科技部補助專題研究計畫成果報告
期末報告

整合知識問答解析與文件結構化摘要技術之知識文件檢索模式(I)

計 畫 類 別 ：個別型計畫
計 畫 編 號 ：NSC 102-2221-E-343-003-
執 行 期 間 ：102 年 08 月 01 日至 103 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：碩士班研究生-兼任助理人員：陳俞佑
碩士班研究生-兼任助理人員：蔡耀昌
大專生-兼任助理人員：鍾晏寧
大專生-兼任助理人員：李後揚
大專生-兼任助理人員：韓進富
大專生-兼任助理人員：莊翔智

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：
1.公開資訊：本計畫可公開查詢
2.「本研究」是否已有嚴重損及公共利益之發現：否
3.「本報告」是否建議提供政府單位施政參考：否

中 華 民 國 103 年 10 月 30 日

中文摘要：近年來，人們已習慣以網路進行資訊與知識之獲取，卻容易造成資訊過量等問題產生，因此延伸出關鍵字搜尋、文件分類等相關研究與技術並協助知識搜尋者能快速且有效取得資訊。雖有資訊搜尋窗口以縮小搜尋範圍，然而面對特定領域網站時，若無相關領域背景之知識搜尋者仍需不斷嘗試並取得回饋，關鍵字搜尋與文件分類乃缺乏有效地協助。因此，本研究乃以「勞工安全衛生知識網」為基礎，針對特定領域知識文件發展一套「整合知識問答解析與文件結構化摘要技術之知識文件檢索」模式，以協助知識搜尋者有效地篩選資料與知識。

本研究所建立「整合知識問答解析與文件結構化摘要技術之知識文件檢索」模式乃包含「知識文件表達項目解析模組」、「知識文件問答解析模組」與「知識文件結構化摘要推論模組」。首先，「知識文件表達項目解析模組」乃針對「勞工安全衛生知識網」之人因工程工作場所研究報告與技術叢書進行文件表達項目解析，以擷取此類型知識文件之表達方式、呈現內容以及相關性語彙，進而建構知識語彙庫並取得文件之觀念性語句，以作為知識文件問答解析之關鍵字語意比對與結構化摘要推論之依據。其次，「問答解析模組」乃進行文件關鍵字詞與搜尋字串語意相似性分析，透過語意相似性進而取得確實文件，最後，「結構化摘要模組」乃計算並擷取具有文件之代表性語句，並以制式摘要方式呈現於搜尋者，以避免個人閱讀偏好影響文件之篩選，進而加強勞工安全衛生知識網搜尋技術以提高知識網知識分享之成效。本研究除發展模式與方法論外，並依此方法論建構一套「整合知識問答解析與文件結構化摘要技術之知識文件檢索」系統以進行案例驗證，從而確認方法論與技術之可行性。

整體而言，藉由知識語彙庫建立並透過觀念性語句擷取，並整合關鍵字索引與結構化摘要，以提供更方便、有效地方式解除知識搜尋者於專業領域中搜尋瓶頸，並能藉以上述整合技術更容易取得專業知識。

中文關鍵詞：勞工安全衛生知識網、知識管理、資料探勘、文件摘要技術、語意分析

英文摘要：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 Knowledge Document Retrieval model based on question and answer analysis and document structured 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, based on Conceptual Vocabulary Determination (CVD) module, 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. It is expected that knowledge demanders can directly read the desired parts according to problems to ensure they can find document they want within a short time. In order to demonstrate applicability of the proposed methodology, a web-based knowledge document retrieval system is also established based on the proposed model. Furthermore, the knowledge documents (i.e., ergonomic technology reports) from the website of "Institute of Occupational Safety and Health" are applied as examples to evaluate the proposed model. As a whole, this research provides an approach for knowledge demanders to efficiently and accurately acquire the domain knowledge documents.

英文關鍵詞： Institute of Occupational Safety and Health, Knowledge Management, Data Mining, Document Summarization Technology, Semantic Analysis

一、報告內容

1. 研究動機與目的

由於網際網路發達，人們已習慣以網路進行資訊與知識之獲取，卻容易造成資訊過量等問題產生，因此延伸出相關研究與技術以協助知識搜尋者能快速且有效取得資訊。為協助知識搜尋者方便取得資訊，多數解決方案乃建立網站將相關領域文件彙整並提供給搜尋者分享，此類網站即具有該領域之代表特性，亦表示知識搜尋者欲查詢相關資訊時，期望可以直覺方式於該網站查詢所需資訊，進而節省資訊搜尋時間。

針對上述，雖有特定領域網站作為資訊搜尋窗口以縮小搜尋範圍，但由於網路資源龐大，從中仍需進行文件分類、關鍵字查詢等方式以準確獲取所需文件。本研究以「勞工安全衛生知識網」為例，該網站彙集各行業職業傷害防制等勞工安全相關知識與文件，雖已針對主題進行文件分類，但主題下所含眾多議題之文件，導致無法立即判斷是否為所需資料，該知識網雖有關鍵字搜尋方式以進行查詢，但知識文件多為專業人員進行訪視、研究後撰寫，其文件內容大多以專業名詞、專業知識描述。因此，若無相關領域背景之知識搜尋者以關鍵字搜尋方式亦無法確切取得回饋，或是知識搜尋者無法定義關鍵字詞反而以揣測方式進行搜尋。此外，搜尋者必須透過文件標題先行篩選再以閱讀摘要方式判定文件取決，且該網頁之文件摘要呈現方式乃以作者自行撰寫，其自由形式呈現方式易影響搜尋者判斷文件取得。

根據上述內容，雖然目前網頁針對文件主題或是定義關鍵字以協助知識搜尋者取得資訊。然而，針對特定領域知識網若以此方式可能因專業知識不足，而導致知識搜尋者無法正確取得資訊；另外，文件摘要之呈現沒有統一格式或是字數之控管，導致搜尋者容易因個人喜好而影響文件篩選，進而降低知識網站分享知識功能。綜合上述，其既有之運作模式如圖 1 之 AS-IS Model 所示。

如圖 1 所示，目前知識搜尋者取得資訊之方式可透過關鍵字搜尋、主題分類與文件摘要閱讀等，但以上述方式若針對特定知識領域仍有部分瑕疵，本研究乃將問題彙整並列點如下：

1. 知識搜尋者無從定義專有名詞以進行關鍵字搜尋：於勞工安全領域中有許多該領域之專有名詞，並且多數文件之題目、關鍵字皆以專有名詞撰寫與定義，於此無相關背景或是間接人員之搜尋者較無明確關鍵字定義並搜尋，以至於無法立即、確切取得所需資訊。
2. 主題分類眾多、自由形式摘要呈現方式不一：雖然勞工安全衛生知識網已針對主題進行文件分類，但主題下所包含議題眾多且複雜，此外文件摘要呈現方式乃由作者自行撰寫，並以自由形式方式呈現，以此方式容易因文字長短、文章編排等因素影響搜尋者閱讀習慣，進而影響知識文件篩選條件。

以工廠之老闆舉例說明，假設工廠老闆欲改善員工工作所造成身體痠痛等職業傷害，透過關鍵字「痠痛」於勞工安全衛生知識網搜尋其結果卻包含許多無正相關之文件，其原因乃該領域所定義之專有名詞為「肌肉骨骼傷害」而非一般搜尋者所解讀之「痠痛」，於此造成關鍵字無法確實發揮其功能。假設工廠老闆已縮小搜尋範圍，但文件摘要尚未結構化，容易因個人習慣選擇摘要字數較短之文件作為取決考量，而錯過重要知識文件。

有鑑於此，為掌握搜尋者需求並針對問題提供確實回饋，本研究認為可在關鍵字搜尋技術與文件摘要呈現方式上加強搜尋者篩選之思維模式，以協助知識搜尋者能快速且確實求得需求文件。因此，如期望之運作模式如圖 2 所示，本研究之研究動機與目的可歸納為以下兩點：

圖 1、知識搜尋者取得資訊之既有模式 As-Is Model

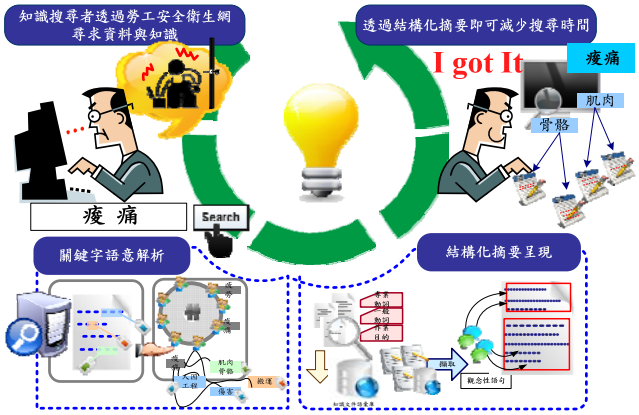

圖 2、知識搜尋者取得資訊之期望模式 To-Be Model

1. 建立知識語彙庫並加強關鍵字語意搜尋

由於勞工安全衛生知識網多為專業知識背景人士所撰寫，因而使用許多專有名詞做為文件之標題與摘要，但此方式造成無相關知識背景之一般搜尋者無法定義關鍵字以進行搜尋。因此，為加強關鍵字語意搜尋技術，本研究乃將文件進行解析並取得文件關鍵字詞，並與使用者搜尋字串進行語意關聯，以提高關鍵字搜尋之效能。

2. 結構化摘要以輔佐知識搜尋者檢視文件

當搜尋者縮小搜尋範圍欲以摘要閱讀方式篩選文件，但其摘要呈現乃由作者自行撰寫，以此方式易根據作者風格、喜好而有不同摘要呈現方式，以至於知識搜尋者容易因個人閱讀偏好，進而影響篩選文件。為改善搜尋者篩選文件之判斷，本研究乃將知識文件進行解析以取得文件結構、表達語彙等，並以文件重要性語彙為基，擷取文件觀念性語彙同時結合文件起承轉合呈現方式以形成結構化摘要。

整體而言，為協助知識搜尋者快速且有效取得資訊與知識，本研究乃解析知識文件之表達方式、呈現內容以及相關性語彙，並以此為基建立知識語彙庫，進而與搜尋字串進行語意關聯，以加強關鍵字搜尋技術。另外，亦可透過知識語彙庫取得文件之觀念性語句，並依照文件結構建立結構化摘要，以避免個人閱讀偏好影響文件之篩選，本研究之研究目的即縮短知識搜尋者搜尋文件之時間，進而加強勞工安全衛生知識網搜尋技術以提高勞工安全衛生知識網知識分享之成效。

