組示意圖

步驟(B1)—取得論壇文章之初始評分

本步驟乃利用目標論壇文章A_T之表面特徵「相異詞語數」作為初始評分之依據，首先乃去除文章重複出現之詞語，將相同之詞語視為1種詞語種類，即可取得目標論壇文章真正所應用之相異詞語數N(FAL_Set)，此外，本步驟亦同時計算訓練文章庫所有論壇文章之相異詞語數N(LDW_Set_p)，以便於後續步驟與目標論壇文章A_T進行相似性比對（如公式(13)所示）。

$$ADW\_Set = \{AL_j \mid AL_j \text{ not exist in } ADW\_Set \; \forall j\}$$
$$LDW\_Set_p = \{L_{p,k} \mid L_{p,k} \text{ not exist in } LDW\_Set_p \; \forall k\} \quad (13)$$

步驟(B2)—計算目標論壇文章之評閱分數

本步驟首先乃將目標論壇文章與訓練文章之共用詞語作為兩文章之相似度$Sim(A_T,L_p)$，並同時計算相似程度之標準差$Sd(A_T,L_\bullet)$，如公式(14)所示，以使評閱分數可收斂於評閱範圍中。接著透過公式(15)計算訓練文章對於目標文章之語意差異程度$Sem(A_T,L_p)$，以得知各訓練文章共用詞語之差異性，最後即以公式(16)計算目標論壇文章之評閱分數PGrade，若評閱分數PGrade低於門檻值Th_PGrade即表示該目標論壇文章之語意表達程度較低。

$$Sim(A_T,L_p)=N(ADW\_Set\cap LDW\_Set_p)$$

$$Sd(A_T,L_\bullet)=\sqrt{\frac{\sum_{all\ p}\left(Sim(A_T,L_p)-\left(\frac{\sum_{all\ p}Sim(A_T,L_p)}{L_\bullet}\right)\right)^2}{L_\bullet-1}} \tag{14}$$

$$Sem(A_T,L_p)=\frac{\left(\frac{N(LDW\_Set_p)-\sum_{p\neq m}N(LDW\_Set_m)}{L_\bullet}\right)}{Sd(A_T,L_p)} \tag{15}$$

$$PGrade=\sum_{all\ p}Sim(A_T,L_p)\times Sem(A_T,L_p) \tag{16}$$

此外，本研究乃提出四種門檻值設定方式以提供使用者針對需求篩選論壇文章，其門檻值Th_PGrade設定方式分別為平均值、中位數、最小值、門檻值直接定義與四分位數等方式，亦表示若評閱分數PGrade小於門檻值Th_PGrade即表示將該論壇文章語句及語意通順度不佳，並將該論壇文章放置合適改寫文章集合CRT_Set中，其各門檻值設定方式如下：

(a)以平均值作為門檻值

首先乃針對訓練文章庫中各篇論壇文章重複公式(14)至公式(16)計算其評閱分數並加總，再除以訓練文章總數求得評閱分數平均值，進而將平均值作為門檻值Th_PGrade_1，若評閱分數低於整體平均值，即表示該網頁文件之語意表達性較低，如公式(17)所示。

$$Th\_PGrade_1=\frac{\sum_{all\ p}LPGrade_p}{L_\bullet} \tag{17}$$

$$CRT\_Set=\{A_T\mid PGrade\le Th\_PGrade_1\}$$

(b)以中位數作為門檻值

為避免評閱分數加總最大值與最小值差距過大而影響整體平均值，因此以整體筆數之中位數作為門檻值Th_PGrade_2，如公式(18)所示，若目標論壇文章評閱分數低於（或相等）中位數，即表示該論壇文章之內容架構須進行重組程序。

$$Th\_PGrade_2=\begin{cases}LPGrade_{\frac{p+1}{2}} & IF\ L_\bullet\ is\ odd\\ \frac{1}{2}\times\left(LPGrade_{\frac{p+1}{2}}+LPGrade_{\frac{p+1}{2}+1}\right) & IF\ L_\bullet\ is\ even\end{cases} \tag{18}$$

$$CRT\_Set=\{A_T\mid PGrade\le Th\_PGrade_2\}$$

(c)直接定義門檻值

使用者亦可自行制定門檻值，以進行篩選合適語句重組之論壇文章。如公式(19)所示，若評閱分數小於自行定義門檻值Th_PGrade_3則將該網頁文件放置合適語句重組文章集合CRT_Set中。

$$CRT\_Set=\{A_T\mid PGrade\le Th\_PGrade_3\} \tag{19}$$

(d)以最小值作為門檻值

訓練文章庫乃經篩選後之文章，其中文章評閱分數最小者極可能文章語意表達性較低，因此，本研究將評閱分數最低者作為門檻值Th_PGrade_4供使用者篩選論壇文章，若評閱分數小於最小值門檻值

Th_PGrade_4，即表示該論壇文章相對於訓練文章之語意表達性較為不佳，故文章須進行語句重組之程序，如公式(20)所示。

$$\begin{aligned} Th\_PGrade_4 &= Min(LPGrade_p)\ \forall p \\ CRT\_Set &= \{A_T \mid PGrade \le Th\_PGrade_4\} \end{aligned} \tag{20}$$

(e)以四分位數作為門檻值

為能更精確挑選且不受極端值影響篩選論壇文章之結果，於此乃制定整體筆數之四分位數作為門檻值，即以第三位之四分位數進行篩選。首先計算第三位之四分位數指標 Q，以界定門檻值 Th_PGrade_5，進而篩選論壇文章之評閱分數，如公式(21)所示。當中，若評閱分數小於最小值門檻值 Th_PGrade_5 則表示該論壇文章語意不佳，並放置合適語句重組文章集合 CRT_Set 中。

$$\begin{aligned} Q &= L_{\bullet} \times 75\% \\ Th\_PGrade_5 &= \begin{cases} LPGrade_{P\times 75\%} & IF\ Q \notin \{X : |X| \in N\} \\ \frac{1}{2} \times \left(LPGrade_{p_{\times 75\%}} + LPGrade_{p_{\times 75\%+1}}\right) & IF\ Q \in \{X : |X| \in N\} \end{cases} \\ CRT\_Set &= \{A_T \mid PGrade \le Th\_PGrade_5\} \end{aligned} \tag{21}$$

步驟(B3)—去除語意相近之語句

由於中文語句常有不同詞彙組成之句子，但語意相同之情形，故本步驟乃去除論壇文章中語意相同之語句，以更精確針對重要語句進行重組。本研究乃將論壇文章目標語句利用中英翻譯器（Denisowski's CEDICT）取得語句中各詞性詞彙之英文詞彙 $ASL_{m,i}$，之後利用Wordnet取得各種詞性之英文詞彙延伸定義詞彙集合 $ASLE\_Set_m$，並針對所有語句之英文詞彙計算其關聯程度 $Re[ASL_{m,i}, ASL_{k,i}]$，如公式(22)所示，最後，將所有目標語句與其他語句之各種詞性詞彙相關程度加總後，即為目標語句與其他語句之歧義相似度 $amp[AS_m, AS_k]$，如公式(23)所示，並以最大歧義相似度之語句作為目標語句之歧義句，而本研究乃去除目標語句與歧義句中詞性種類最少之語句，以取得論壇文章之重要語句。

$$CL\_Set_i = ASLE\_Set_{m,i} \cap ASLE\_Set_{k,i}\ \forall i$$

$$Re[ASL_{m,i}, ASL_{k,i}] = Max(-\log(\frac{N(CL\_Set_i[EL_u])}{ASL_{\bullet,i}}))\ \forall u \tag{22}$$

$$amp[AS_m, AS_k] = \sum_{all\ i} Re[ASL_{m,i}, ASL_{k,i}] \tag{23}$$

步驟(B4)—取得重要語句與代表語句之候選填充句

先前步驟中所取得之重要語句及代表語句（具代表事件之語句），本步驟首先依據語句於論壇文章之位置順序，彙集成擴充語句集合 Ex_Set （如公式(24)所示），接著將擴充語句集合 Ex_Set 中所有語句與訓練文章庫中所有語句 $L_{\bullet\bullet}$ 進行相似性比對（如公式(25)所示），並以相似程度最高語句之前後語句作為候選填充句，此外，為使改寫論壇文章後可與原先論壇文章欲表達之情緒一致，故本步驟將過濾與論壇文章原先情緒不一致之候選填充句，如公式(26)所示。

$$Ex\_Set = \{AS_1, AS_2, AS_3, \cdots, AS_m \mid Max(amp[AS_m, AS_k]) \text{ and } Max(ASL_{m,\bullet})\} \tag{24}$$

$$Sim(Ex\_Set[IS_g], L_b) = \frac{Ex\_Set[IS^{\omega}{}_g] \cdot L^{\omega}{}_b}{|Ex\_Set[IS^{\omega}{}_g]| \cdot |L^{\omega}{}_b|} \tag{25}$$

$$Con\_Set = \{L_{b+1}, L_{b-1} \mid Max(Sim(Ex\_Set[IS_g], L_b)) \text{ and } LS_{p,b,d} \in AS\_Set_v\}\ \forall g, w \tag{26}$$

步驟(B5)—取得候選填充句之多重組合句

考量候選填充句 $Con\_Set[CS_q]$ 與重要語句及代表語句組合後可能產生句子不連貫之情況，因此，本步驟乃將候選填充句之關鍵詞（普通名詞、情緒詞彙、地方詞及形容詞）與訓練文章庫中所有文章之語句進行比對，以取得候選填充句之填充詞 $LF_{p,b,x}$（關鍵詞與關鍵詞間之詞彙），進而彙集候選填充句之多重組合句 $CS\_MUS_{q,z}$，如公式(27)所示。

$$CS\_MUS_{q,z} = \left\{ \begin{matrix} LF_{p,b,x} \mid Con\_Set[CS_{q,r}] \text{ exist in } L_{p,b} \\ \text{and } Con\_Set[CS_{q,r}] \text{ not exist in } LF_{p,b,x} \end{matrix} \right\} \forall r, x \quad (27)$$

步驟(B6)—重組論壇文章架構

根據前一步驟所取得候選填充句之多重組合句，本步驟乃依據語句之順序，將目標論壇文章候選填充句之多重組合句$CS_MUS_{q,z}$、代表語句及重要語句進行組合，如公式(28)所示，以將合適重組之論壇文章進行語句重組，此外，因候選填充句具有多個組合句，故重組後之論壇文章將有數個不同之語句架構，為使從中取得架構較完整之論壇文章，本步驟乃將重組後之論壇文章藉由步驟(B1)與步驟(B2)取得評閱分數Rec_Grade（彙整如表 7 所示），並由評閱分數最高之語句架構作為目標論壇文章重組後之內容。

$$Rec\_A_n = \begin{bmatrix} CS\_MUS_{1,1} \\ CS\_MUS_{1,2} \\ \vdots \\ CS\_MUS_{1,c} \end{bmatrix} \rightarrow \begin{bmatrix} CS\_MUS_{2,1} \\ CS\_MUS_{2,2} \\ \vdots \\ CS\_MUS_{2,a} \end{bmatrix} \rightarrow \begin{bmatrix} CS\_MUS_{3,1} \\ CS\_MUS_{3,2} \\ \vdots \\ CS\_MUS_{3,v} \end{bmatrix} \rightarrow \cdots \rightarrow \begin{bmatrix} CS\_MUS_{q,1} \\ CS\_MUS_{q,2} \\ \vdots \\ CS\_MUS_{q,z} \end{bmatrix} \rightarrow \cdots \quad (28)$$

where $\text{Max}(Rec\_Grade_n)$

表 7、論壇文章架構與評閱分數彙整表

論壇文章語句架構	語句架構內容	評閱分數
Rec_A_1	$CS\_MUS_{1,1} \rightarrow CS\_MUS_{2,2} \rightarrow \cdots \rightarrow CS\_MUS_{q,3}$	Rec_Grade_1
Rec_A_2	$CS\_MUS_{1,1} \rightarrow CS\_MUS_{2,3} \rightarrow \cdots \rightarrow CS\_MUS_{q,4}$	Rec_Grade_2
…	…	…
Rec_A_{14}	$CS\_MUS_{1,5} \rightarrow CS\_MUS_{2,1} \rightarrow \cdots \rightarrow CS\_MUS_{q,2}$	Rec_Grade_2
…	…	…
Rec_A_n	$CS\_MUS_{1,c} \rightarrow CS\_MUS_{2,a} \rightarrow \cdots \rightarrow CS\_MUS_{q,v}$	Rec_Grade_n

針對論壇文章語句重組部分，過去研究大多乃根據論壇文章之語意結構及流暢度進行語句重組，但大多數未加入文章撰寫者欲表達之情緒因素，使得重組後之內容常與文章撰寫者欲表達之情緒不一致，是故，本研究乃先行判斷論壇文章之語意程度，並藉由論壇文章之情緒類別分析重要語句及代表語句，以取得論壇文章重組所需之候選填充句（包含情緒語句），最後透過候選填充句之多重組合句重組論壇文章之語句架構，以期望協助文章撰寫者提升論壇文章內容之流暢度，其內容亦可與原先欲表達之情感一致。

4. 系統應用流程

根據第三章發展之方法論，本研究乃開發一套提升論壇知識利用價值之論壇文章情感解析及語句結構重組系統以確認模式可行性，當中，本研究將系統使用者分為一般使用者與系統管理者，並依權限而有不同執行權力，如圖9所示，此外，系統管理者乃針對論壇文章（以Mobile01論壇為例）進行訓練文章蒐集，如圖10及圖11所示，以作為後續解析之基礎。

一般使用者與系統管理者上傳社群文章

當使用者進入此功能輸入如標題「台灣與…」及內容「這個頭銜…」，並輸入作者「undio」，上傳自定義路徑如「D:\data」論壇文章原始檔案後（如圖12所示），系統將進行斷詞之資料預處理，斷詞後取得如「這(Nep)　個(Nf)　頭銜(Na)…」等結果，並將結果儲存與維護於資料庫中（如圖13所示）。

文章表達情緒推論

➢ 「文章代表事件解析功能」

管理者針對標題「台灣與FBI.」之文章執行文章代表事件解析功能，系統即自動抓取目標文章之候選事件，並得知候選事件「台灣」於標題中出現「1」次後，計算標題加權分數皆為「0.16」，

如圖14所示，接著，U依據前後語句詞彙之分佈值，取得候選事件之前後文配對分數「0.4」，如圖15所示，之後，系統根據詞性與前後文詞彙計算所有候選事件前文之分佈平均值為「0.24」，如圖16所示，待完成三項分數之解析後，系統即會將候選事件之三項總分予以正規化後即可得知候選事件「犯罪」之代表分數「0.459」為最大值，故可得知「犯罪」為文章之代表事件，如圖17所示。

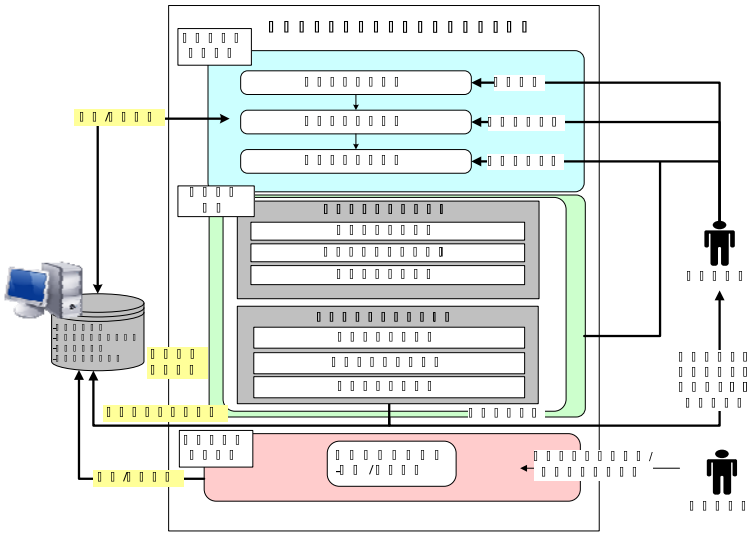

圖 9、提升論壇知識利用價值之論壇文章情感解析及語句結構重組系統運作架構

圖 10、Mobile01 論壇之文章(1)

圖 11、Mobile01 論壇之文章(2)

圖 12、文章基本資料輸入

圖 13、文章資料預處理與新增

圖 14、候選事件標題加權分數解析

圖 15、候選事件前後文配對分數解析

圖 16、事件詞性相關性分數解析

圖 17、文章代表事件解析介面

➢ 「情緒詞彙隸屬係數判定功能」

系統使用者執行情緒詞彙隸屬係數判定功能，將區分已判定情緒類別隸屬係數之詞彙個數為「7」，未判定詞彙數為「8」，並依據情緒類別之詞彙數、詞彙於情緒類別語句之頻率數，計算情緒詞彙「驚詫」於情緒類別「恐懼」、「驚奇」、「厭惡」與「焦慮」之關係係數，分別為「0.1」、「0.77」、「0.3」、「0.2」，而剩餘 3 個情緒類別之關係係數皆為「0」，如圖 18 所示，最後，系統乃將詞彙之關係係數正規化後即可得知情緒詞彙「驚詫」最大值類別隸屬係數為「0.56」，所對應之情緒類別為「驚奇」，權限內使用者亦可點選「查看」鈕得知詞彙「驚詫」於各情緒類別之隸屬分佈情況，如圖 19 所示。

圖 18、情緒詞彙之關係係數解析結果

圖 19、情緒詞彙與隸屬係數解析

➢ 「情緒詞彙隸屬係數判定功能」

系統管理者執行文章表達情緒判定功能後，系統乃分析文章中具有代表事件「台灣」系統利用向量空間模型之餘弦函數，算代表語句之相似值，之後，系統判斷相似語句中是否存在轉折詞與否定詞，同時根據相似語句分析語句於情緒類別之情緒機率值，如圖 20 所示，最後，系統將依據代表語句對應之相似語句，計算「恐懼」情緒類別之情緒穩定值為「-0.726」、「驚奇」為「-0.202」、「焦慮」為「-0.152」，並以最趨近於 0 之穩定值「-0.152」為代表語句所對應之情緒類別「焦慮」及「高興」，如圖 21 所示。

圖 20、取得相似語句之情緒機率值

圖 21、取得文章表達之情緒類別

發文者文章語句重組

➢ 「文章評閱分數判定」

系統管理者執行「文章評閱分數判定」，於此本研究乃以文字數較長之文章為例，並選定標題「台灣這些球員還嫌...等」之文章，系統即分析文章中不同之詞語數有「172」個，同時根據相異詞語數計算評分標準差為「10」，如圖 22 所示，之後，系統將依據評分標準差計算目標文章與訓練文章內容「最近親人都在問說...等」之語意差異程度為「-12.7」、共同詞語數為「29」、與訓練文章內容「他就說不高興那就不...等」之語意差異程度為「12.7」、共同詞語數為「49」，最後，系統乃依據語意差異程度與共同詞語數計算目標文章評閱分數為「264」，如圖 23 所示。

圖 22、文章初始分數計算介面

圖 23、文章評閱分數計算介面

➢ 「文章多重組合句建立」

系統管理者執行「文章多重組合句建立」後，系統乃分析目標語句之英文詞彙延伸定義集合與其他語句「小弟住的紐約天氣..」之交集詞彙為「Snowfall」，詞彙於文章出現頻率為 3 次，且文章英文代表名詞詞彙總數為 21 後，系統即根據上述數據分析得知目標語句與重要語句之關聯程度為「0.65」，如圖 24 所示，接著，系統即會根據門檻條件將語意關聯程度「0.85」、語句內容為「但還是有餘雪結冰的狀況」予以去除，如圖 25 所示，之後系統將重要語句「但這這個禮拜下了一場雪」進行相似性比對得知語句「但上個禮拜下了一場大雪」之相似度為「0.707」，該語句相似度為最高值，因此系統將判定該語句為重要語句「但這這個禮拜下了一場雪」之候選填充句，如圖 26 所示，之後，系統乃依據候選填充句「小弟住的..」之關鍵詞彙「天氣、紐約、穩定」，與訓練文章庫比對相符之語句，將具有關鍵詞彙之語句「最主要的原因...」、「紐約天氣..」等五個組合句建立成候選填充句之多重組合句，如圖 27 所示。

圖 24、取得文章各語句之關聯程度介面

圖 25、去除語意相似之語句介面

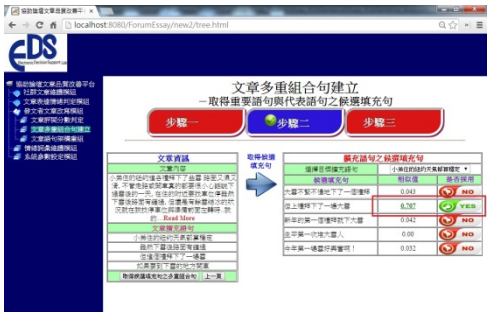

圖 26、取得語句之候選填充句介面

圖 27、取得候選填充句之多重組合句

➢ 「文章語句架構重組」

完成文章多重組合句之建立後，系統首先取得目標文章中欲替換語句共四句，，當中包含「但這個禮拜下了一場雪」有其他 5 組組合句、「如果要到下雪的地方開車」有其他 1 組組合句，得知文章語句共有「12」種不同之語句架構，當選擇第 1 組組合時，系統即會將「但這個禮拜下了一場雪」替換為「但上禮拜下了場大雪」、「如果要到下雪的地方開車」替換為「如果要再下雪天開車」並將新的文章架構呈現於介面中，如圖 28 所示，而第 2 組新的文章架構內容如圖 29 所示，之後，權限內使用者即可執行最後步驟，並點選取得最高評閱分數之文章鈕後，系統乃根據相異詞語數計算得知第 1 組文章架構相似程度標準差為「1138.01」，並得知所有語句組合中最高評閱分數為第 4 組「1183.62」，如圖 30 所示，最後，系統乃顯示第 4 組之語句架構，如圖 31 所示。

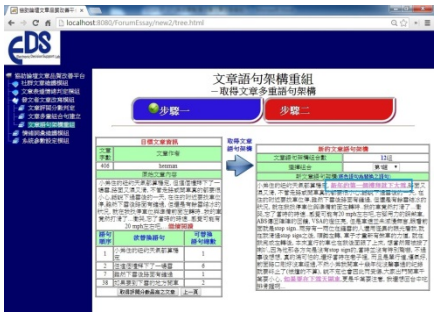

圖 28、取得文章多重語句架構-第 1 組

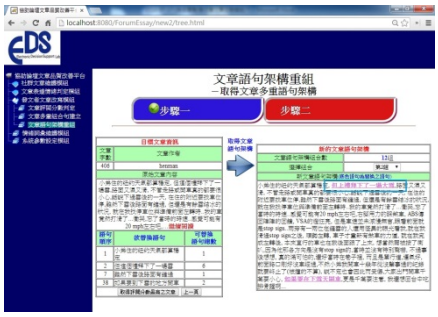

圖 29、取得文章多重語句架構-第 2 組

圖 30、取得評閱分數最高之文章組合

圖 31、評閱分數最高之文章資訊與內容

5. 系統驗證與評估

於「文章表達情緒判定」之驗證中，本研究乃分為「比較研究」與「系統自我績效評估」兩部分進行驗證；於「發文者文章語句重組」之驗證中，本研究乃以「Mobile01 論壇」之文章作為驗證與訓練文章之樣本驗證系統之績效，此外，本研究所有之驗證皆以召回率、正確率與 F 值（F-Measure）作為驗證指標。