2. 文獻回顧

本研究所涉及之研究主題乃包括「問答應用與技術分析」及「摘要應用與技術分析」等兩大研究方向，以下即針對此兩項主題之相關研究進行文獻回顧及探討。

2.1 問答應用與技術分析

對於問答應用與技術議題而言，本研究乃問答應用類型以及問答技術進行相關文獻探討，期望於其中觀察應用於不同類型之不同角度與層面解析，以更深層瞭解問答應用與技術特性。

(I)問答應用類型分析

於問答應用領域探討中，本研究乃針對「問答系統」以及「資訊檢索」等兩主題進行相關文獻探討，期望從中探討問答應用所涉及之範圍與領域。以下即針對「問答系統」以及「資訊檢索」進行文獻探討。

(A)問答系統

針對問答應用類型大多以搜尋者提出詢問句方式提出問題，並取得回覆句作為答案以回應，常以一問一答之方式進而取得答案，以此方式即可應用於醫學領域上協助搜尋者擷取複雜、無格式化之臨床報告重點，如 Cao 等人（2011）乃針對醫學上臨床報告建立一套線上問答系統（AskHERMES），以並將擷取重點總結後形成答案，而 Terol 等人（2007）則提出一套醫療互聯網問答系統，該研究透過自然語言處理以及廣泛解析器（Broad Coverage Parser，MINIPAR）解析問句中各詞性與依賴關係（Dependency Relationship），進而結合統一醫學語言系統（Unified Medical Language System，UMLS）與醫療術語網（WordNet）選擇相關資訊並擷取作為回覆句，以確保問答系統之資訊正確性，且能同時處理專有名詞與自然語言之問答，並以互聯網方式更能廣泛使用。Guo 及 Zhang（2009）藉由語意關係探索提出資料庫於領域知識上表示本體語意之方法，以醫療領域為主題，並透過自然語言分析（Natural Language；NL），將醫療領域之語言及完整口語問答方式進行關鍵連結。另外，為了取得最確實答案，Dalmas 和 Webber（2007）建立一套問題解答模式（QA Answer Model，QAAM），即避免以單一答案作為回覆句易導致其他資訊價值流失，進而問題分類並透過模式圖型程序（Model-View-Controller，MVC）進行準則分析，以確立回覆句與回覆句之間之關聯性，除了篩選回覆句外，亦分析回覆句之背景並作為評估準則，以確定回覆句能於不同背景下使用，並表示以組合回覆句方式更具有資訊參考能力。Scheffer（2004）先制定回答選項類別，並將客戶提問之電子郵件以支援向量機（Support Vector Machine；SVM）擷取郵件內容特徵，再以天真貝氏分類法（Naive Bayes Classifier）分析郵件特徵，以區分客戶於郵件中所提問題之所屬類別，再根據郵件所屬回答類別回應客戶對應回答內容，以達到客戶提問自動化回覆，Lorch 等人（2001）認為傳統檢索系統乃著重於快速回應使用者需求，而非針對問題給予確實回應。因此，提出一套醫學問答系統（Medical Definitional Question Answering System，MedQA)，該研究以監督學習方式將問題自動分類（如：分成 What、How 等，What 表示問定義而 How 則是問作法），使用者即可確實從問題中取得答覆，改善傳統檢索系統

以關鍵字方式快速回應。

(B)資訊檢索

資訊檢索多以定義關鍵字詞進行搜尋，待搜尋完後即可取得相關文件或資訊之回饋。對此，大多研究乃針對生物醫學領域進行資訊檢索，Huang 等人（2006）乃針對生物醫學文獻之關聯性分析，並提出一套混合式關係模式以擷取生物醫學文獻，以加強文獻之間關聯性同時提高資訊檢索之正確性。Hahu 等人（2002）基於醫療調查結果提出一套醫療自然語言系統（MedSynDiKATe），使檢索功能更加口語化。透過詞彙連結關係先將文件分析，並根據概念、語意關係將知識概念轉換成邏輯結構以整合成醫療知識庫。Uzuner 等人（2010）則針對臨床記錄提出以問題為導向之醫療記錄模式，以建立病歷之關聯性並應用於醫療記錄索引，進而可提供醫護人員較為正確之醫療訊息。該研究藉由醫學語言統一系統（Unified Medical Language System，UMLS）定義疾病類型與症狀，再以藉由支援向量機（Support Vector Machines，SVM）將病歷之重要語句、特點進行排序與分類，並透過特徵向量取得詞彙關聯性。Rindflesch 和 Fiszman（2003）以 Hypernymic 概念為基提出一套語言結構以解釋並應用生物醫學領域檢索系統，透過統一醫學語言系統（Unified Medical Language System，UMLS）與自然語言處理以建立語意網路，再加入 Hypernymic 生物分類學概念使語意更具體，以此方式能提高語意精確度亦能應用於資訊檢索技術以減少系統錯誤。此外，針對資訊量過於龐大，故無法針對使用者確切篩選文件之問題，Mingxin（2011）以語意網為基建立一套資訊檢索系統（Info rmation retrieval，IR），該研究以本體論先針對網頁文件之 HTML 語法以資訊描述架構（RDF）提取重要訊息，並將具有相同概念之資訊相互關聯，從中亦針對過去搜尋紀錄以取得網頁與關鍵字之關聯性並建立資料庫，其代理人即可根據資料庫與搜尋字串進行關聯並提供正確回應於使用者。Teng 等人（2010）則透過群集文件從中尋找主題方式，使文件更具關聯性。該研究乃蒐集被多人點選閱讀之文件，同時研究文件之關聯性與共同特徵（即基本要素 Basic Element，BE：包含語句之開頭、修飾詞與連結關係），利用共同特徵作為文件篩選條件並進行群集分析。當中，Yin（2011）即以引用圖特性建立生物醫學領域文獻檢索模式，以取得文獻檢索之最佳路徑。該研究先以連鎖分析演算法（Degree Distribution）分析文獻上下文之內容被引用機率，並透過資訊關聯方式程度（Degree）作為程度範圍，最後以關鍵字權重（Okapi）作為檢索基準線，以判斷該圖之節點是否為範圍內，進而增進文獻檢索之效果。

綜上所述，問答應用領域可針對特定知識領域進行詢問方式取得回應，或以檢索方式取得資訊，皆有許多文獻進行研究各領域之問答機制效果。

(II)問答技術分析

於問答技術中，過去研究多數都針對文件先行解析並取得文件之特徵與主題等，再進行回覆之任務；因此本研究乃針對「以文件主題分類」、「以文件特徵分類」及「以文件語意分類」進行相關文獻探討。

(A)以文件主題分類

針對文件主題分析部分，Oh 等人（2012）乃提出一套問答系統學習機制，透過現有問答文件以分析結構如（詞性與專有名詞等），再以詞義消歧法進行語意分析，從中取得問句與回覆句之組合（回答格式、回答主題、目標與預期回答內容），經由學習機制即可取得問句之重要字詞並連結相關回覆文件以取得回覆句，以此方式能更精確解決問題，同時亦提高問題解決之效率。Yangarber 等人（2000）提出一套非監督式自動化關聯評估模式（The Discovery Procedure，ExDIsCO），首先將文件進行文法結構拆解（即主詞、動詞與受詞），再根據詞彙鏈、詞頻等方式找出一組候選模式（即種子模式 Seed Patterns），並透過種子模式與文件內結構單字比對以判斷文件所屬類別，最後即可根據類別進行文件關聯分析，以此方式無需透過讀者自行解讀或是標記即可自動關聯相關資料以加強篩選。Han 等人（2007）提出以先確定問題類型並建立各類型相關詞彙，於此即可先從問題中判決問題目標並分析問題檢索之類別，以進行擴大查詢並提取過去共同出現之資訊作為外部資源。該研究乃以功能附屬分析器（Conexor FDG Parser）提取動詞或名詞，並以擴展技術獲取相關資訊之語句，最後以各種準則（冗餘或具有術語）進行評分與排序與擷取，以此方式能消除過於冗餘以精簡語句，並遞補更具意義之回覆句作為參考。

(B)以文件特徵分類

藉由文件特徵分析，即可根據分析結果與條件進行關聯以及相似性連結。因此，Jones 和 Love（2007）提出於文件關聯中具有主題角色扮演模式，若關係越具相似性則表示兩者之間存在一個共同

角色，此外亦有其他扮演關聯中之某角色（即掠食者則會有狩獵關係作用出現），透過背景環境、關係為匹配準則以取得文件之共同關係。透過模式以角色作用推理出具有相似之文件，其方式能更深入與擴展關聯分析機制。當中針對詢問句可包含多種問題條件，如：世界上最長的河流是什麼？當中條件有區域性（世界上）、比較性（最長的）以及詢問物（河流），因而無法以單一角度進行回覆。於此，Oh 等人（2011）乃提出一套複合式問答系統（Compositional QA），針對問題格式（單一問項或是多重問項）、主題、問題限制（時間或地點等）作為判斷準則並以此為基以得知回覆句之類型與格式，進而以語意匹配模式（Lexico-Semantic Patterns，LSPs）針對問題條件分別找尋答案，最後再將答案組成並回覆。Ko 等人（2004）提出以重要性語句作為文件分類之根據，以加強文件分類技術。該技術之概念乃由文件摘要延伸，透過語句內容與標題符合度、語句辭頻等權重計算取得重要性得分，根據重要性得分數取得文件分類之依據。於過去研究常以關鍵字詞進行分類，而忽視作者資訊亦可成為分類準則。因此，Legara 等人（2011）提出一套根據作者署名自動化分類專欄文章，根據作者寫作風格生成四個準則：(1)句法、(2)結構、(3)詞彙、(4)內容特點，並依照上述準則進行頻率排序（Frequency Ranking Method，FRM)）以及影響大小排序（Effect Size Ranking Method，ESRM)，最後乃以排序結果作為分類標準，該研究將此作法應用於意見專欄中並發現以此方式結果乃優於頻率分類法。