5.1 驗證資料之蒐集與建置

本研究於「文章表達情緒判定」中，比較研究之驗證資料乃選定「痞客邦」，而系統自我績效評估則選定「奇摩新聞」作為驗證資料之來源，此外，於「發文者文章語句重組」中，本研究乃選定「Mobile01 論壇」作為驗證與測試資料之樣本，以驗證本系統之可行性。本研究乃於「痞客邦」中隨機蒐集部落客所發表之文章（如圖 32 所示）以作為訓練文章，於蒐集之過程中乃篩選字數過少之文章以使訓練文章更為確實；「奇摩新聞」乃存有數量龐大之新聞文章，以及各篇新聞皆可投票閱讀後之情緒感受等特性，是故，本研究乃隨機於各領域中蒐集投票情緒感受高於 30 人次之「新聞文章」資料；最後，本研究乃於「Mobile01 論壇」前文章數前五大之主題中隨機蒐集訓練文章，實際文章如圖 33 所示，於蒐集過程中亦篩選字數過少之文章，以建構語句較為完整之「候選填充句集合」，進而於後續取得更具準確性推論與驗證之結果。

圖 32、痞客邦之文章

圖 33、奇摩新聞之文章

5.2 文章表達情緒判定之驗證

針對「文章表達情緒判定」本研究乃分為「比較研究」與「系統自我績效評估」兩方式驗證系統之績效，以下乃針對各驗證方式細述與說明驗證設計、驗證績效與結果。

(A-1)與其他研究比較之驗證方式說明

本研究乃參考 Tung 與 Lu（2012）於相同驗證情境與驗證資料來源，於「痞客邦」部落格中隨機蒐集 1300 筆驗證資料（如表 8 以 2 份實際文章為例），同時額外蒐集 500 筆訓練文章，並以人工方式判斷 1300 筆驗證資料之實際情緒類別，而情緒類別與比較研究相同分為「生氣」、「快樂」、「思考」、

「害怕」、「悲傷」、「擔心」、「驚嚇」7 大情緒類別，爾後，將 500 筆訓練文章逐一匯入系統中，以使系統能自動建置與推論情緒詞彙之情緒類別係數；完成情緒詞彙隸屬係數之推論後，再逐一推論 1300 筆測試資料之情緒類別，最後，將推論結果與實際結果進行統計藉此得知系統推論之績效，再與 Tung 與 Lu（2012）比較驗證之績效。

表 8、痞客邦實際測試文章（以 2 份為例）

編號	新聞文章標題	新聞文章內容	實際情緒類別
1	回到令人害怕的工作崗位	6 點就起床 中壢 陰冷的天氣 外面飄著毛毛雨捨不得離開溫暖的被窩~~還有溫暖的人工暖爐~~~再怎麼不捨得 還是要振作!在車上吃了我家大頭買的愛心飯團和熱豆漿	害怕
2	不說才最難過傷心	自己也不清楚到底怎麼回事，許多解釋都很像藉口累積了許多情緒，又會害怕以文章的方式面對，太清楚了，即使很希望也很喜歡這個出口，因為真的太清楚了	悲傷

(A-2)與其他研究之比較-驗證結果分析

本研究於 500 份訓練文章之數量下，系統乃針對 1300 份測試資料進行情緒類別之判定，其中，本研究系統判定結果於 1300 筆測試資料中，推論正確實際情緒類別之個數為 1068，其召回率、正確率與 F 值皆為「82.1%」，而 Tung 與 Lu（2012）推論之三項指標皆為「72.5%」，因此，本研究於相同之資料與驗證方式下，三項指標之推論數據皆略高於 Tung 與 Lu（2012）近 10%之績效（推論數據之比較如圖 34 所示），亦即表示本研究所提出之方法論與開發之系統，推論成效將略優於 Tung 與 Lu（2012）。

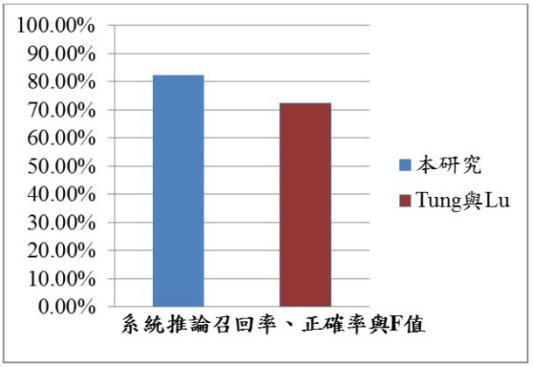

圖 34、本研究與 Tung 與 Lu（2012）推論數據之比較

(B-1)系統自我評估之驗證方式說明

首先，本研究乃於「奇摩新聞」中隨機蒐集 1000 筆訓練文章，並隨機挑選 20 份奇摩新聞文章作為測試資料（如表 9 以 3 份示之），當中測試資料皆高於 30 人次投票閱讀後情緒感受之文章，此外，本研究乃依據奇摩新聞網站所分類之情緒，將情緒類別分為 7 大類別，接著，本研究乃將系統驗證過程規劃為兩階段，於系統驗證第一階段乃於 1000 筆訓練文章中隨機挑選 200 筆匯入至系統中，並針對 20 份測試資料進行推論，以觀察系統初期之判定績效。待完成上述之第一階段系統績效驗證後，即進行系統測試第二階段之驗證，於此階段乃分為 8 個週期，每週期持續匯入 100 筆訓練文章資料（共計 800 筆），藉由持續匯入以分析系統於不同訓練測試資料下之推論結果，於各個週期乃利用前述 20 份測試資料重新進行推論，以分析系統之長期學習趨勢。

表 9、奇摩新聞之實際測試文章（以 2 份示之）

編號	文章標題	文章內容	實際情緒類別
1	子孫為領 18 趴 老父變活死人	（中央社記者陳清芳台北 6 日電）陽明大學附設醫院醫師陳秀丹今天說，有位老校長 7、8 年來靠呼吸器飽	難過
2	《阿嬤 揹巾》情深無怨	「人有兩撇好寫，人卻難做！」64 歲婦人郭秀酌為了照顧患有多重身障的 12 歲孫女「奕妡」，從出生第一天	感人

(B-2)系統自我績效評估-驗證結果分析

本研究乃將系統自我績效評估分為「第一階段驗證結果分析」與「第二階段驗證結果分析」兩大項目，以下即針對各階段說明系統驗證過程與結果。

第一階段驗證結果分析

於第一階段系統驗證中，於 200 筆訓練文章基礎之下，系統針對 20 筆實際測試文章進行判定，

獲得情緒類別推論平均召回率、正確率與 F 值為 45%，當中，實際文章情緒類別數為 20 個，系統推論之情緒類別數為 20 個，推論之正確數為 9 個，此階段詳細推論結果與三項指標之分佈趨勢如圖 35 所示。整體而言，於此階段中，召回率、正確率與 F 值平均分佈趨勢呈現兩極化之狀態，但目前階段尚有過半之驗證資料無法判定正確，是故，文章情緒類別推論之準確率與績效欠佳。

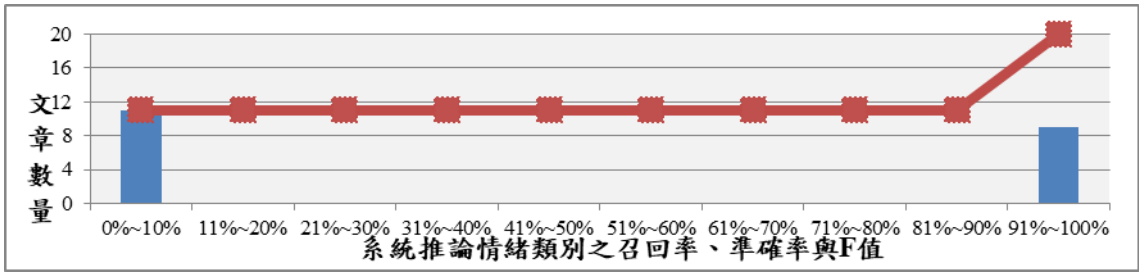

圖 35、第一階段情緒類別推論三項指標推論之分佈趨勢

第二階段驗證結果分析

本研究將第二階段分為 8 個週期，並於各週期匯入 100 份訓練文章，以觀察隨訓練文章之增加，各週期驗證指標變化之趨勢，於此，各週期驗證相關結果整理如表 10 所示，而各驗證週期文章情緒類別三項指標推論之績效分佈趨勢如圖 36 所示。藉由對表 10 之觀察可得知，以每週期增加 100 份訓練文章為單位，平均每週期三項驗證指標之整體成長率約為 4.4%，而針對最後第九週期（共匯入 1000 份訓練文章）之驗證結果言之，三項指標分別由第一週期之 45%提升至 80%，因此，綜合上述之觀察，本研究所開發之文章表達情緒判定乃具備學習能力與相當程度之正確性。

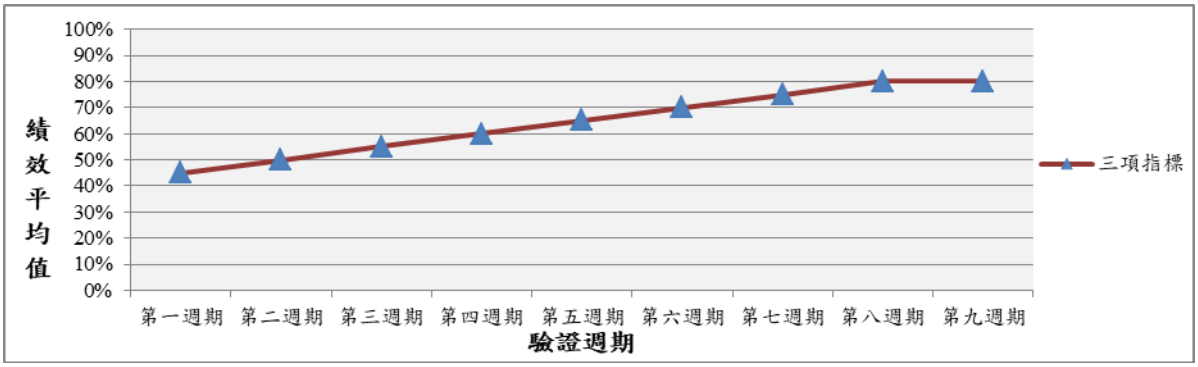

圖 36、系統自我績效評估-各驗證週期推論績效之分佈趨勢

表 10、系統自我績效評估之推論結果彙整

文章表達情緒判定-系統自我績效評估		各週期情緒類別推論之驗證-訓練文章匯入數量									平均
		第一階段	第二階段(各週期)								
		第一週期 200 份	第二週期 300 份	第三週期 400 份	第四週期 500 份	第五週期 600 份	第六週期 700 份	第七週期 800 份	第八週期 900 份	第九週期 1000 份	
召回率	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
正確率	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
F 值	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%

5.3 發文者文章語句重組之驗證

本研究所提之「發文者文章語句重組模組」乃針對違規文章之語句進行重組，以避免違反發文規範。故於驗證前，本研究乃先行建置驗證樣本，並將驗證樣本作為系統驗證之對照樣本。

發文者文章語句重組驗證方式說明

本研究乃於「Mobile01」論壇中隨機蒐集 600 筆訓練文章，而於驗證資料部分乃於論壇中隨機收集 20 份違規文章（如表 11 以 2 份示之）。將違規文章依標點符號分割語句並隨機排列，以形成 20 份測試文章，接著，本研究乃將系統驗證規劃為兩階段，於系統驗證第一階段乃於 600 筆訓練文章中隨機挑選 200 筆匯入至系統中，並針對 20 份測試資料進行推論，以觀察系統初期之判定績效，進而觀察本研究所提方法論之正確性。待完成上述之第一階段系統績效驗證後，即進行系統測試第二階段之驗證，於此階段乃分為 4 個週期，每週期持續匯入 100 筆訓練文章資料（共計 600 筆），藉由持續匯入以分析系統於不同訓練測試資料下之推論結果，於各個週期乃利用前述 20 份測試資料重新進行推論，以分析系統之長期學習趨勢。

表 11、測試文章編號 1 之偏激語句判斷問卷

語句編號	語句內容	請給予一個情緒
1	這樣對岸打過來要怎麼辦	☐ 偏激情緒 ☐ 非偏激情緒
2	不過軍方在接受媒體詢問時卻表示	☐ 偏激情緒 ☐ 非偏激情緒

於驗證前本研究乃需建置測試資料之「實際需被改寫語句數」以作為系統推論之對照樣本，因此，首先乃邀請 30 位長期使用資訊論壇且擁有資訊背景、熟悉論壇發文規範之學生（皆屬於資工系與資管系）作為初期受測者，針對 20 份測試文章挑選將可能被歸類為違規文章之語句，以形成測試文章之實際違規語句，此外，為了驗證受測者對於閱讀違規文章語句之主觀差異性與一致性，本研究乃以 20 份違規文章內所有之語句為依據，形成各測試文章之違規語句判斷問卷（如表 11 以 2 份為例），同時隨機組合 1 份重複性測試文章（語句數為 20 份測試文章語句數之平均值），將重複性測試文章隨機安排至 20 份測試文章後，再請受測者針對 20 份測試文章進行「實際違規語句」之評估，並以均方根（Root Mean Square；RMS）測試受測者主觀差異性，以「受測者重複性」、「受測者正確性」與「整體主觀差異值」三項指標，評估 30 位初期受測者閱讀語句之「主觀一致性與差異性」，最後，本研究乃以「整體主觀差異值」排名最好之前 20 位作為最終受測者（受測結果如表 12 以 10 位受測者為例），並以此些受測者所挑選之違規語句，作為 20 份驗證文章之實際違規語句，三項受測者主觀評估指標定義與說明如下：

受測者重複性指標（如公式(29)所示）為「受測者第一次挑選之違規語句數」與「受測者第二次挑選之違規語句，與第一次挑選之違規語句相符個數」之差異平均；受測者正確性指標（如公式(30)所示）為「受測者第二次挑選之違規語句個數」與「受測者第一次挑選之違規語句相符合個數」之差異平均；受測者整體主觀差異指標（如公式(31)所示）為「受測者重複性乘以受測者正確性再乘以二」與「受測者重複值加受測者正確性」之比例，期望透過此項指標評估受測者主觀程度，當中，主觀程度愈低（即愈趨近於 0），即標示受測者具備良好的主觀感受，相關符號定義如下：

OTS　受測者受測次數

ORF_i　第 i 位受測者於所有測試文章中，第一次挑選之違規語句個數

ORS_i　第 i 位受測者於所有測試文章，第二次挑選之違規語句與第一次之相符個數

ORE_i　第 i 位受測者之受測重複性

OAF_i　第 i 位受測者於所有測試文章中，受測者第二次挑選之違規語句個數

OAS_i　第 i 位受測者於所有測試文章，第二次挑選之違規語句與第一次之相符個數

OAC_i　第 i 位受測者之正確性

$$ORE_i = \sqrt{\frac{(ORF_i - ORS_i)^2}{OTS}} \quad (29)$$

$$ORe_i = \sqrt{\frac{(ORF_i - ORS_i)^2}{OTS}} \quad (30)$$

$$OSU_i = \frac{2 \times ORE_i \times ORA_i}{ORE_i + ORA_i} \quad (31)$$

表 12、10 位受測者主觀差異性調查結果

受測者	1	2	3	4	5	6	7	8	9	10
第一次挑選語句總數	64	36	41	54	60	50	52	57	63	61
第二次挑選語句總數	47	58	54	61	58	62	54	61	60	59
第二次挑選語句與第一次之相符總個數	30	30	37	45	50	39	47	39	56	48
受測者重複性	24.0	4.2	2.8	6.4	7.1	7.8	3.5	12.7	4.9	9.2
受測者正確性	12.0	19.8	12.0	11.3	5.7	16.3	4.9	15.6	2.8	7.8
整體主觀差異指標(愈趨近於 0 愈好)	16.03	6.99	4.58	8.15	6.29	10.52	4.12	14.00	3.60	8.43
排名	2	13	19	12	16	7	21	3	24	11
不採納之使用者	不採納					不採納		不採納		

驗證結果分析

本研究乃將系統自我績效評估分為「第一階段驗證結果分析」與「第二階段驗證結果分析」兩大項目，以下即針對各階段說明系統驗證過程與結果。

第一階段驗證結果分析

於第一階段系統驗證中，於 200 筆訓練文章基礎之下，系統針對 20 筆實際測試文章進行判定，

獲得情緒類別推論平均召回率、正確率與 F 值為 45%，當中，實際文章情緒類別數為 20 個，系統推論之情緒類別數為 20 個，推論之正確數為 9 個，此階段詳細推論結果與三項指標之分佈趨勢如圖 37 所示。於此階段中召回率、正確率與 F 值平均分佈趨勢呈現兩極化之狀態，是故文章情緒類別推論之準確率與績效欠佳。

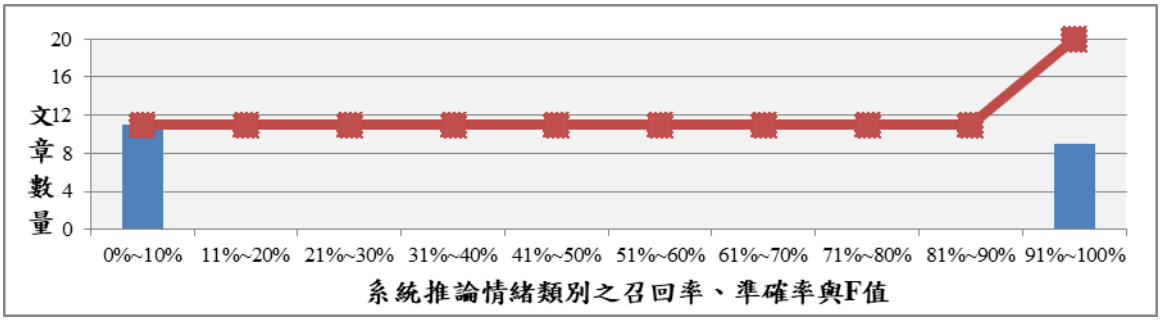

圖 37、第一階段情緒類別推論三項指標推論之分佈趨勢

第二階段驗證結果分析

本研究將第二階段分為 8 個週期，並於各週期匯入 100 份訓練文章，以觀察隨訓練文章之增加，各週期驗證指標變化之趨勢，於此，各週期驗證相關結果整理如表 13 所示，而各驗證週期文章情緒類別三項指標推論之績效分佈趨勢如圖 38 所示。

藉由對表 13 之觀察可得知，以每週期增加 100 份訓練文章為單位，平均每週期三項驗證指標之整體成長率約為 4.4%，而針對最後第九週期（共匯入 1000 份訓練文章）之驗證結果言之，三項指標分別由第一週期之 45%提升至 80%，因此，綜合上述之觀察，本研究所開發之文章表達情緒判定乃具備學習能力與相當程度之正確性。

表 13、系統自我績效評估之推論結果彙整

文章表達情緒判定-系統自我績效評估		各週期情緒類別推論之驗證-訓練文章匯入數量									平均
		第一階段	第二階段(各週期)								
		第一週期 200 份	第二週期 300 份	第三週期 400 份	第四週期 500 份	第五週期 600 份	第六週期 700 份	第七週期 800 份	第八週期 900 份	第九週期 1000 份	
召回率	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
正確率	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%
F 值	平均值	45%	50%	55%	60%	65%	70%	75%	80%	80%	64%
	成長率	-	5%	5%	5%	5%	5%	5%	5%	0%	4.4%

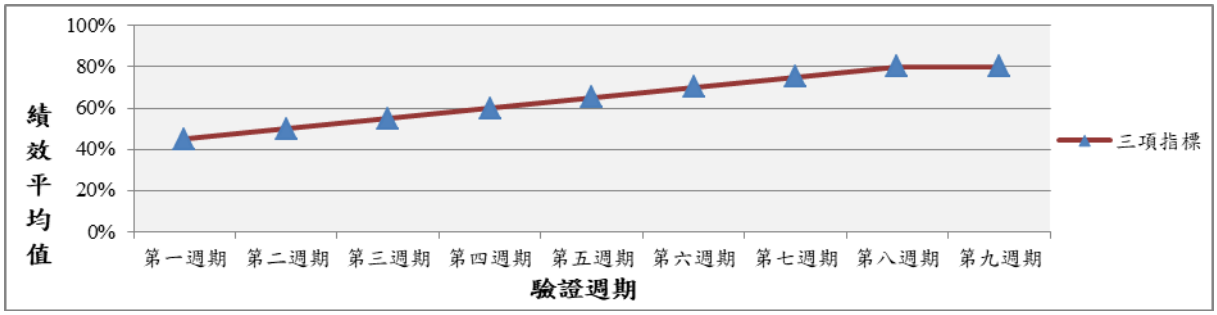

圖 38、系統自我績效評估-各驗證週期推論績效之分佈趨勢

6. 結論

近年因網路之蓬勃發展及論壇之崛起，使用者可透過發言規範與社群管理員之審核，即可輕易地發表各項文章並討論其內容。然而，文章撰寫者所發表之文章篇幅與詞語用法不盡相同，為維持論壇文章之品質，論壇乃以現有之文字比對技術及發文規範，審核文章之內容，藉此過濾違規文章，但因網路創造甚多奇異詞彙，當文章具有獨特用詞時，將降低社群之文字比對方法過濾違規文章之成；因此，對論壇管理者逐筆審視所有之文章，可從中過濾違規之文章或給予修正，但此舉將花費大量之人力成本與時間；對於文章撰寫者而言，可能因個人疏忽之關係，而於無意識中將帶有偏激之詞語寫入文章內，因而違規而遭致論壇平台或論壇管理者移除，將降低文章撰寫者再發文之意願。

為改善既有論壇審核違規文章機制所面臨之問題，並可立即修正違規文章之內容，藉由對近期相關文獻之借鑑與改良，本研究乃發展一套「提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式」，藉由論壇文章之基礎資訊，結合語句相似度分析技術、自動化情緒隸屬係數建立技術、語

句評閱技術、多重組合句之建立等，推論文章隱含之情緒資訊，並根據偏激情緒之語句重組文章語句之內容以協助論壇使用者與管理者避免文章違反論壇規範之可能性。另一方面，本研究除發展模式與方法論外，亦根據此方法論建構一套以網際網路為基之「提升論壇知識利用價值之論壇文章情感解析及語句結構重組系統」以進行案例驗證，從而確認方法論與技術之可行性。