(C)以文件語意解析

大多問答系統（Question Answering，QA）乃先定義特定名詞（Named Entities，NEs）以建立問答機制，但若以生活語言（Common Noun，CN）為基之詢問句則無法確實取得回覆。是故，Moreda 等人（2011）以語意規則與語彙網（WordNet）為基建立一套語意問答系統。該研究首先進行詢問句演算分析以了解詢問句之語意角色，進而加入語彙網（WordNet）語意模式以取得過去類似問題之回覆句，再以主要關鍵標記（Prop Bank，PB）方式進行專有名詞與生活語言關聯。Dunlavy（2007）等人提出一套整合資訊查詢系統（The Query, Cluster, Summarize，QCS），該系統先根據文句位置、文件內容等特性標記文件之重要語句，透過潛在語意索引技術（Latent semantic indexing，LSI）進行相關性查詢分析，再以修剪法（Trimming）去除具有相關語句群集之冗長語句以形成摘要，最後以相關性將遞減列出提高查詢效益。Erdogan 等人（2005）將語法結構結合語意並提出三種方法以建立語言模式加強語音辨識能力。透過潛在語意分析、雙連語意詞彙模型（Two-Level Semantic-Lexical Modeling）解析資料之相關性並以機率模型計算詞彙序列，最後以最大熵法（Maximum Entropy，ME）結合成語言模式，經由此模式能增加語意之辨識能力亦能提高詞彙整合之緊密度。Ruiz-Casado 等人（2007）提出一套以詞彙模式自動識別語意關係，首先以最小距離法計算兩者間之相似性並以矩陣方式呈現，並從中去除無相關性，最後整合成一套特定關係，以擷取語意間特殊關係（例如：木星之於太陽系關係：木星是太陽系的一部分）亦能基於文件、特定領域形成整體-部分之關係（Holonymy）。透過統計分析可自動預測兩事件之關聯性，但由於語意關係易造成語句具有多種解釋，因此 Dorr 和 Gaasterland（2007）提出一套考量時態與語意關係之結合模式，基於時序關係、事件觀點等將相關事件結合。首先乃根據動詞時態（Basic Tense Structure，BTS，包含現在、過去、未來）轉成矩陣並以相關性形成時態結構（Complex Tense Structure，CTS），從中加入語法約束時態結構形成延伸時態結構（Constraint on Derived Tense Structures，CDTS），最後以組合排序法組織主題之關聯性即可擷取相關性語句，由於該模式乃將因果關係以及時序連結作為考量準則，透過重新分析與解釋可提高多筆事件關聯之整合性。　　綜合上述，針對問答技術層面皆有許多文獻研究，無論根據文件之主題、文件之特徵或根據文件語意方式，皆使資訊解析方式更加多元、其結果更加精準。

2.2 摘要應用與技術分析

對於摘要應用與技術議題而言，本研究針對摘要應用方式與摘要建立技術進行相關文獻探討，當中摘要應用可分為領域分析與文件類型分析與摘要技術分析三部分，期望於其中觀察摘要文件之解析方式，以更深一層瞭解文件潛在之資料特性，從中建立摘要。

(I)摘要領域分析

於摘要應用領域探討中，本研究乃針對「特定專業知識領域」以及「線上即時網際網路領域」等兩主題進行相關文獻探討，期望從中探討問答應用所涉及之範圍與領域。以下即針對「特定專業知識領域」以及「線上即時網際網路領域」進行文獻探討。

(A)特定專業知識領域

對於文件摘要大多乃應用於資料過量大或是具有過多相似主題文件之領域，如醫療領域、線上新聞與法律領域等。針對醫學領域文刊數量過多問題，Elhadad 等人（2005）乃建立一套統一摘要模式，基於搜尋檢索技術將其結果進行摘要總結以協助使用者更有效瀏覽。該系統乃以主題樹概念將主題作為連結節點，先將文件進行主題分群再根據檢索主題找出相關文件，再將相關文件進行多文件摘要以形成搜尋總結。於此，該模式能過濾大量醫學資訊，並提供使用者特定查詢之具體總結。Ling 等人（2007）提出一套以基因學角度為基之半結構化摘要技術，醫學學者可藉由此技術於醫學資料庫中將文獻進行半結構化摘要，更為方便與查詢於該領域最新資訊。該技術經由字典檢索、過濾同義詞方式訓練模型取出文件關鍵字並作為目標基因，透過目標基因作為語句擷取條件，再以空間向量模型（Vector Space Model，VSM）、機率語言模型（Probabilistic Language Model）訓練模型於語意部分之評分，最後乃擷取語句並形成摘要。Zhou 等人（2006）基於自然語言處理提出一套自動化擷取醫療術語模式，藉由專門於醫療領域之自然處理語言（Medical Language Extraction and Encoding，MedLEE）建立語意結構，經由圖形介面工具建立視覺化摘要以產生概念與主題，透過主題樹即可判定該術語之語意類型並擷取。該模式可改善人工建立醫療術語方式，透過自然語言處理將文件摘要視覺化呈現亦可減輕開發人員之負擔。此外，針對法律領域文刊數量過多問題，為改善手寫法案之效率，Moens（2007）將法院判決紀錄等形式文書予以電子化，並提出一套自動化個案摘要與檢索功能，使用者可根據相關案例找尋論點與諮詢。由於法院判決紀錄屬高度結構化文件，並具有固定出現概念序列，該研究將法院判決文件劃分為：(1)基本要素（包含被告、原告、案由、投訴類型等）、(2)背景、(3)相關法律問題、(4)辯護論點、(5)裁決與處置，根據上述畫分作為結構摘要關鍵字擷取依據。藉由分配概念將文件之內容分配至所屬概念，並基於圖形概念方法將鏈結相關性形成摘要，此作法不僅節省人力與時間，更能替律師將相關案件進行總結以方便尋找論據。Moens 等人（2005）提出一套透過通用於單一文件與多文件摘要模式，並以層級化方式呈現主題、摘要及細節部分。該模型先以詞性標記（Part-Of-Speech，POS）標籤具有文法結構和語意關係之語句，再以演算法根據主題、詞彙於文件之頻率、結束符號等，將文件層級化分割建立主題樹，透過主題樹取得語句以及對應語句以形成摘要。此外，針對多文件該研究將群聚技術結合詞性標記技術，即可藉由主題樹方式可依照層級形成摘要。該模式可應用於不同類型文件（如英文雜誌、荷蘭百科全書等）；此外，該模式亦可將文件壓縮以取得重要資訊並應用於新聞頭條上。Xie 和 Liu（2010）則針對特定會議記錄提出以迴歸監督式學習方式統計分類擷取具代表改善會議記錄總結，透過支援向量機（Support Vector Machines，SVM）監督學習方式標記並分類具有代表性文句，並以增加取樣（Up-sampling）、減少取樣（Down-sampling）和重複取樣（Re-sampling）三種不同取樣方式，以提高於少數類別（Minority Class）中擷取文句之預測準確性。

(B)線上即時網際網路領域

針對大量網頁文件搜尋不易之問題，Bouras 等人（2008）提出一套個人化新聞索引系統（PeRSSonal）。該研究將網路新聞、文章自動化摘要並透過分類以產生文字標籤（Text Labeling），結合網路新聞 RSS 訂閱服務即可進行個人化服務。藉由字頻計算取得新聞文件之關鍵字，並以餘弦相似性權重計算與分類新聞，最後選擇具有代表性語句形成摘要。針對網路新聞於手持行動裝置呈現與負載等問題，Yang 和 Wang（2007）針對手持行動裝置之文件進行自動化摘要。該研究利用碎形理論（Fractal Theory）形成摘要模型，並以樹狀結構或是分層方式加強手持行動裝置視覺化操作。該研究乃以碎形理論可將文件結構進行層級化（章節、段落、文句、字詞等），藉由分層即可取得章節區塊並計算文句於父節點之價值，找尋該章節之代表，再根據壓縮比決定文件摘要擷取文句之數量。此外，Zajic 等人（2007）為克服新聞標題文字長度之限制，運用文件壓縮技術將擷取單一文件之摘要語句，以形成候選語句並彙整成多文件摘要。其壓縮方式可透過兩種方法進行：(1)解析與整理（Parse-And-Trim）、(2)基於隱藏式馬可夫鏈模型（Hidden Markov Model，HMM）之改良模型：HMM Hedge（Hidden Markov Model HEaDline GEnerator）。透過解析與整理可去除冗長語句（即動詞、連接詞、備註等）同時解析語句長度是否符合限制規定，直到符合規定則擷取。其考量因素包含：字詞出現位置、字詞出現位置與開頭之距離、語意偏差，透過模型訓練進行語句壓縮與過濾最後擷取最符合之標題。Lin 和 Liang（2008）表示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，待時間過去讀者難以追朔該新聞事件。因而提出將事件主

軸納入摘要建立條件機制（Story-line based Topic Retrospection，SToRe），透過事件主軸能使讀者更了解事件發展與概念。該研究乃先界定事件再由事件建立議題，透過議題發展出脈絡並去除較無相關性事件，從中再擷取代表性語句並以議題主軸為依據形成摘要。該研究所提出之機制能提供更了解議題發展，同時亦可作為索引提高讀者閱讀效率。

另外，針對資料量過大之議題或是具有相似主題文件，Ye 等人（2007）提出一套以文件概念格（Document Concept Lattice，DCL）方式建立多文件摘要。該研究乃將相關主題、議題之對應語句於概念格中，從中挑取最具概念之語句形成摘要。該研究運用詞彙鏈以及反詞頻方式刪除冗長、不必要之詞彙，並將具有意義之詞彙形成一個有方向但非循環之概念圖。此外，Kuperberg 等人（2006）表示許多文章乃增加描述語句（即無相關語句）或是銜接語句（即中度相關語句）以強調閱讀重點（即高度相關語句），但過多描述反延長閱讀時間。因此，為了加強語句順暢性，利用功能性磁振造影（Functional Magnetic Resonance Imaging，fMRI）研究於事件之因果關係提出以簡短訊息（即兩句背景以及一句關鍵語句）確立與了解主要訊息。Sweeney 等人（2008）表示由於自動化文件摘要多為壓縮文件內容而擷取重要片段，但此方式易影響原始文件之正確解釋。因此，提出以新語句方式建立摘要，並透過延伸閱讀方式（Show Me More）給予更多資訊以完整表達文件。該研究先以搜尋者查詢為需求條件建立摘要（Query-biased Summarisation Methods）並基於文件結構（標題、開頭）、詞頻和查詢條件等條件進行得分計算，再依據得分排序形成摘要順序；此外基於類似文件即去除重複性語句並形成摘要。Ko 和 Seo（2008）考量非所有文件皆以特定格式（如標題與位置）撰寫於此影響文件摘要之形成，因而針對無標題樣式文件進行摘要，該研究提出上下文資訊以及混合統計方式，透過雙連搜尋（Bi-gram）技術將文件之兩個連續語句結合成一個虛擬語句，再以詞頻計算方式（TF-based Query Method）篩選出關鍵字詞，將關鍵字詞作為語句相似度衡量依據，以取得出最重要、最具代表性之語句以形成摘要。

(II)摘要技術分析

於摘要建立技術部分，過去研究將摘要建立技術分為「監督式機器學習」方式與「非監督式機器學習」方式等兩種技術進行探討。

(A)監督式學習

針對監督式機器學習可細分為各種技術，如詞彙鏈、支援向量機或主題樹等技術。透過詞彙進行標記以形成詞彙鏈，但語意排序正確性乃影響摘要之可讀性，若語意排序錯誤可能造成閱讀混淆或是形成錯誤觀念。因此，Bollegala 等人（2010）提出一套由下而上方式排序語句以結合兩份文件摘要技術，該研究以年代、主題適配度、繼承、優先等級作為擷取標準，並以監督式訓練模型計算語句間之方向性與強度排序，同時結合支援向量機（Support Vector Machine，SVM），以進行分類即可群集兩份文件之語句與排序並擷取形成摘要。Jung 等人（2005）透過假設語句結合統計計算方式以提出一套自動化摘要模式，藉由雙向連詞（Bi-gram）結合兩兩相鄰語句以建立假設語句，並以假設語句為基計算該語句於文件之重要性（即標題、位置等準則）以刪除不必要語句，最後再將假設語句分離以進行群聚分析取得相似性，取得更具代表性語句已形成摘要。Benedí 和 Sánchez（2005）將 SCFGs 技術結合 N 連詞提出一套結合語言模型，透過 N 連詞擷取詞彙關聯性，再以隨機語法技術進行關聯性分類，以此方式即可降低詞彙關聯複雜性進而提高 SCFGs 之運算。Li 和 Chen（2010）利用統計語言模型先進行判斷文件相關性，再以機率排序和潛在馬可夫鏈模式擷取出具代表性之片段。藉由此方式能精確取得文字片段開頭與結尾以提高語意之協調性，亦可針對使用者需求應用於文件自動化摘要。Oliva 等人（2011）針對簡短文件提出一套基於結構（即名詞、形容詞等）之語意相似度系統（Syntax-based Measure For Semantic Similarity，SyMSS），透過 WordNet 概念模擬人類共同詞彙以強化語意合理性與相似性，同時利用註解測量（Gloss-based Measures）方式比較字詞語意、相似性與詞性，並以加總方式計算語句之間相似性。Ouyang 等人（2011）建立一套以特徵為基擷取模型，其方式乃預先定義七個篩選條件（根據語意、與主題之相關性、字詞出現頻率、是否具有結尾語意以及語句存在位置）並作為評分準則，透過支援向量迴歸模型（SVR）學習方式將具有相同得分條件之語句關聯群集以完成分類，從中進行二元決策方法篩選語句，根據篩選即可留下最具價值、重要性語句之最佳組合。針對多文件摘要技術大多針對需求者之要求、主題建立，進而要求自動化摘要能於有限字數內表達涵義。是故，Vanderwende 等人（2007）以多文件摘要系統 SumBasic 為基提出一套多文件摘要總結系統，透過擷取、語句簡化、和詞彙擴充方式並以任務為目的進行多文件主題摘要。基於 SumBasic 系統主要

透過詞頻並考量人類手動摘要方式提出多文件摘要，於此該研究加入語句壓縮以及加強同義詞等詞彙關係以擷取重要內容，並以簡化方式滿足需求者，經由此系統能將主題詞彙擴充以提高相關性文件之擷取。

(B)非監督式學習

針對非監督式機器學習技術，可分為向量空間模型與潛在語意分析兩種。Chan（2006）基於潛在語意分析提出一套擷取最具代表性語句之量化模式，透過潛在語意加強人類理解模式並將具有代表性或連結性之語彙組織成詞彙網路圖以表示文件語句間之關聯性，並以此為基建立模式，透過此模式能加強文件摘要生成之連續性，並以關聯方式加強語句相關性。He 等人（2009）提出一套自動化摘要評估系統，為減少篩選時間藉由潛在語意分析（Latent Semantic Analysis，LSA）以矩陣方式取得文字出現頻率並以餘弦定理確定相似性，同時結合 N 連詞共生（N-gram Co-occurrence）以評估、篩選作品，再以機器翻譯排名系統（BLEU Algorithm）機器學習方式標記作品並藉由匹配訓練以提高精確度，透過 LSA 以及 N 連詞共生方式乃優於自動彙整評估系統（Recall-Oriented Understudy for Gisting Evaluation，ROUGE）評估更能加強輔佐以提高篩選分級效率。Steinberger 等人（2007）以潛在語意分析（Latent Semantic Analysis，LSA）為基建立自動化摘要系統，並文件壓縮百分比方式提出摘要字數之限制。該研究以 TERM-BASED 方法找出具重要性詞彙並延伸文件之詞彙關係，同時結合潛在語意分析（Latent Semantic Analysis，LSA）以自動擷取出相關語句生成摘要，以此方式能提高多文件摘要之彙整性亦提高摘要之一致性。針對未涉獵之領域文件資料，文件摘要多數需以人工手動標記方式形成有系統性之規則，且非常費力與耗時。因此，Nomoto 和 Matsumoto（2001）提出一套非監督式多樣性多文件摘要技術，首先利用群聚演算（K-means）並以最小距離（Minimum Description Length Principle，MDLP）於各領域中尋找相關主題文件，再以 Zechner 所提出之加權模型（Z-model）去除多餘語句同時亦擷取最具代表性語句以自動形成多樣性多文件摘要，以此方式不僅能取代人力亦能應用於檢索技術，透過檢索結果形成多文件摘要以加強篩選。Yeh 等人（2005）分別為以「改良語彙庫為基技術」（Modified Corpus-based Approach，MCBA）與「潛在語意分析 LSA 為基結合 TRM 技術」，以建立文件摘要。當中，MCBA 方式乃以特徵評分（即位置、正向關鍵字、排除負向關鍵字、向心性以及與標題相似性等），計算結合基因演算法（Genetic Algorithm，GA），以擷取得最佳組合形成摘要；第二種技術則透過潛在語意分析（LSA）則取得文件或語彙庫之語意矩陣，並建立語意關聯圖（Text Relationship Map，TRM），以取得顯著語意並形成摘要。Ko 等人（2003）利用詞彙群聚提出一套以主題為基之文件摘要技術，首先將文件進行上下關聯分析並以空間向量呈現，從中取得詞彙群集並確定群集核心以產生主題和關鍵字，最後則以壓縮比率作為之擷取語句之固定數量形成摘要。此外，Yangarber（2003）以特定名詞（人、位置、組織及日期等）進行分類並解析文件結構（主詞、動詞與受詞）提出一套非監督式學習語意模式。首先以權重配置方式取代二元決策進行相關性匹配，即可以權重準確取決具有相關性候選文件，透過此方式即可自然延伸出停止學習之終點，以提高非監督式學習之精確性。

2.3 小結

綜上所述，本研究之主要目的乃以知識文件為基礎，以進行專業語彙語一般詞彙之語意關聯性以及知識文件結構化摘要推論，本研究藉由此二大方向進行文獻探討，因此，由2.1節之文獻回顧可知，過去「問答應用與技術」之相關研究可發現針對問答系統之可以文件主題、文件特徵或是文件語意作為主要分析因素並等部分進行探勘，以透過問項類型，以進行問題之解答。由2.2節之文獻回顧可知，過去「摘要應用與技術」之相關研究，可得知過去針對摘要之技術大多乃應用於資料量過大或是具有相似主題之文件，此外根據摘要建立之技術可以機器學習方式或是根據文章內容與結構評估等方式。

整體而言，有別於先前之研究，本研究藉由2.1節之文獻回顧所探討「問答應用與技術」議題，可得知文件主題、特徵以及語意作為分析因素，並以此為依據建構「知識文件表達項目解析模組」進行表達項目分析與觀念性語句擷取，以及建構「知識文件問答解析模組」進行主要詢問詞分析，並透過詢問詞與回覆詞配對解析找出具有關聯性之語意詞，進而取出對應知識文件，接著藉由2.2節之文獻回顧所探討「摘要應用與技術」議題，所得知摘要可應用於知識領域以輔佐知識文件之呈現，此外針對摘要呈現方式不同進而影響搜尋者閱讀觀感。因此本研究乃建構「知識文件結構化摘要推論模組」進行知識文件摘要建立，對於知識文件呈現方式，本研究乃以簡略結構化摘要描述文件之主要動機與目

的，另以詳述結構化摘要描述文件之細部內容。最後知識搜尋者則有效縮短知識搜尋者搜尋文件之時間，以及提高知識分享之成效。

3. 整合知識問答解析與文件結構化摘要技術之知識文件檢索模式

本研究所提出之「整合知識問答解析與文件結構化摘要技術之知識文件檢索模式」乃以勞工安全知識網所提供之技術叢書與研究報告為分析基礎，先行解析知識文件之結構特性以提出八個表達項目、二十八個表達細項同時建立「知識語彙庫」，透過知識語彙庫即可取得專有名詞並進行語意關聯，針對搜尋字串進行主要詢問詞解析並找出具有關聯或相關語意詞，進而取出對應知識文件以提高檢索準確性；此外亦可以知識語彙庫為基建立語彙法則，以取得觀念性語句並根據結構化摘要之法則擷取具有該文件之代表性語句以建立摘要，即形成一份制式知識文件以釐清與完整表達報告內容。因此本研究之主要流程可分為四大部份，分別為如圖 3 之 Part1「知識文件表達項目解析模組」、Part2「問答解析模組」、Part3「結構化摘要模組」所示。