參考文獻

1. Afraz, Z. S., Muhammad, A. and Ana M. M. E., 2010, "Lexicon based sentiment analysis of Urdu text using SentiUnits," *Advances in Artificial Intelligence*, Vol. 6437, pp. 32-43.
2. Alavi, M. and Leidner, D. E., 2012, "Review knowledge management and knowledge management systems Conceptual foundations and research issues," *MIS Quarterly*, Vol. 25, No. 1, pp. 107-136.
3. Bai, X., 2011, "Predicting consumer sentiments from online text," *Decision Support Systems*, Vol. 50, No. 4, pp. 732-742.
4. Chan, S. W. K., 2006, "Beyond keyword and cue-phrase matching: A sentence-based abstraction technique for information extraction," *Decision Support Systems*, Vol. 42, No. 2, pp. 759-777.
5. Danushka, B., Naoali, O. and Mitsuru, I., 2012, "A preference learning approach to sentence ordering for multi-document summarization," *Information Sciences*, Vol. 217, pp. 78-95.
6. Das, Dipankar. and Bandyopadhyay, S., 2010, "Sentence level emotion tagging on blog and news corpora," *Advances in Pattern Recognition*, Vol. 6256, pp. 332-341.
7. David, D. R., Fernando, M., Ramon, L. H., Stephan, M., Ricardo, G., Cerferino, M. and Francisco, D. P., 2011, "Conflict and cognitive control during sentence comprehension Recruitment of a frontal network during the processing of Spanish object-first sentences," *Neuropsychologia*, Vol. 49, No. 3, pp. 382-391.
8. Ercan, G. and Cicekli, I., 2007, "Using lexical chains for keyword extraction," *Information Processing and Management*, Vol. 43, No. 6, pp. 1705-1714.
9. Fan, N., Cai, W. D., Zhao, Y. and Li, H. X., 2009, "Extraction of sentiment topic sentences of Chinese texts," *Journal of Computer Applications*, Vol. 29, No. 4, pp. 1171-1174.
10. Fang, Y. H. and Chiu, C. M., 2010, "In justice we trust: Exploring knowledge sharing continuance intentions in virtual communities of practice," *Computers in Human Behavior*, Vol. 26, No. 2, pp. 235-246.
11. Fu, X., Liu, G., Guo, Y. and Wang, Z., 2013, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowledge-Based Systems*, Vol. 37, pp 186-195.
12. Ge, S. L. and Chen, X. X., 2009, "Cluster analysis of college English writing in automated essay scoring," *Computer Engineering and Applications*, Vol. 45, No. 6, pp. 145-148.
13. Gregory, P. S. and Daniel, N. A., 2007, "Emotion intensity and categorization ratings for emotional and nonemotional words," *Cognition and Emotion*, Vol. 22, No. 1, pp. 114-133.
14. Gu, Y. and Grossman, R. L., 2010, "Sector: A high performance wide area community data storage and sharing system," *Future Generation Computer Systems*, Vol. 26, No. 5, pp. 720-728.
15. Guido, B. and Leendert, V. D. T., 2006, "Security policies for sharing knowledge in virtual communities," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 36, No. 3, pp. 439-450.
16. Hamouda, A., Marei, M. and Rohaim, M., 2011, "Building machine learning based senti-word lexicon for sentiment analysis," *Journal of Advances in Information Technology*, Vol. 2, No. 4, pp. 199-203.
17. Hao, X. Y., Li, J. H., You, L. P. and Liu, K. Y., 2007, "A research on building of Chinese reading comprehension corpus," *Journal of Chinese Information Processing*, Vol. 21, No. 6, pp. 29-35.
18. Hsu, C. L. and Lin, C. C., 2007, "Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation," *Information & Management*, Vol. 45, No. 1, pp. 65-74.
19. Huang, J., Tian, S. W., Yu, L. and Feng, G. J., 2012, "Sentence sentiment analysis based on Uyghur sentiment word," *Computer Engineering*, Vol. 38, No. 9, pp. 182-185.
20. Ichifuji, Y., Konno S. and Sone, H., 2010, "An advisory method for BBS users and evaluation of BBS comments," *Procedia Social and Behavioral Sciences*, Vol. 2, No. 1, pp. 218-224.
21. Jiang, H., Huang, G. Q. and Liu, J. G., 2011, "The research on CET automated essay scoring based on data mining," *Advances in Computer Science and Education Applications*, Vol. 202, pp. 100-105.
22. Kaila, S., Huang, H., Wu, X. and Zhang, S., 2006, "A logical framework for identifying quality knowledge from different data sources," *Decision Support Systems*, Vol. 42, No. 3, pp. 1673-1683.
23. Kakkonen, T., Myller, N., Sutinen, E. and Timonen, J., 2008, "Comparison of dimension reduction

methods for automated essay grading," *Educational Technology & Society*, Vol. 11, No. 3, pp. 275-288.
24. Ko, Y. and Seo, J., 2008, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognition Letters*, Vol. 29, No. 9, pp. 1366-1371.
25. Lai, L. F., 2007, "A knowledge engineering approach to knowledge management," *Information Sciences*, Vol. 177. No. 19. pp. 4072-4094.
26. Lee, K. J., Choi, Y. S. and Kim, J. E., 2011, "Building an automated English sentence evaluation system for students learning English as a second language," *Computer Speech and Language*, Vol. 25, No, 2 pp. 246-260.
27. Li, C. L., Zhu, Y. H. and Xu, Y. Q., 2011, "Research of attribute word extraction method in Chinese product comment," *Computer Engineering*, Vol. 37, No. 12, pp. 26-29.
28. Li, G. and Liu, F., 2012, "Application of a clustering method on sentiment analysis," *Journal of Information Science*, Vol. 38, No. 2, pp. 127-139.
29. Li, N. and Wu, D. D., 2010, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," *Decision Support System*, Vol. 48, No. 2, pp. 354-368.
30. Li, S., Ye, Q., Li, Y. j. and Luo, S. Q., 2008, "Information inference based sentiment classification for online news comments," *Journal of Chinese Information Processing*, Vol. 23, No. 5, pp. 75-79.
31. Li, Y. M., Liao, T. F. and Lai, C. Y., 2012, "A social recommender mechanism for improving knowledge sharing in online forums," *Information Processing and Management*, Vol. 48, No. 5, pp. 978-994.
32. Liao, C. H., Kuo, B. C. and Pai, K. C., 2012, "Effectiveness of automated Chinese sentence scoring with latent semantic analysis," *The Turkish Online Journal of Educational Technology*, Vol. 11, No, 2 pp. 80-87.
33. Lin, F. R. and Liang, C. H., 2008, "Storyline-based summarization for news topic retrospection," *Decision Support Systems*, Vol. 45, No. 3, pp. 473-490.
34. Lin, Y., Ye, Q., Li, J., Zhang, Z. and Wang, T., 2011, "Snippet-based unsupervised approach for sentiment classification of Chinese online reviews," *International Journal of Information Technology & Decision Making*, Vol. 10, No. 6, pp. 1097-1110.
35. Liu, J., Wang, B., Lu, H. and Ma, S., 2008, "A graph-based image annotation framework," *Pattern Recognition Letters*, Vol. 29, No. 4, pp. 407-415.
36. Liu, W. P., Zhu, Y. H., Li, C. L., Xiang, H. Z. and Wen, Z. Q., 2009, "Research on building Chinese basic semantic lexicon," *Journal of Computer Application*, Vol. 29, No. 10, pp. 2875-2877.
37. Liu, W., Yan, H. L., Xiao, J. G. and Zeng, J. X., 2010, "Solution for automatic web review extraction," *Journal of Software*, Vol. 21, No. 12, pp. 3220-3236.
38. Liu, Z, M. and Liu, L., 2012, "Empirical study of sentiment classification for Chinese microblog based on machine learning," *Computer Engineering and Applications*, Vol. 40, No. 1, pp. 1-4.
39. Matsubara, D., Miki, K., Inouchi, H. and Hoshi, T., 2006, "File management using virtual directory architecture for central managed P2P information sharing system (NRBS)," *Information and Media Technologies*, Vol. 1, No. 1 pp.514-523.
40. Matsumoto, K., Konishi, Y., Sayama, H. and Ren, F., 2011, "Analysis of Wakamono Kotoba emotion corpus and its application in emotion estimation," *International Journal of Advanced Intelligence*, Vol. 3, No. 1, pp. 1-24.
41. Miao, Y., Su, J., Liu, S. and Zhang, J., 2012, "Bootstrapping-based method for chinese sentiment lexicon construction," *International Conference on Information Engineering Lecture Notes in Information Technology*, Vol. 25, pp. 248-253.
42. Oh, H., Fiorito, S. S., Cho, H. and Hofacker, C. F., 2008, "Effects of design factors on store image and expectation of merchandise quality in web-based stores," *Journal of Retailing and Consumer Services*, Vol. 15, No. 4, pp. 237-249.
43. Pang, H. J., 2010, "Text sentiment analysis-oriented commodity review detection," *Journal of Computer Applications*, Vol. 32, No. 7, pp. 2038-2041.
44. Peng, H., Shi, Z. Z., Qiu, L. L. and Chang, L., 2008, "Matching algorithm of semantic web service based on similarity of ontology concepts," *Computer Engineering*, Vol. 34, No. 15, pp. 51-53.
45. Peng, J., Yang D. Q., Tang, S. W., Fu, J. and Jiang, H. K., 2007, "A novel text clustering algorithm based on inner product space model of semantic," *Chinese Journal of Computers*, Vol. 30, No. 8, pp. 1354-1363.
46. Pérez, M. S., Sánchez, A., Robles, V. and Herrero, P., 2007, "Design and implementation of a data mining grid-aware architecture," *Future Generation Computer Systems*, Vol. 23, No. 1, pp. 42-47.

47. Rorissa, A., 2008, "User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory," *Information Processing and Management*, Vol. 44, No. 5, pp. 1741-1753.
48. Tractinsky, N., Cokhavi, A., Kirschenbaum, M. and Sharfi, T., 2006, "Evaluating the consistency of immediate aesthetic perceptions of web pages," *International Journal of Human-Computer Studies*, Vol. 64, No. 11, pp. 1071-1083.
49. Tung, C. M. and Lu, W. H., 2012, "Predict depression tendency of web posts using negative emotion evaluation model," *In ACM SIGKDD Workshop on Health Informatics*.
50. Wang, C. Y., Yang, H. Y. and Chou S. C. T., 2008, "Using peer-to-peer technology for knowledge sharing in communities of practices," *Decision Support Systems*, Vol. 45, No. 3, pp. 528-540.
51. Wang, S. and Noe, R. A., 2010, "Knowledge sharing: A review and directions for future research," *Human Resource Management Review*, Vol. 20, No. 2, pp. 115-131.
52. Wang, S. G. and Li, W., 2010, "Research on sentiment classification problem of sino-Japanese relations forum," *Computer Engineering and Applications*, Vol. 43, No. 32, pp. 174-177.
53. Wang, S. G., Li, D. Y., Wei, Y. J. and Song, X. L., 2009, "A synonyms based word sentiment orientation discriminating," *Journal of Chinese Information Processing*, Vol. 23, No. 5, pp. 68-74.
54. Wang, X. D., Liu, Q. and Tao, X. J., 2010, "Sentiment ontology construction and text orientation analysis," *Computer Engineering and Applications*, Vol. 46, No. 30, pp. 117-120.
55. Wang, X. D., Liu, Q. and Zhang, Z., 2011, "Topic semantic orientation compute based on sentiment words ontology," *Computer Engineering and Applications*, Vol. 47, No. 27, pp. 147-151.
56. Wang, Z. H. and Jiang, W., 2012, "Online reviews sentiment analysis model based on rough sets," *Computer Engineering*, Vol. 38, No. 16, pp. 1-4.
57. Win, K. T. and Zhang, M., 2006, "Enhancing information quality through a semantic approach in health information retrieval," *International Journal of Electronic Business Management*, Vol. 4, No. 1, pp. 8-15.
58. Xie, L., Zhou, M. and Sun, M., 2012, "Hierarchical structure based Hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction," *Journal of Chinese Information Processing*, Vol. 26, No. 1 pp. 73-83.
59. Xu, C., Wang, M., He, T. T. and Zhang, Y., 2008, "Automatic summarization method based on extracting sentences from local topics," *Computer Engineering*, Vol. 34, No. 22, pp. 49-51.
60. Yang, F, Peng, Q. K. and Xu, T., 2010, "Sentiment classification for online comments based on random network theory," *Acta Autimatica Sinica*, Vol. 36, No. 6, pp. 837-844.
61. Yang, J., Peng, S. Y. and Hou, M., 2011, "Recognizing sentiment polarity in Chinese reviews based on top sentiment sentences," *Application Research of Computers*, Vol. 28, No. 2, pp. 569-572.
62. Yin, C. X. and Peng, Q. K., 2012, "Identifying word sentiment orientation for free comments via complex network," *Acta Automatica Sinica*, Vol. 38, No. 3, pp. 389-390.
63. Yu, Z. T., Fan, X. Z., Guo, J. Y. and Geng, Z. M., 2006, "Answer extraction for Chinese question-answer system based on latent semantic analysis," *Chinese journal of Computers*, Vol. 29, No. 10, pp. 1888-1893.
64. Zan, H. Y., Zuo, W. S., Zhang, K. L. and Wu, Y. F., 2011, "Sentiment analysis based on rules and statistics," *Computer Engineering and Science*, Vol. 33, No. 5, pp. 146-150.
65. Zhan, P., Zhao, Z. W. and Zhuo, J. W., 2011, "Method of sentence semantic orientation distinction based on semantic weighted sentiment word," *Computer Engineering and Applications*, Vol. 47, No. 35, pp. 161-164.
66. Zhang, C., Zeng, D., Li, J., Wang, F. Y. and Zuo, W., 2009, "Sentiment analysis of Chinese documents: From sentence to document level," *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 12, pp. 2474-2487.
67. Zhang, J., 2009, "An overview of factors influencing knowledge sharing in virtual communities," *Journal of Modern Information*, Vol. 29, No. 7, pp. 222-225.
68. Zhang, Y. X., Zhang, M. and Deng, Z. H., 2009, "Feature-driven summarization of customer reviews," *Journal of Computer Research and Development*, Vol. 46, No. 2, pp. 520-525.
69. Zhao, Y. Y., Bing, Q. and Ting, L., 2010 "Integrating intra- and inter-document evidences for improving sentence sentiment classification," *Acta Automatica Sinica*, Vol. 36, No. 10, pp. 1417-1425.
70. Zhou, J., Lin, C. and Li, B. C., 2010, "Research of sentiment classification for Netnews comments by machine learning," *Journal of Computer Application*, Vol. 30, No. 4, pp. 1011-1014.

出席國際學術會議心得報告

106 年　　07 月　　10 日

計畫編號	MOST 105-2221-E-343-003
計畫名稱	提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式(I)
出國人員姓名 服務機關及職稱	楊士霆 南華大學資訊管理學系 助理教授
會議時間地點	July 4-7, 2017. Venue: Osaka, Japan
會議名稱	International Conference on Innovation and Management (IAM2017S)
發表論文題目	A Domain Knowledge Document Retrieval Platform

一、參加會議經過與心得

此次 2017 International Conference on Innovation and Management（IAM2017S）研討會乃安排於日本（Japan）之大阪（Osaka）舉辦，研討會舉辦時間為 7/4 至 7/7 共四天，配合研討會主辦單位之行程規劃與可行機位安排，個人於 07/02 上午七點，即出發前往桃園國際機場，進行登記作業且於上午 09:10 由桃園國際機場起飛，並於下午 12:50 抵達關西空港航站 Kansai Airport Station 國際機場，辦理入境手續後即搭乘「關西特急（Haruka）」至京都，並於當地下午 15：30 左右抵達住宿飯店「京都飯店（京都四条)」(Hotel Mystays)，台北與大阪時差為 1 小時；此程搭機、轉乘地鐵時間總計七小時左右，故辦理 Check in 手續並稍做休息後，於本日及 07/02、07/03 日上、下午即參觀京都伏見稻荷大社、河原町通、新京極、清水寺、寧寧之道、高台寺、圓山公園、八板神社(祇園路上)等周遭景點，以瞭解日本大阪、京都之交通、飲食習慣、友善的風俗民情及北海道之建築、古蹟等風貌。

7/04 上午隨即搭乘地鐵，前往大阪蒙特利格拉斯米爾飯店（Hotel Monterey Grasmere Osaka)，由於個人於大阪住宿在「蒙特利格拉斯米爾」，離研討會會場（RIHGA Royal Hotel Osaka）需搭「地下鐵」千日前線，從大阪難波站至阿座波站，以及步行，共需 30 分鐘交通時間，因此在辦理 Check in 後，下午隨即前往研討會會場，並利用時間於周遭參觀（如大阪天王寺、新世界區等)。再者，於 07/05 上午 09：30 隨即前往研討會飯店報到，完成報到手續（如圖 1 至圖 3 所示)，並與國外與會學者、專家進行互動。

此次研討會之規模可算是小型且精緻規模，研討會總計發表篇數約為 60 篇左右，議程數為 6 個左右，皆為 Oral 議程。大會正式行程日期為 7/4 至 7/7 四日，7/4 乃為提前報到日期，正式發表日期亦為 7/5 至 7/7 三日。本次研討會內容乃安排與此次會議主題相關之企業、組織、電子商務的管理、科技創新與發展專題演講與論文發表，再依不同論文主題每天分至 2 個時

段進行發表。個人的論文被安排於發表日 7/5 的下午(13:30-16:00)場次(編號 P0103)「Session B」發表，由於此研討會主題乃著重於企業的管理、科技創新與發展，個人研究（資料探勘、知識管理、系統開發）與部份學者甚為相似，故於發表後其他學者亦表示對此研究的高度興趣，詢問本研究之網頁（知識文件）探勘技術、問答解析（詞彙解析）分析與應用、系統智慧推論技術與其他研究之差異，個人並作完整回答，互動甚佳。此外個人亦參加多場與研究興趣較相關之發表場次，並對於其他學者發表內容提出詢問，對於知識管理、資訊科技管理等課題觸發新的研究靈感（如圖 3 至圖 8 所示）。

除會議發表時間外，在其他交流活動時，個人與國際/國內學者亦有良好交流，於此次研討會認識多位先進，藉由討論瞭解許多國際/國內工業工程、資訊管理學者之研究方向，並規劃未來合作之可能作法，收穫極大，此對於個人學術經歷尚屬資淺而言，乃一大助益。

圖 1、抵達 IAM2017S 會場並註冊(1)

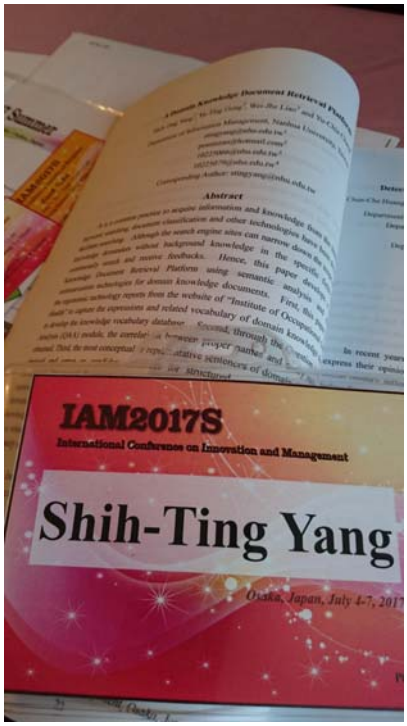

圖 2、抵達 IAM2017S 會場並註冊(2)

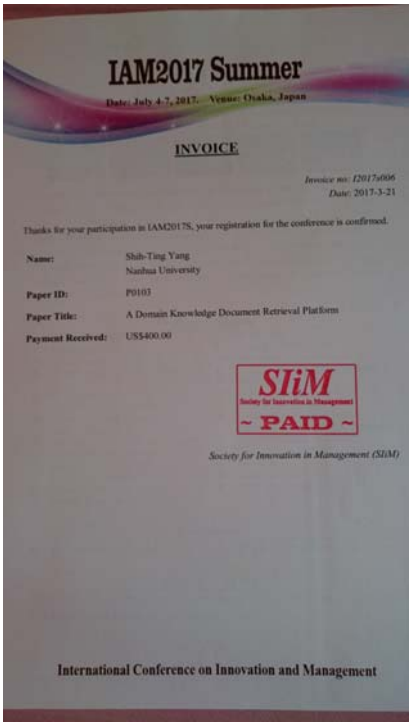

圖 3、抵達 IAM2017S 會場並註冊(3)

圖 3、論文發表與研討(1)

圖 4、論文發表與研討(2)

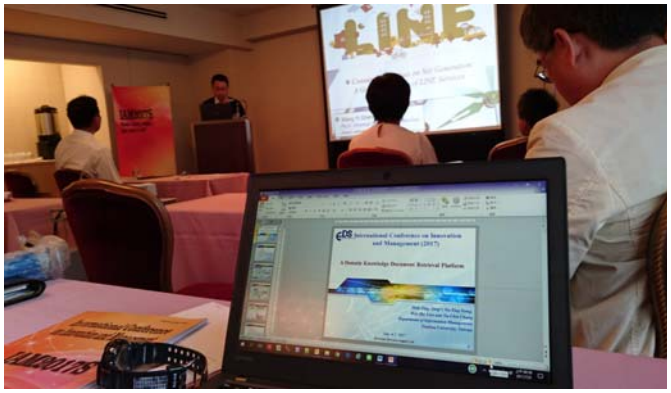

圖 3、論文發表與研討(1)

圖 4、論文發表與研討(2)

圖 5、論文發表與研討(3)

圖 6、論文發表與研討(4)

待研討會圓滿結束後，個人於當日（7/7）即搭機場巴士前往關西空港航站 Kansai Airport Station 國際機場，並搭機回桃園國際機場，結束此次 IAM2017S 學術研討活動。

三、建議

此次會議中的各項活動安排都可發現主辦單位頗為用心，對於遠道造訪之學者給予多項貼心之服務，為國內學校爭取主辦國際型研討會可加以參考之長處。然而，雖然主辦單位之用心可見，由於此次研討會乃屬小型（精緻型）之規模，雖然各與會學者之於會場中研討之熱絡，然相較於一般中大型國際研討會之會場可能規畫數個地點，或者數個樓層，此次研討會僅舉辦於 RIHGA Royal Hotel Osaka 之六樓，故需受限於飯店之場地限制，如各議程場地較為狹小、且無提供休息區，空間規劃皆不甚完美，此可提供國內學者於辦此類中、大型學術研討會之借鏡。

整體而言，本次大會舉辦頗為用心，個人於此行收穫豐富，且結識多位國際學者，希望能於未來建立更長遠的交流與合作。

四、攜回資料名稱及內容

1. 研討會論文集：含議程集 1 本、論文摘要集 1 本、論文全文電子檔（光碟一張）。
2. 國內外學者學術交流名片。

A Domain Knowledge Document Retrieval Platform

Shih-Ting Yang[1], Yu-Ting Gong[2], Wei-Jhe Liao[3] and Yu-Chia Chang[4]

Department of Information Management, Nanhua University, Taiwan
stingyang@nhu.edu.tw[1]
possiezan@hotmail.com[2]
10225066@nhu.edu.tw[3]
10225079@nhu.edu.tw[4]
Corresponding Author: stingyang@nhu.edu.tw

Abstract

It is a common practice to acquire information and knowledge from the Internet; thus, keyword searching, document classification and other technologies have been developed to facilitate searching. Although the search engine sites can narrow down the scope of search, knowledge demanders without background knowledge in the specific fields need to continuously search and receive feedbacks. Hence, this paper develops a Domain Knowledge Document Retrieval Platform using semantic analysis and document summarization technologies for domain knowledge documents. First, this paper analyzes the ergonomic technology reports from the website of "Institute of Occupational Safety and Health" to capture the expressions and related vocabulary of domain knowledge documents to develop the knowledge vocabulary database. Second, through the Question and Answer Analysis (QAA) module, the correlations between proper names and query strings can be obtained. Third, the most conceptual or representative sentences of domain documents can be derived and serve as candidate sentences for structured summarization. Finally, the Document Structured Summarization (DSS) module is used to calculate and retrieve representative sentences of the documents and integrate them into summary for knowledge demanders. In order to demonstrate applicability of the proposed methodology, a web-based knowledge document retrieval system is also established based on the proposed methodology. As a whole, this research provides an approach for knowledge demanders to efficiently and accurately acquire the domain knowledge documents.