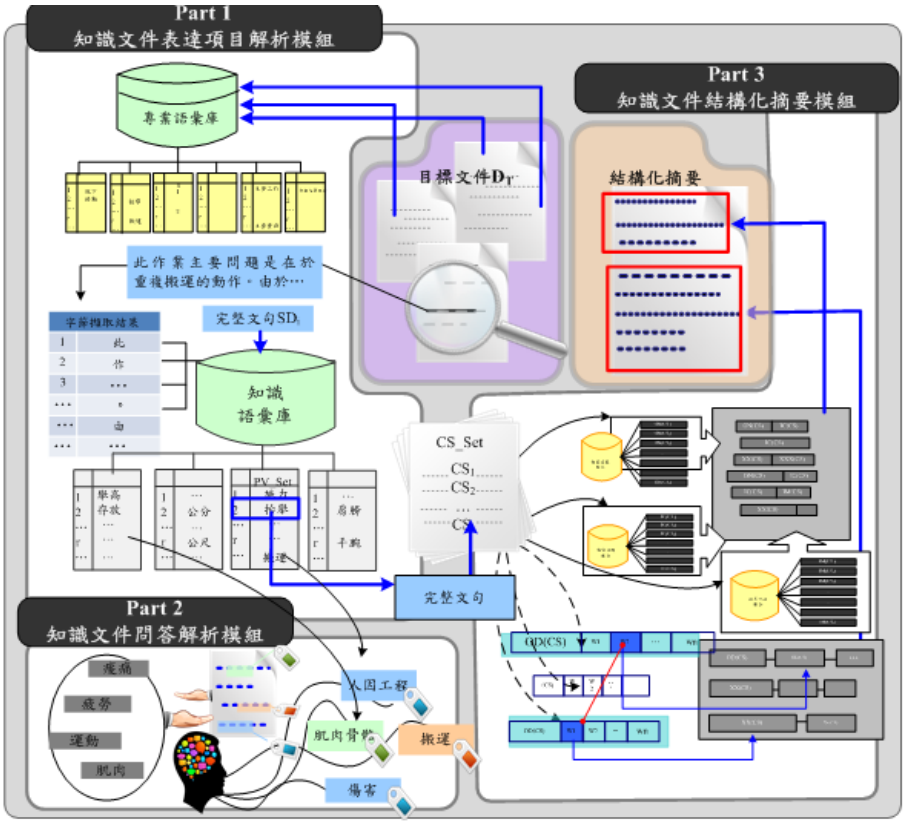

圖3、整合知識問答解析與文件結構化摘要技術之知識文件檢索模式之流程架構

3.1知識文件表達項目解析模組

由於人因工程工作場所研究報告皆由專業人員進行訪視後撰寫，多以專業名詞、專業知識描述報告，為協助無背景知識之使用者進行搜尋。因此，針對表達項目建立乃參照人因工程改善報告與技術叢書內文，如於技術叢書「人因工程危害預防技術」提到造成肌肉骨骼傷害之五個主要成因為過度施力、高重複動作與震動等因素，本研究與人因工程相關人員進行討論，並於討論後將改善報告與技術叢書內文中重複出現或重要之描述詞彙彙整，再進行分項以建立知識文件之表達項目。待本研究解析知識文件並建立表達項目與觀念性語句擷取，針對各不同表達項目與表達細項之細部說明如下：

3.1.1建立知識文件之表達項目

針對人因工程工作場所研究報告文件之內容解析，可分為八個表達項目、十九個表達細項，為加強語句順暢性，本研究另外增加九個表達細項進行語句輔助。故「知識語彙庫」共計包含二十八個語彙集合，其二十八個語彙集合如表1與表2所示：

表 1、表達細項與對應內容及表達方式

表達項目	表達細項	表達方式
作業領域	作業領域	表達行業類別之集合；其內容包含「農林漁牧業」、「礦業及土石採取業」、「食品製造業」、「紡織成衣業」等。
作業名稱	作業名稱	表達作業在此行為之名稱集合；其內容包含「更換模具做業」、「納箱作業」、「搬運作業」。
作業身分	作業員性別	形容該作業之作業員性別語彙之集合；其內容包含「男性」「女性」或是「男女不拘」。
	作業員年齡	形容該作業之作業員年齡範圍之語彙，如「20~30 歲」；其內容包含「中年」、「青年」、「無年齡限制」等。
	作業員職稱	為形容該作業之作業員職位之語彙，其內容包含「護士」、「技術師」、「作業員」等。
作業環境	設備語彙	為描述該作業所使用之設備語彙之集合，如「採血床」、「保溫櫃」、「熱鍛機台」等。
	設備佈置	即設備擺放與佈置語彙之集合，如「輸送帶的高度為 75 公分」。
	工具介紹	為描述作業或是改善流程中所使用之工具語彙集合；其內容包含「蝴蝶籠」、「手臂支撐靠墊」、「升降台車」、「手推車」等。
作業行為	作業目的	為作業目標對應作業名稱語彙之集合；其內容包含「主要工作」、「主要功能」、「主要重點」等。
	作業描述	為對應作業名稱語彙、作業目的語彙、作業工具語彙之集合。
	專業動詞	為描述作業員對應作業描述語彙之姿勢集合；其內容包含「站姿」、「彎腰」、「低頭」、「施力」、「抬舉」。
作業時間	作業次數/天	為描述作業員在此作業所重複動作之次數；其內容包含「一天需要」、「一天必須」等。
	作業時間/次	為描述作業員在此作業所花費時間之集合；其內容包含「一次所花時間」、「一次需要」、「一次必須」等。
	作業距離/次	為描述作業員在此作業所需距離之集合；其內容包含「距離」、「最近距離」、等。
傷害成因	傷害因素	為描述傷害成因之集合；其內容包含「過度施力」、「高重複性動作」、「震動」、「低溫」、「不良工作姿勢」。
	傷痛部位	表示作業員對應作業描述語彙之姿勢部位及傷痛部位之集合；其內容包含「頭頸」、「軀幹」、「手部」、「手腕」、「腿部」。
改善方法	改善目的	表明對應傷害成因語彙之改善目的之集合；其內容包含「主要改善」、「有效改善」、「大程度地」、「明顯降低」等。
	改善流程	描述流程改善語彙之集合；其內容包含「考慮」、「利用」、「建議」、「只要」等。
	改善評估	解釋其作業改善後評估語彙之集合；其內容包含「行動水準」、「評級點數」、「檢核總分」、「風險等級」、「負荷明顯降低」等。

表 2、輔助表達細項與對應內容及表達方式

表達細項	表達方式
連結語彙	其內容包含「與」、「和」、「但是」、「以」。
一般性動詞	一般性動詞之語彙集合；其內容包含「是」、「主要是」、「為」、「舉高」、「放下」、「移動」、「存放」。
數字語彙	記錄「0」至「9」及其所組合之語彙。
金錢單位語彙	包含「元」、「台幣」等。
年齡單位語彙	包含「歲」。
長度單位語彙	包含「距離」、「公分」、「公尺」、「長」等。
時間單位語彙	包含「時間」「分鐘」、「小時」等。
重量單位語彙	包含「公斤」、「公克」、「噸」等。
頻率單位語彙	包含「次」、「下」等。

3.1.2 觀念性語句擷取

先將目標文件D_T進行字節斷句方式取得完整語句SD_i，並以領域專家所建立之「知識語彙庫」為基增加語彙法則，再將其完整語句SD_i與語彙法則進行觀念性語彙比對，進而取得觀念性語句並配置於各自歸屬語彙集合中。如公式(1)所示，首先以字節擷取方式取得完整語句SD_i。

$$SD_i = \{SD_{i,1}, SD_{i,2}, SD_{i,3} \cdots, SD_{i,j}, \cdots\} \quad (1)$$

為更具完整性與精確性，本研究乃增加語彙法則以明確定義觀念性語句之結構。針對各語彙法則建立八種篩選規則，其詳細說明與定義語彙法則之表達方式如下：

1. 作業領域語彙法則（R_OF）：如公式(2)所示，若完整語句存在於作業領域觀念語彙中，即表示該完整語句 SD_i 為作業領域之觀念語句 OF_Set。

$$\text{IF } SD_{i,j} \text{ exist in OF(CS)} \ \forall j \ \text{ Then } \ SD_i \in OF\_Set \quad (2)$$

2. 作業名稱語彙法則（R_ON）：如公式(3)所示，若完整語句存在於作業名稱觀念語彙中，即表示該完整語句 SD_i 為作業名稱之觀念語句 ON_Set。

$$\text{IF } SD_{i,j} \text{ exist in ON(CS)} \ \forall j \ \text{ Then } \ SD_i \in ON\_Set \quad (3)$$

3. 作業身分語彙法則（R_OR）：在此階段將設置寬鬆法則以及嚴謹法則乃確保擷取無誤。寬鬆法則：能以一個至兩個語彙即能表示作業身分之觀念，如公式(4)至(6)所示。嚴謹法則：以多個語彙形成嚴謹結構以表達其作業身分之觀念，如公式(7)所示。

$$\text{IF } SD_{i,j} \text{ exist in ORT(CS)} \ \forall j \ \text{ Then } \ SD_i \in OR\_Set \quad (4)$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{ORT(CS) and ORA(CS)}) \forall j \ \text{ Then } \ SD_i \in OR\_Set \quad (5)$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{ORT(CS) and ORS(CS)}) \forall j \ \text{ Then } \ SD_i \in OR\_Set \quad (6)$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{ORT(CS) and N(CS) and AU(CS)}) \forall j \ \text{ Then } \ SD_i \in OR\_Set \quad (7)$$

4. 作業環境語彙法則（R_OE）：根據公式(8)至公式(9)所示，對於作業環境之描述須以設備語彙 F(CS)結合設備佈置語彙 FL(CS)，同時結合數字語彙 N(CS)以及長度單位語彙 LU(CS)、乃明確表達作業設備與工具之規格。

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{F(CS) and FL(CS)} \\ \text{and N(CS) and LU(CS)} \end{pmatrix} \forall j \ \text{ Then } \ SD_i \in OE\_Set \quad (8)$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{OT(CS) and N(CS) and LU(CS)}) \forall j \ \text{ Then } \ SD_i \in OE\_Set \quad (9)$$

5. 作業行為語彙法則（R_OV）：根據公式(10)至公式(11)所示，對於作業目的之表達描述須以作業目的語彙 OG(CS)結合一般動詞語彙 GV(CS)、作業工具語彙 OT(CS)。以及專業動詞語彙 PV(CS)。

$$\text{IF } SD_{i,j} \text{ exist in } (\text{OG(CS) and GV(CS)}) \forall j \ \text{ Then } \ SD_i \in OV\_Set \quad (10)$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{OG(CS) and GV(CS) and PV(CS)}) \forall j \ \text{ Then } \ SD_i \in OV\_Set \quad (11)$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{ON(CS) and GV(CS)} \\ \text{and PV(CS) and OT(CS)} \end{pmatrix} \forall j \ \text{ Then } SD_i \in OV\_Set \quad (12)$$

6. 作業時間語彙法則（R_OH）：作業時間語彙法則即表示作業之頻率與所花時間。其描述方式包含作業頻率語彙 OFQ(CS)、作業時間語彙 OH(CS)、作業距離語彙 ODT(CS)等。透過公式(13)、公式(14)、公式(15)篩選並將其符合之語句列為作業時間語彙之觀念語句 OH_Set。

$$\text{IF } SD_{i,j} \text{ exist in} (\text{OFQ(CS) and PV(CS) and N(CS)}) \forall j \ \text{ Then } SD_i \in OH\_Set \quad (13)$$

$$\text{IF } SD_{i,j} \text{ exist in} (\text{OH(CS) and N(CS) and FU(CS)}) \forall j \ \text{ Then } SD_i \in OH\_Set \quad (14)$$

$$\text{IF } SD_{i,j} \text{ exist in} \begin{pmatrix} \text{OT(CS) and ODT(CS)} \\ \text{and N(CS) and LU(CS)} \end{pmatrix} \forall j \ \text{ Then } SD_i \in OH\_Set \quad (15)$$

7. 傷害成因語彙法則（R_IC）：傷害成因語彙法則乃表達作業所造成之傷害，其表達方式包含傷害成因語彙以及部位語彙，根據公式(16)找尋具有傷害成因語彙之觀念語句 IC_Set。

$$\text{IF } SD_{i,j} \text{ exist in} (\text{IC(CS) and B(CS)}) \forall j \ \text{ Then } SD_i \in IC\_Set \quad (16)$$

8. 改善方法語彙法則（R_IM）：根據公式(17)至公式(18)所示，改善方法語彙法則表達方式包含改善目的 IG(CS)、改善流程 GV(CS)、改善評估等 IR(CS)，並結合一般動詞 GV(CS)、專業動詞語彙 PV(CS)以及作業工具語彙 OT(CS)表達之。

$$\text{IF } SD_{i,j} \text{ exist in } (\text{IG(CS) and GV(CS) and PV(CS)}) \forall j \quad \text{Then } SD_i \in IM\_Set \tag{17}$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{IR(CS) and GV(CS) and OT(CS)}) \forall j \quad \text{Then } SD_i \in IM\_Set \tag{18}$$

$$\text{IF } SD_{i,j} \text{ exist in R(CS) } \forall j \quad \text{Then } SD_i \in IM\_Set \tag{19}$$

$$\text{IF } SD_{i,j} \text{ exist in } (\text{RV(CS) and ORT(CS) and PV(CS)}) \forall j \quad \text{Then } SD_i \in IM\_Set \tag{20}$$

在此模組可得八種觀念語句之集合，分別為作業領域、作業名稱、作業身分、作業環境、作業行為、作業時間、傷害成因以及改善方法等。透過知識文件表達項目解析階段乃擷取知識文件表達項目之觀念性語句，即為後續問答解析與結構化摘要之應用。

3.2 問答解析模組

由於知識搜尋者所輸入之搜尋字串，大多屬知識搜尋者以個人直覺且不屬於知識文件專業語彙之詢問詞。針對傳統關鍵字檢索技術，若以專業語彙搜尋即可找到所屬相關知識文件；反之，自行定義之詢問詞可能因不具有明確定義，則有找出非所屬相關文件。因此，為加強自然語言搜尋彈性，本研究提出知識文件問答解析模組，針對知識搜尋者之搜尋字串進行主要詢問詞分析，並透過詢問詞與回覆詞配對解析找出具有關聯性之語意詞，進而取出對應知識文件以提高檢索準確性。

3.2.1詢問詞隱含目的判斷

本步驟以詢問（Question）與回覆答案（Answer）方式從中取得詢問詞所隱含目的與相關詞，以訓練詢問詞與相關語意詞之相似機率。如公式(21)所示，當知識搜尋者輸入詢問句（QW_i）經由斷詞取出詞性序列，即可從詢問句透過字詞拆解以擷取有效詢問詞（$QW_{i,j}$），根據勞工安全衛生研究所制定之雙語詞彙進而取得相關語意詞，即根據專有名詞「擦傷」可根據附註說明「覆蓋於身體的表面組織（如皮膚或黏膜）被擦掉或刮掉」，以取得相關語意詞「皮膚」或「黏膜」等，並藉由語意詞作為詢問詞以取得之回覆句（AW_k）亦由多個回覆詞（$AW_{k,m}$）所組成。藉由公式(22)計算詢問句（$QW_{i,j}$）與回覆詞（$AW_{k,m}$）之詞彙出現機率，以取得相關性詞彙並形成詢問詞集合並將各詢問詞相對之回覆詞彙。最後如公式(23)所示，藉由詢問詞相對於回覆詞語意門檻值 ω(AW,QW)篩選與該回覆詞句有相關性之詢問詞以成集合。

$$QW_i = \{QW_{i,1}, QW_{i,2}, QW_{i,3}, \cdots, QW_{i,j}, \cdots\} \quad and \quad AW_k = \{AW_{k,1}, AW_{k,2}, AW_{k,3}, \cdots, AW_{k,m}, \cdots\} \tag{21}$$

$$P(AW_{k,m} \mid QW_{i,j}) = \frac{C(AW_{k,1} \mid QW_{i,1}) + C(AW_{k,1} \mid QW_{i,2}) \cdots + C(AW_{k,1} \mid QW_{i,j})}{C(QW_{i,j})} \bullet C(AW_{k,1}) \tag{22}$$

$$\text{IF } P(AW_{k,m}, QW_{i,j}) > \varpi(AW, QW) \text{ Then } AW_{k,m}, QW_{i,j} \in QWAW^{\varpi}{}_k \tag{23}$$

3.2.2詞彙類別相似度判斷

本研究乃透過目標文件解析模組所提出之表達項目為基以建立文件關鍵字集（$D_{i,q}$），並如公式(24)將每個表達項目之表達語彙定義為專業語彙。本步驟以向量空間模型（Vector Space Model，VSM）之餘弦函數（Cosine）計算該文件與回覆詞集合之相似度，並以公式(25)判斷該回覆詞集合與文件之相似程度 $Sim(D_q|QWAW_k)$，若相似度大於門檻值 ω 且值越趨近於 1，即亦表示該群組越具有文件之解釋意義。

$$D_i = \{D_{i,1}, D_{i,2}, D_{i,3}, \cdots, D_{i,q}, \cdots\} \tag{24}$$

$$D_q^{\varpi} = [w_1, w_2, \cdots, w_q]^T, QWAW^{\varpi}{}_k = [w_1, w_2, \cdots, w_k]^T, \text{Sim}(D_q \mid QWAW_k) = \frac{D^{\varpi}{}_q \cdot QWAW_k^{\varpi}}{|D^{\varpi}{}_q| \cdot |QWAW_k^{\varpi}|} \tag{25}$$

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文件，如公式(26)至公式(29)，門檻值設定可為平均值、中位數、四分位數與直接定義，若大於門檻值 ω(Dq,QWAWk)即表示專業語彙與該文件具有連結性，並放置候選文件集合（CandidateDoc_Setd），其設定方式如表 3 所示。

表 3、門檻值設定方式(1)

門檻值	公式(26)至公式(29)	解釋
平均值	$\omega(CT,D_i)=\frac{\sum_{all\ n} Sum(D_i)}{N(CT_n)}$ ， IF $Sum(D_i)\geq\omega(CT,D_i)$ Then $D_i\in CondidateDoc\_Set_d$ (26)	加總專業語彙比對次數，並除以總數求得平均值
中位數	$\varpi(CT,D_i)=\begin{cases}Sum(D_{\frac{i+1}{2}}) & \text{IF } N(CT_n) \text{ is odd}\\ \frac{1}{2}\times\left(Sum(D_{\frac{i+1}{2}})+Sum(D_{\frac{i+1}{2}+1})\right) & \text{IF } N(CT_n) \text{ is even}\end{cases}$ (27) and IF $Sum(D_i)\geq\varpi(CT,D_i)$ Then $D_i\in CondidateDoc\_Set_d$	以整體筆數之中位數作為門檻值
四分位數	$Q=N(CT_n)\times 75\%$ $\varpi(CT,D_i)=\begin{cases}Sum(D_{i\times 75\%}) & \text{IF } Q\notin\{X:\lvert X\rvert\in N\}\\ \frac{1}{2}\times(Sum(D_{i\times 75\%})+Sum(D_{i\times 75\%+1})) & \text{IF } Q\in\{X:\lvert X\rvert\in N\}\end{cases}$ (28) *and* IF $Sum(D_i)\geq\varpi(CT,D_i)$ Then $D_i\in CondidateDoc\_Set_d$	以第三四分位數進行篩選。
直接定義	IF $Sum(D_i)\geq\varpi(CT,D_i)$ Then $D_i\in CondidateDoc\_Set_d$ (29)	使用者自行制訂

透過「知識文件問答解析」模組乃藉由詢問詞隱含目的判斷以取得該詢問詞之目的與相關回覆詞，再以回覆詞與專業語彙關聯分析，間接取得詢問詞與文件之相似關聯程度，進而取得具有相關性之相關文件。

3.3 結構化摘要模組

透過觀念性語句擷取階段本研究已得具有觀念性之語句，但考量摘要字數之限制以及摘要描述之完整性。在此階段仍需建立結構化摘要之法則，乃確保其觀念性語句具有該文件之代表性。考量知識搜尋者需求以及文章敘述完整性等因素，本研究乃將上述集合畫分為二部分，分別為簡略部分以及詳述部份。使用者可先行透過簡略結構化摘要快速瀏覽以得之文件之動機與目的等重點描述，若對該份文件有興趣即可再以詳述結構化摘要了解文件細部內容，即可更了解文件內容進而判斷文件之重要性。

3.3.1簡略部分建立

針對簡略摘要推論部分，主要以簡短描述方式表達文件之動機與目的。因此，於簡略摘要階段，乃擷取最具有觀念性與代表性之語句，因而計算語句之向心性程度，假設該語句同時存在多種觀念性語彙即表示該語句具有代表性，並將其語句列為候選語句以進行語句之結構強度計算，同時亦考量文章閱讀性且納入語句之結構強度作為準則判斷，即表示語句之主詞、受詞與動詞之間之關聯程度）。在此所擷取之語句集合包含：作業名稱集合、傷害成因集合以及改善方法之集合。

其條件代碼包含：具有作業名稱之觀念性語句且存在作業目的語彙之條件（W_1）、具有作業名稱之觀念性語句且存在作業工具語彙之條件代碼（W_2）、具有傷害成因之觀念性語句且存在專業動詞語彙之條件代碼（W_3）、具有傷害成因之觀念性語句且存在作業工具語彙之條件代碼（W_4）、具有改善方法之觀念性語句且存在改善目的語彙之條件代碼（W_5）、具有改善方法之觀念性語句且存在作業工具語彙之條件代碼（W_6）與具有改善方法之觀念性語句且存在改善動詞語彙之條件代碼（W_7）。