Keywords: Institute of occupational safety and health, knowledge management, data mining, document summarization technology, semantic analysis

1. Introduction

As the Internet becomes a popular source of information acquisition, many researches have been conducted and technologies have emerged to help users searching for and accessing information quickly and efficiently. To help users obtaining information conveniently, websites have been established to collect literature of related fields; these websites are representative in the given and specified fields. In other words, when users are searching for information, they conduct searches on the website intuitively. Although webpages at present allow users to obtain information by search topic or keywords, the search may not be successfully if the users lack domain knowledge in the specified field. In addition, as contents are in free format without restrictions in the number of words, document filtering of users may be affected by personal browsing and reading preferences, thereby reducing the knowledge sharing function of these websites.

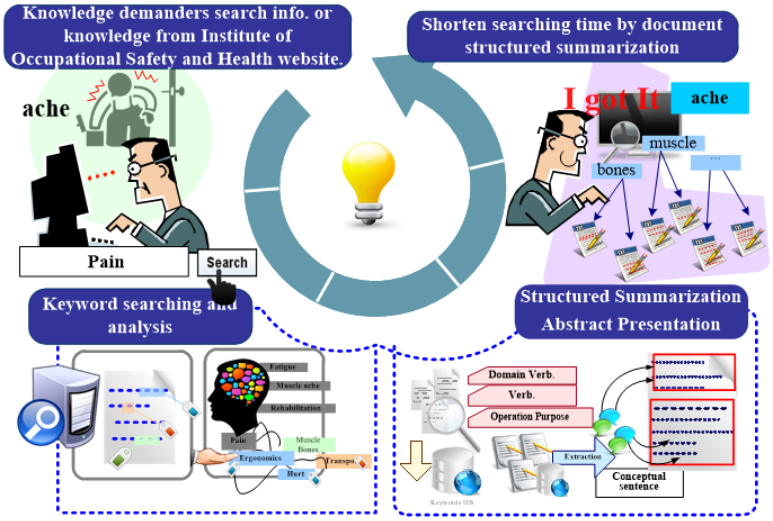

Figure 1: To-Be model

To solve above problem, this paper proposes that the methodology of user filtering in the search technology and document abstract presentation should be enhanced in order to help users to rapidly and efficiently obtain the documents needed. Based on the website of Institute of Labor, Occupational Safety and Health, this paper analyzes the knowledge document expressions and related vocabulary regarding the research reports (documents) or technical books on the knowledge websites, and establishes the knowledge vocabulary set. In this way, semantic association with the search word series can thus be established to enhance the keyword semantic search technology. In addition, the conceptual sentences of the documents can be obtained by using the knowledge vocabulary library. Meanwhile, structured summarization based on the document structure can be established to help users quickly determining the documents and avoiding the impact of personal reading preferences on the filtering of documents. The proposed domain knowledge document retrieval

methodology, combining semantic parsing and document summarization technology, can strengthen the search word series semantic determination by using the representative vocabulary of the document. The proposed platform can also avoid the impact of personal reading preferences on document filtering by using the concept of structured summarization, and thus enhancing the knowledge sharing effectiveness of the specified field website. The To-Be model can be shown in Figure 1.

2. Literature Review

2.1 Q&A application and technology

Regarding the topic of Q&A application and technology, this paper conducts literature review relating to Q&A application and Q&A technology, expecting to observe the different perspectives and dimensional analysis for different types for understanding of the characteristics of Q&A application and technology.

The types of Q&A application can be divided into the Q&A system and retrieve system. Regarding Q&A system, Oh et al. (2011) proposed a compositional Q&A system using criteria judgment for question analysis. The question format (single or multiple question items), subject, question limitations (time or location) are used as the judgment criteria to learn about the types and formats of feedback sentences. Cao et al. (2010) established an online Q&A system (AskHERMES) for medical clinical reports to capture the key points of the complex clinical reports without fixed format. Regarding information retrieval, Huang et al. (2006) proposed a composite relational model to capture biomedical literature by using the shallow parsing to develop the grammatical and semantic structure, and using the greedy method for matching to acquire the theme of the biomedical literature through the training model. According to the document association and common features' BE (Basic Element), Teng et al. (2010) established a user-oriented document summarization retrieval system.

In terms of Q&A technology, this paper categorized various considerations. The factors for consideration may be based on subject, document characteristics or the semantics for analysis. Concerning subject for analysis, Oh et al. (2012) proposed a Q&A learning mechanism to analyze the structure through the existing Q&A documents, and used the word meaning disambiguation for semantic analysis to obtain the combinations of questions and answers (answer format, answer subject, target and expected answer content). Han et al. (2007) determined question types to establish various types of relevant vocabulary, allowing the users to determine the problem targets and analyze the question retrieval category, in order to expand the question. Jones and Love (2007) argued that if the relationships of the documents are more similar, it means that there is a common role in between the two documents. Through the background environment, with the relationship as the matching criteria, the common relationship of the documents can be obtained. Ko et al. (2004) used the important sentences as the basis for document classification in order to enhance document

classification effective. For semantics, Dorr and Gaasterland (2007) proposed a composite model considering tense and semantic relationship to associate relevant events based on time sequence relationship and event viewpoints. Dunlavy et al. (2010) proposed an integrated information question system to conduct the relevant question analysis according to main sentences with characteristic market documents, such as the sentence location and document content, by the potential semantic index technology.

2.2 Summarization application and technology

For the summarization application fields, most applications are in the fields of too much data or too many similar thematic documents, such as in the medical field, the online news and legal field. Regarding the problem of too much literature in the medical field, Elhadad et al. (2005) established a uniform summarization model based on search and retrieval technology to summarize the abstracts of the results, thus making the browsing process more efficiency. To overcome the limitation of textual length of news headlines, Zajic et al. (2007) used the document compression technology to form the reserve sentences from the captured abstract sentences from a single document and summarize the multi-document summarizations.

The summarization establishment technology is discussed in the "supervised machine learning" model and the "non-supervised machine learning" model. Regarding the supervised learning, Bollegala et al. (2010) used the support vector machine for classification to group the sentences of two documents for sequence and abstract. Li and Chen (2010) used the statistical language model to determine the documents correlation, the probability sequencing, and potential Markov chain model to capture the representative segments. Through the assumed sentences combined with statistical computation, Jung et al. (2005) proposed an automatic summarization model, which can establish the assumed sentences by linking words combining the neighboring sentences and compute the importance of documents according to titles and locations. The research also conducted the clustering analysis pattern to obtain the similarity and form the abstract. Regarding the non-supervised learning, based on the potential semantic analysis, Chan (2006) proposed a quantitative model to capture the most representative sentences, which can strengthen human understanding models through potential semantics, and organize the most representative or linkage vocabulary into the lexicological network to present the association between sentences of the documents. Steinberger et al. (2007) established the automatic summarization system based on the potential semantic analysis and proposed the limitation of the number of words of the summarization by document compression percentage. Ko et al. (2003) proposed a subject-based document summarization technology upon the vocabulary clustering, which conducts the textual association analysis of the document and present by using the spatial vectors to obtain the vocabulary clusters and determine the cluster core to produce subject

and keywords. The compression percentage was used to capture fixed sentences to form the summarization.

3. Domain Knowledge Document Retrieval Methodology

The proposed domain knowledge document retrieval methodology uses the technical books and research reports on the professional website as the basis for analysis. This paper analyzes the structural characteristics of the knowledge documents, and proposed eight expression items and 28 detailed expression items while establishing the knowledge vocabulary set. Through knowledge vocabulary set, this paper obtains the terminological terms, proper names for semantic association, analyzes the main question words based on search word series, and determines the associated or related semantic words. The corresponding knowledge document can be found to enhance retrieval accuracy. In addition, the vocabulary rules can be established on the basis of knowledge vocabulary set to obtain the conceptual sentences and capture the representative sentences of the documents, according to the rules of structured summarization. In other words, the development of a formal knowledge document can clarify and completely express the report contents. Therefore, the main methodology can be divided into the following parts as shown in Figure 2, including Part 1 Knowledge Document Expression Item Analysis module, Part 2 Question and Answer Analysis (QAA) module, and Part 3 Document Structured Summarization (DSS) module.

3.1 Knowledge Document Expression Item Analysis Module

This paper consults ergonomics staffs and summarizes the repetitive or important descriptive words in the improvement reports and technical books for the establishment of expression items of the knowledge documents. After the analysis of the knowledge documents, the establishment of expression items and the capturing of the conceptual sentences, the expression items and the details of the detailed expression items are illustrated. The main calculations include Step (A1)-Establishment of the expression items of the knowledge documents and Step (A2)-Conceptual Sentence Acquisition.

3.2 Question and Answer Analysis (QAA) module

As most of the query strings input by the users are intuitive or colloquial query string words of the users and do not belong to the domain vocabulary of the knowledge document. The domain vocabulary search is used to find out the relevant knowledge document; on the contrary, the self-defined query strings (i.e., colloquial query string) may have no clear definitions, it may result in finding some irrelevant documents. Hence, to enhance the natural language search flexibility, this paper proposes a knowledge document Q&A Analysis (QAA) module to conduct the analysis of the main question words of the colloquial query strings of the users, and find out the semantic words of association by matching and parsing of question words and answer words, and thus capturing the corresponding knowledge

documents and enhancing the retrieval accuracy. The main calculations include Step (B1)-Determination of the implied goals of the question words and Step (B2)-Determination of the vocabulary category similarity.

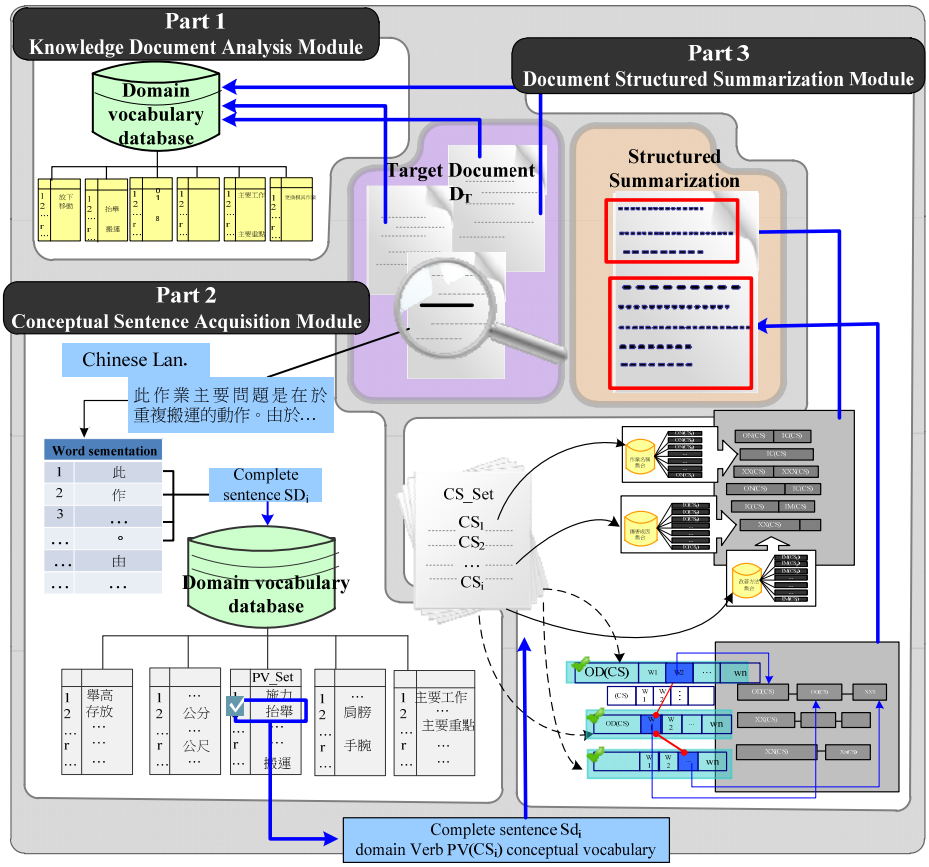

Figure 2: Architecture of domain knowledge document retrieval platform

3.3 Document Structured Summarization (DSS) Module

For the completeness of the textual descriptions, the DSS module can be divided into two parts including the establishment of brief part and the detailed part.