步驟(A1) 具有觀念性語句之向心性計算

根據簡略階段法則之定義擷取作業名稱、傷害成因與改善方法之觀念性語句乃為首要動作。在此步驟乃先計算作業名稱、傷害成因與改善方法之觀念性語句向心性，首先將其各表達項目之具有觀念性語句（即$ON(CS_i)$、$IC(CS_i)$與$IM(CS_i)$）拆解成二至六字詞以形成集合表示（即$ON(CS_{i,j})$、$IC(CS_{i,j})$與$IM(CS_{i,j})$）。各表達項目之向心性判斷方式如公式(30)至公式(32)所示，首先以字詞進行各項語彙比對，並以條件代碼為1若不存在則為0方式累計向心性分數，最後乃以Score ON(CS)、ScoreIC(CS)與ScoreTM(CS)呈現觀念語句之向心性分數，並將具有向心性語彙放置於各表達項目之候選語彙集合。

$$
\begin{aligned}
&ON(CS_i)=\{CS_{i,1},CS_{i,2},CS_{i,3},\cdots,CS_{i,j},\cdots\}\\
&IF\ ON(CS_{i,j})\ exist\ in\ OG(CS)\ \forall j\ \ Then\ W_1=1\ ,Otherwise\ W_1=0\\
&IF\ ON(CS_{i,j})\ exist\ in\ OT(CS)\ \forall j\ \ Then\ \ W_2=1\ ,Otherwise\ W_2=0\\
&Score\ ON(CS_i)=W_1+W_2\\
&IF\ \ 1<ScoreON(CS_i)\le 2\ Then\ ON(CS_i)\in CandidateON
\end{aligned}
\tag{30}
$$

$$
\begin{aligned}
&IC(CS_i)=\{CS_{i,1},CS_{i,2},CS_{i,3}\cdots,CS_{i,j},\cdots\}\\
&IF\ IC(CS_{i,j})\ exist\ in\ PV(CS)\ \forall j\ Then\ W_3=1\ ,Otherwise\ W_3=0\\
&IF\ IC(CS_{i,j})\ exist\ in\ B(CS)\ \forall j\ \ Then\ \ W_4=1\ ,Otherwise\ W_4=0\\
&Score\ IC(CS_i)=W_3+W_4\\
&IF\ 1<ScoreIC(CS_i)\le 2\ Then\ IC(CS_i)\in CandidateIC
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
&IM(CS_i)=\{CS_{i,1},CS_{i,2},CS_{i,3}\cdots,CS_{i,j},\cdots\}\\
&IF\ IM(CS_{i,j})\ exist\ in\ IG(CS)\ \forall j\ \ Then\ W_5=1\ ,Otherwise\ W_5=0\\
&IF\ IM(CS_{i,j})\ exist\ in\ OT(CS)\ \forall j\ Then\ \ W_6=1\ ,Otherwise\ W_6=0\\
&IF\ IM(CS_{i,j})\ exist\ in\ RV(CS)\ \forall j\ Then\ \ W_7=1\ ,Otherwise\ W_7=0\\
&Score\ IM(CS_i)=W_5+W_6+W_7\\
&IF\ 2<ScoreIM(CS_i)\le 3\ Then\ IM(CS_i)\in CandidateIM
\end{aligned}
\tag{32}
$$

步驟(A2) 具有觀念性語句之結構強度計算

經由候選語彙集合從中計算集合各語句之結構強度 $Score(TS,TV,TO)$。針對各表達項目之主詞（TS）、動詞（TV）以及受詞（TO）呈現不同，可分為三種語句結構。第一、完整結構語句，即表示該語句具三種結構元素之結合，其該語句之結構之分數 $Score(TS,TV,TO)$ 為 2。第二、半結構語句，假設該語句具有動詞但結合受詞或是主詞，即表示該語句乃須連結其他語句，其該語句之結構之分數 $Score(TS,TV,TO)$ 為 1。第三、不完全結構語句，即表示該語句不具有動詞，同時亦表示該語句難以進行前後語句之連結，其該語句之結構之分數 $Score(TS,TV,TO)$ 為 0，各表達項目之結構強度計算方式如公式(33)至公式(41)所示。

表 4、各表達項目之結構強度計算

觀念性語句	結構性	公式(33)至公式(41)
作業名稱	完整結構	$ONCS_i=\{CS_{i,1},CS_{i,2},CS_{i,3},\cdots,CS_{i,j},\cdots\}$ $IF\ ON(CS_{i,j})\ \ exist\ in\ (PV(CS))\ OR\ \ (GV(CS))\ \forall j\ And\ \ ON(CS_{i,j})\ \ exist\ in\ (OT(CS))\ \ \forall j$ $And\ \ IF\ ON(CS_{i,j})\ exist\ in\ (ON(CS))\ \ \forall j$ (33) $Then\ (Score(TS,TV,TO))=2$
	半結構	$ONCS_i=\{CS_{i,1},CS_{i,2},CS_{i,3},\cdots,CS_{i,j},\cdots\}$ $IF\ ON(CS_{i,j})\ \ exist\ in\ (PV(CS))\ OR\ \ (GV(CS))\ \forall j\ \ And\ \ ON(CS_{i,j})\ \ exist\ in\ (OT(CS))\ \ \forall j$ $Or\ \ IF\ ON(CS_{i,j})\ exist\ in\ (ON(CS))\ \ \forall j$ (34) $Then\ (Score(TS,TV,TO))=1$
	不完全結構	$ONCS_i=\{CS_{i,1},CS_{i,2},CS_{i,3},\cdots,CS_{i,j},\cdots\}$ $IF\ ON(CS_{i,j})\ not\ exist\ in\ (PV(CS))\ OR\ \ (GV(CS))\ \forall j$ (35) $Then\ (Score(TS,TV,TO))=0$

傷害成因	完整結構	$ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\,IC(CS_{i,j})\ existin(PV(CS))\,OR\,(GV(CS))\,\forall j\ And\ IC(CS_{i,j})\ existin(IC(CS))\,\forall j$ $And\ IF\,IC(CS_{i,j})\,existin(IC(CS))\ \forall j$ $Then(Score(TS,TV,TO)) = 2$	(36)
	半結構	$ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\,IC(CS_{i,j})\ existin(PV(CS))\,OR\,(GV(CS))\,\forall j\ And\ IC(CS_{i,j})\ existin(IC(CS))\,\forall j$ $OR\ IF\,IC(CS_{i,j})\,existin(IC(CS))\ \forall j$ $Then(Score(TS,TV,TO)) = 1$	(37)
	不完全結構	$ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\ IC(CS_{i,j})\ not\ exist\ in\ (PV(CS))\,OR\ (GV(CS))\,\forall j$ $Then\ (Score(TS, TV, TO)) = 0$	(38)
改善方法	完整結構	$IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\ IM(CS_{i,j})\ \ exist\ in\ PV(CS)\ OR\ GV(CS)\ OR\ RV(CS)\ \forall j$ $And\ \ IM(CS_{i,j})\ \ exist\ in\ (OT(CS))\ \forall j\ And\ \ IF\ IC(CS_{i,j})\ exist\ in\ (ORT(CS))\ \ \forall j$ $Then\ (Score(TS, TV, TO)) = 2$	(39)
改善方法	半結構	$IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\ IM(CS_{i,j})\ \ exist\ in\ PV(CS)\ \ OR\ \ GV(CS)\ \ OR\ \ RV(CS)\ \ \forall j$ $And\ \ IM(CS_{i,j})\ \ exist\ in\ (OT(CS))\ \forall j\ \ And\ \ IF\ IC(CS_{i,j})\ exist\ in\ (ORT(CS))\ \ \forall j$ $Then\ \ (Score(TS,\ \ TV,\ TO)) = 1$	(40)
	不完全結構	$IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}$ $IF\ IM(CS_{i,j})\ not\ exist\ in\ PV(CS)\ \ OR\ \ GV(CS)\ \ OR\ \ RV(CS)\ \ \forall j$ $Then\ \ (Score(TS,\ \ TV,\ TO)) = 0$	(41)

步驟(A3) 具有觀念性語句之權重值計算

考量結構化摘要之簡略部分主旨乃在於使知識搜尋者以短時間內篩選文件，為突顯語句向心性之重要性，本研究對其向心性與結構性分數各別加予權重值WS_1、WS_2，欲表示簡略性部分之摘要乃以語句之觀念向心性為主要考量，其次為語句之結構性，最後將其具有向心性分數並完整結構之語句擷取至各表達項目之選擇語彙集合，各表達項目之權重計算方式如公式(42)至公式(44)所示。

$$TotalScore\ ON(CS_i) = (Score\ ON(CS_i) * WS_1) + ((Score(TS, TV, TO)) * WS_2)$$
$$IF\ \ LowerLimit < TotalScoreON(CS_i) \le UpperLimit\ Then\ ON(CS_i) \in SelectON \quad (42)$$

$$TotalScore\ IC(CS_i) = (Score\ IC(CS_i) *\ WS_1\ ) + (Score((TS, TV, TO)) * WS_2\ )$$
$$IF\ \ LowerLimit < TotalScoreIC(CS_i) \le UpperLimit\ Then\ IC(CS_i) \in SelectON \quad (43)$$

$$TotalScore\ IM(CS_i) = (Score\ IM(CS_i) * WS_1) + ((Score(TS, TV, TO)) * WS_2)$$
$$IF\ LowerLimit < TotalScoreIC(CS_i) \le UpperLimit\ Then\ IC(CS_i) \in SelectON \quad (44)$$

呈上述步驟其觀念乃將具有作業名稱、傷害成因、改善方法等觀念性語句權重之計算與擷取。根據語句之向心性以及語句之結構完整性作為簡略部份之條件，並以條件之權重分數篩選以及擷取語句，以形成簡略部分結構性摘要。

3.3.2 詳述部分建立

針對詳述摘要推論部分，主要進行文件內容之細部描述。於此，詳述結構化摘要主要加入詞彙鏈概念，以表達文件考量文件流程性與時序性之描述。針對詳述部分所擷取集合包含：作業行為集合、