3.3.1 Establishment of the Brief Part

The brief part of the structured summarization is mainly to calculate the centrality of the sentences before carrying out the selection by sentence structural integrity. If the sentence contains a variety of conceptual vocabularies, it means the sentence is representative of the text, and thus the sentence is listed as a candidate sentence for sentence structural strength calculation. The sentence structural strength calculation is to consider the readability of the abstract, hence, the sentence structural strength (namely, the relevance between the subjective, predictive, and verb of the sentence) is calculated to acquire sentences of completeness. The set of the acquired sentences include: the operation name set, the injury cause set and the improvement method set. The main calculations include Step (C1): Calculation of the Centrality of Conceptual Sentences, Step (C2): Calculation of Structural Strength of Conceptual Sentences and Step (C3)-Calculation of the Weights of the Conceptual Sentences

3.3.2 Establishment of the Detailed Description Part

The contents of the detailed description part include the description part and the assessment part with detailed descriptions of operation, operation environment and operation hour. According to the above, the set of the sentences in the description part includes: operation set and operation environment set, and the assessment part acquires mainly the set of improvement methods. The description part is mainly of the operation and operation environment sets with implied vocabularies including: operation goal, domain verb, force application level, operation title and frequency; the assessment part is mainly of the improvement method followed by injury cause, operation tool, assessment verb vocabularies. The main calculations include Step (D1): Acquisition of Conceptual Sentences with Operation Definition Vocabulary, Step (D2): Calculation of Mutual Dependence of Conceptual Vocabularies and Step (D3): Calculation and Acquisition of Conceptual Sentences

4. Domain Knowledge Document Retrieval Platform

In order to verify the feasibility of the domain knowledge document retrieval platform in the practical application, this paper uses the improvement reports on the website of "Institute of Labor, Occupational Safety and Health" as the case verification samples, and the core functional modules of the knowledge document retrieval platform to evaluate the feasibility of the proposed methodology and the developed platform. The users can realize the function of "knowledge document uploading" through "knowledge document management module". After the knowledge document uploading, through the "knowledge document expression item analysis module", the platform can capture the conceptual sentences of various expression items of the knowledge document as shown in Figure 3 and Figure 4. According to the relevant conceptual sentences captured by the expression items, such as the conceptual sentences of the expression item of "operation identity" including "…the joint lifting by the operators…", the platform can obtain the keywords of the document such as "storage box" and "handling", based on which the platform can analyze the question goal and the relevant answer word combinations. As shown in Figure 5, according to the parsing of the question word, answer sentences and answer words matching, the platform can obtain the relevant answer words of the question word "pain" such as "construction industry" and "age", as shown in Figure 6. Furthermore, the platform can connect relevant documents according to the question goal of "pain", and the related answer words of "construction industry" and "age" via question and answer analysis (QAA) module, or select the documents through structured abstract. For document structured summarization (DSS) module, in the case of the brief structured abstract, the platform may compute the centrality score and the structural score. For example, the sum of the weighted score of "the operators do not need to bend…" is "12". Then, according to the weight, the simplified structured abstract can be established according to "document name", and "operation name" (see Figure 7). Regarding the

establishment of the detailed structured abstract, through the coefficient, it can form the vocabulary chain including “warehouse, transportation, and other main problems”. The detailed structured abstract can be obtained according to “operation description”, “problem description”, and “way of improvement” (see Figure 8).

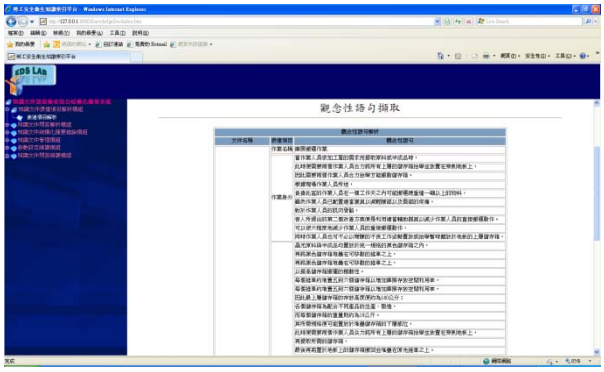

Figure 3: Document expression items analysis result (1)

Figure 4: Document expression items analysis result (1)

Figure 5: Q&A analysis result (1)

Figure 6: Q&A analysis result(2)

Figure 7: Brief structured abstract result

Figure 8: Detailed structured summarization result

5. Conclusions

As the specific knowledge fields are too professional, general users can hardly know and define the key words. As a result, the document search will take more time and have more obstacles. In addition, document abstracts are often presented in free form without the

limitations on number of words. Therefore, it can easily affect the selection of documents due to personal browsing and reading preferences, and thus reducing the knowledge sharing function of the knowledge websites. Hence, this paper proposes a domain knowledge document retrieval methodology and platform. The methodology and platform can process knowledge document semantic Q&A and realize the semantic association of general vocabulary and professional vocabulary through Q&A semantic parsing. Hence, users can search the general vocabulary and get the relevant knowledge documents and then convert the knowledge document in free form into formatted document abstract by way of structured abstract to enhance the reading and selection of users. In this way, the proposed platform can help users to access to the information on the professional website of Institute of Occupational Safety and Health, and thus enhancing the knowledge field search effectiveness.

References

Bollegala, D., Okazaki, N., & Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing and Management*, *46(1)*, 89-109.

Cao, Y. G., Liu, F., Simpson, P., Antieau, L., Bennett, A. Cimino, J. J., Ely, J., & Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, *44(2)*, 277-288.

Chan, S. W. K. (2006). Beyond keyword and cue-phrase matching: A sentence-based abstraction technique for information extraction. *Decision Support Systems, 42(2),* 759-777.

Dorr, B. J., & Gaasterland, T. (2007). Exploiting aspectual features and connecting words for summarization-inspired temporal-relation extraction. *Information Processing and Management, 43(6)*, 1681-1704.

Dunlavy, D. M., O'Leary, D. P., Conroy, J. M. and Schlesinger, J. D. (2007). QCS: A system for querying, clustering and summarizing documents. *Information Processing and Management, 43(6)*, 1588-1605.

Elhadad, N., Kan, M. Y., Klavans, J. L., & McKeown, K. R. (2005). Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine, 33(2)*, 179-198.

Han, K. S., Song, Y. I., Kim, S. B., & Rim, H. C. (2007). Answer extraction and ranking strategies for definitional question answering using linguistic features and definition terminology. *Information Processing & Management, 43(2)*, 353-364.

Huang, M., Zhu, X., & Li, M. (2006). A hybrid method for relation extraction from biomedical literature. *International Journal of Medical Informatics, 75(6)*, 443-455.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology, 55(3)*, 196-231.

Jung, W., Ko, Y., & Seo, J. (2005). Automatic text summarization using two-step sentence extraction. *Lecture Notes in Computer Science, 3411*, 71-81.

Ko, Y., Kim, K., & Seo, J. (2003). Topic keyword identification for text summarization using lexical clustering. *IEICE transactions on information and systems, 86(9)*, 1695-1701.

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management, 44(1)*, 65-79.

Li, Q., & Chen, Y. P. (2010). Personalized text snippet extraction using statistical language models. *Pattern Recognition, 43(1)*, 378-386.

Oh, H. J., Myaeng, S. H., & Jang, M. G. (2012). Effects of answer weight boosting in strategy-driven question answering. *Information Processing and Management, 48(1)*, 83-93.

Oh, H. J., Sung, K. Y., Jang, M. G., & Myaeng, S. H. (2011). Compositional question answering: A divide and conquer approach. *Information Processing and Management*, *47(6)*, 808-824.

Steinberger, J., Poesio, M., Kabadjov, M. A., & Jezek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management, 43(6)*, 1663-1680.

Teng, C., Xiong, N., He, Y., Yang, L. T., & Liu, D. (2010). A behavioural mode research on user-focus summarization. *Mathematical and Computer Modelling, 51(7-8)*. 985-994.

Zajic, D., Dorr, B. J., Lin, J., & Schwartz, R. (2007). Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management, 43(6)*, 1549-1570.

105年度專題研究計畫成果彙整表

計畫主持人：楊士霆	**計畫編號：**105-2221-E-343-003-
計畫名稱：提升論壇知識利用價值之論壇文章情感解析及語句結構重組模式(I)	

成果項目					量化	單位	質化 （說明：各成果項目請附佐證資料或細項說明，如期刊名稱、年份、卷期、起訖頁數、證號...等）
國內	學術性論文	期刊論文			0	篇	
		研討會論文			3		第28屆國際資訊管理學術研討會(ICIM 2017)
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		
		積體電路電路布局權			0		
		著作權			0		
		品種權			0		
		其他			0		
	技術移轉	件數			0	件	
		收入			0	千元	
國外	學術性論文	期刊論文			1	篇	International Journal of Data Warehousing and Mining (IJDWM)
		研討會論文			1		International Conference on Innovation and Management (IAM2017 Summer), Osaka, Japan, July 4-7. Paper ID: P0103.
		專書			0	本	
		專書論文			0	章	
		技術報告			0	篇	
		其他			0	篇	
	智慧財產權及成果	專利權	發明專利	申請中	0	件	
				已獲得	0		
			新型/設計專利		0		
		商標權			0		
		營業秘密			0		

		積體電路電路布局權	0		
		著作權	0		
		品種權	0		
		其他	0		
	技術移轉	件數	0	件	
		收入	0	千元	
參與計畫人力	本國籍	大專生	7	人次	廖偉哲、張育嘉、張婉瑄、蔡旻晉、徐櫻綺、廖瑜哲、謝函恩
		碩士生	0		
		博士生	0		
		博士後研究員	0		
		專任助理	0		
	非本國籍	大專生	0		
		碩士生	0		
		博士生	0		
		博士後研究員	0		
		專任助理	0		
其他成果 （無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。）					

科技部補助專題研究計畫成果自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現（簡要敘述成果是否具有政策應用參考價值及具影響公共利益之重大發現）或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估
   ■達成目標
   □未達成目標（請說明，以100字為限）
   　□實驗失敗
   　□因故實驗中斷
   　□其他原因
   說明：

2. 研究成果在學術期刊發表或申請專利等情形（請於其他欄註明專利及技轉之證號、合約、申請及洽談等詳細資訊）
   論文：■已發表　□未發表之文稿　□撰寫中　□無
   專利：□已獲得　□申請中　■無
   技轉：□已技轉　□洽談中　■無
   其他：（以200字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以500字為限）
   (A)理論方法層面
   (A.1)以Tung與Lu（2010）、向量空間模型、Miao等人（2012）以及為鑒借，建立自動化情緒詞彙庫、文章代表事件之解析、語句相似度之解析、文章情緒類別之解析、文章評閱分數解析，分析得知目標論壇文章之情緒類別與語句重組之內容，以協助論壇管理者快速取得違規之文章。
   (A.2)藉由Chen等人（2010）與向量空間模型之鑒借，經本研究之改良後，整合並發展出一套適用於發文者文章語句重組模組，藉以分析文章評閱分數，並自動多重組合句以重組文章語句之內容，以能針對違規但具有獨特看法之文章重組其內容，進而保留有意義之文章。
   (B)實務應用層面
   (B.1)於論壇管理者層面，可針對特定情緒之文章可快速取得違規文章並將改寫違規之語句，以節省逐筆審核違規文章之時間。
   (B.2)於文章撰寫者層面，可自動將違規文章之語句進行修正，以節省修改違規文章之時間與人力，並可立即得知文章違規之可能性。
   (B.3)於整體論壇之層面，可協助論壇管理者快速過濾違規之文章，並避免文章撰寫者撰寫之文章觸犯社群規範，進而促使使用者持續使用之意願。

4. 主要發現

本研究具有政策應用參考價值：■否　□是，建議提供機關

（勾選「是」者，請列舉建議可提供施政參考之業務主管機關）

本研究具影響公共利益之重大發現：□否　□是

說明：（以150字為限）