作業環境集合以及改善方法集合為主。

步驟(B1) 擷取具有作業描述之觀念性語句

依據階段部分可得欲擷取語句集合包含：作業行為集合以及作業環境集合。首先將作業行為觀念性語彙以及各隱含語彙進行TF-ISF字詞相似度計算，TF-ISF字詞權重計算之方法如公式(45)所示，並以作業描述（OD）為判定語彙，並與其他潛在語彙進行詞頻計算乃求得作業描述於各潛在語彙之相似關係（$WOD_{i,j}$），再以較高分數之語彙作為候選語句$Res(S_i)$並儲存於候選語句之集合ResSet中（如公式(46)）。

$$WOD_{i,j} = TFOD_{i,j} \cdot \log \frac{NumSet}{ODF_i} \tag{45}$$

$$Res(S_i) = \{ S_{i,1}, S_{i,2}, S_{i,3}, \cdots, S_{i,j}, \cdots \} \tag{46}$$

步驟(B2) 計算具有觀念性語彙之相依程度

為避免前後文語意之問題，本研究乃以語彙與語彙出現機率關係為基，透過語彙與語彙出現順序之關係，作為語句連結之依據以預測下一順序之語句。即候選語句 $Res(S_i)$與後續連接語句 $Res(S_{i-1})$之連結機率性並以 $P(Res(S_i)|Res(S_{i-1}))$表示之，假設兩語句與具有關聯性即表示連結機率值越大；反之，假設若不具有相關性或不符合語意其機率值越小。

假設在第 i 個段落中存在第 k 個語彙 $CW_k[P_i]$之語彙具有觀念性（以 A(CS)表示之），且下一個出現語彙 $CS_{k+1}[P_i]$亦具有其觀念性（以 B(CS)表示之），則標記 I[i,k]為 1 表示具有順序關係，若為 0 則無關係(如公式(47)所示)。待標記完成後，再將段落中出現之語彙順序關係進行加總並以 $F[CW_k,CWk_{+1}]$表示之，根據公式(48)即可計算語彙 k 至語彙 k+1 之發生頻率。另加入該語彙發生於段落之起始與結束之頻率以判定語彙整體先後順序關係，並如公式(49)至公式(50)所示以 $P^{From}(k)$與 $P^{To}(k)$之比值 $R[CW_k]$表示，並根據比值大小進而排序整體趨勢順序（SCSW）。

$$I[i,k] = \begin{cases} 1, IF\ CW_k[P_i] \in A(CS)\ , CS_{k+1}[P_i] \in B(CS) \\ 0, Otherwise \end{cases} \tag{47}$$

$$F[CW_k, CW_{k+1}] = \sum_{all\ i, all\ k} I\ [i,k] \tag{48}$$

$$P^{From}(k) = \sum_{all\ k+n} F[CW_k, CW_{k+n}], \quad P^{To}(k) = \sum_{all\ k} F[CW_{k+n}, CW_k] \tag{49}$$

$$R[CW_k] = \frac{P^{From}(k)}{P^{To}(k)},\ SCSW = CW'_{k-1} \rightarrow CW'_k \rightarrow \cdots \rightarrow CW'_{k+n} \rightarrow \cdots$$
$$Where\ R[CW'_{k-1}] \geq R[CW'_k] \geq \cdots \geq R[CW'_{k+n}] \geq \cdots \tag{50}$$

步驟(B3) 計算與擷取觀念性語句

透過上述步驟即可以$R[CW_k]$與該段落存在語彙之比值判定該語彙。若其比值為最大值$MaxR[CW_k]$即表示該語彙為起之語彙，若為最小值$MinR[CW_k]$即為結束語彙之可能性。於本步驟乃以$Res(S_{i,j})$判斷該語句是否符合結構化摘要建立之條件，並以結構化摘要符合標記值$Mk_{i,j}$表示之，其原則以及相對應之公式如公式(51)至公式(56)所示。具有整體趨勢順序排序SCSW之語彙k，則結構化摘要符合標記值$Mk_{i,1}$=1；存在連結語彙C(CS)，則標記值$Mk_{i,2}$=1; 存在銜接語彙LinkC(CS)則標記值$Mk_{i,3}$=1;存在語彙之評斷係數$R[CW_k]$與出現語彙總數量之比值為最大值，則標記值Mki,4=1;存在語彙之評斷係數$R[CW_k]$與出現語彙總數量之比值為最小值，則標記值$Mk_{i,5}$=1; 存在結尾符號集合中EndMark_Set，則標記值$Mk_{i,6}$=1。

$$MK_{i,1} = \begin{cases} 1, IF\ Res(S_{i,j})\ exit\ in\ SCSW\ \forall j \\ 0, Otherwise \end{cases} \tag{51}$$

$$MK_{i,2} = \begin{cases} 1, IF\ Res(S_{i,j})\ exit\ in\ C(CS)\ \forall j \\ 0, Otherwise \end{cases} \tag{52}$$

$$MK_{i,3} = \begin{Bmatrix} 1, IF\ \mathrm{Re}s(S_{i,j}) exit\ in\ LinkC(CS)\ \forall j \\ 0, Otherwise \end{Bmatrix} \quad (53)$$

$$MK_{i,4} = \begin{Bmatrix} 1, IF\ \mathrm{Re}s(S_{i,j})\ exit\ in\ \mathrm{CW_k}\ and\ \mathrm{R[CW_k]} = Max\mathrm{R[CW_k]}\ \forall j \\ 0, Otherwise \end{Bmatrix} \quad (54)$$

$$MK_{i,5} = \begin{Bmatrix} 1, IF\ \mathrm{Re}s(S_{i,j})\ exit\ in\ \mathrm{CW_k}\ and\ \mathrm{R[CW_k]} = Min\mathrm{R[CW_k]}\ \forall j \\ 0, Otherwise \end{Bmatrix} \quad (55)$$

$$MK_{i,6} = \begin{Bmatrix} 1, IF\ \mathrm{Re}s(S_{i,j})\ exit\ in\ EndMark\_Set\ \forall j \\ 0, Otherwise \end{Bmatrix} \quad (56)$$

如公式(57)所示，根據結構化摘要符合標記值$Mk_{i,j}$彙整並作為擷取候選語句之考量，假設該候選語句之$Mk_{i,4}$為1即表示具有摘要之開頭，透過公式(58)與公式(59)以$Mk_{i,2}$與$Mk_{i,3}$之加總$Sum1(Res(S_i))$作為描述階段之篩選，最後則以$Mk_{i,5}$與$Mk_{i,6}$之加總$Sum2(Res(S_i))$作為評估階段之擷取根據。最後，以公式(60)SCSW結果作為主要依據，並根據其他結構化摘要建立之條件（即連結語彙或是屬於開頭、結束語彙），作為為摘要歸屬位置之輔佐。其詳述部份結構化摘要擷取語句以及各語句歸屬位置。

$$MK = \begin{bmatrix} \mathrm{MK_{1,1}} & \mathrm{MK_{2,1}} & \ldots & \mathrm{MK_{i,1}} \\ \mathrm{MK_{1,2}} & \mathrm{MK_{2,2}} & \ldots & \mathrm{MK_{i,2}} \\ \mathrm{MK_{1,3}} & \mathrm{MK_{2,3}} & \ldots & \mathrm{MK_{i,3}} \\ \mathrm{MK_{1,4}} & \mathrm{MK_{2,4}} & \ldots & \ldots \\ \mathrm{MK_{1,5}} & \mathrm{MK_{2,5}} & \ldots & \mathrm{MK_{i,5}} \\ \mathrm{MK_{1,6}} & \mathrm{MK_{2,6}} & \ldots & \mathrm{MK_{i,6}} \end{bmatrix} \quad (57)$$

$$Sum1(\mathrm{Re}s(S_i)) = \sum_{j=2}^{3} MK_{i,j} \quad (58)$$

$$Sum2(\mathrm{Re}s(S_i)) = \sum_{j=5}^{6} MK_{i,j} \quad (59)$$

$$SCSW = \mathrm{CW'_{k-1}} \rightarrow \mathrm{CW'_k} \rightarrow \cdots \rightarrow \mathrm{CW'_{k+n}} \rightarrow \cdots \quad Where\ \mathrm{R[CW'_{k-1}]} \geq \mathrm{R[CW'_k]} \geq \cdots \geq \mathrm{R[CW'_{k+n}]} \geq \cdots \quad (60)$$

4. 系統應用流程

針對前一章節所發展之方法論與模式，本研究乃開發一套整合知識問答解析與文件結構化摘要技術之知識文件檢索系統，以確認方法論與模式之可行性。此系統之功能重點乃透過使用者上傳知識文件，系統管理者先行設定系統參數與新增詢問詞與回覆詞，進而執行知識文件表達項目分析、問答解析與結構化摘要等主功能。本章即針對本研究所提之「整合知識問答解析與文件結構化摘要技術之知識文件檢索系統」，分別以系統核心架構、系統應用等主題進行說明。

4.1整合知識問答解析與文件結構化摘要技術之知識文件檢索系統之流程架構

本研究所開發之系統依其運作流程可分為「知識文件上傳」、「知識文件表達項目解析」、「知識文件問答解析」及「知識文件結構化摘要推論」等階段，此系統之運作流程架構如圖4所示，各功能層次之詳細流程說明如下。

4.2 系統應用

為驗證知識文件結構化摘要系統於實務應用之可行性，本研究乃以勞工安全衛生知識網之改善報告為案例驗證樣本，並以知識文件結構化摘要系統之核心功能模組，以評估本研究所發展之方法論與開發系統之可行性。使用者可透過「知識文件管理模組」以進行「知識文件上傳」功能。待知識文件上傳之後，系統即可透過「知識文件表達項目解析模組」擷取知識文件各表達項目之觀念性語句，如圖5與圖6所示，再根據表達項目擷取相關之觀念性語句，如表達項目「作業身分」之觀念性語句包含「…作業人員合力抬舉…」等，同時取得文件之關鍵字詞「儲存箱」、「搬運」等，並以此為基連結先行分析詢問目的以及相關回覆詞組合，再建立結構化摘要。如圖7所示，根據詢問詞、回覆句與回覆

詞配對解析，取得詢問詞「酸痛」相關回覆詞為「營造業」與「年紀」等，如圖8所示即可根據詢問目的為「酸痛」，並以相關回覆詞「營造業」與「年紀」等進而連結相關文件，或是透過結構化摘要以進行文件篩選，於簡略結構化摘要乃計算向心性分數與結構性分數，如「同時作業人員也可不必以彎腰的…」加權後之加總分數「12」，再根據權重值依照「文件名稱」、「作業名稱」等建立簡略結構化摘要（如圖9）。於詳述結構化摘要推論，即透過係數即可形成詞彙鏈為「庫房、搬運、…、主要問題等」進而依序「作業描述」、「問題描述」以及「改善方式」取得詳述結構化摘要（如圖10）。

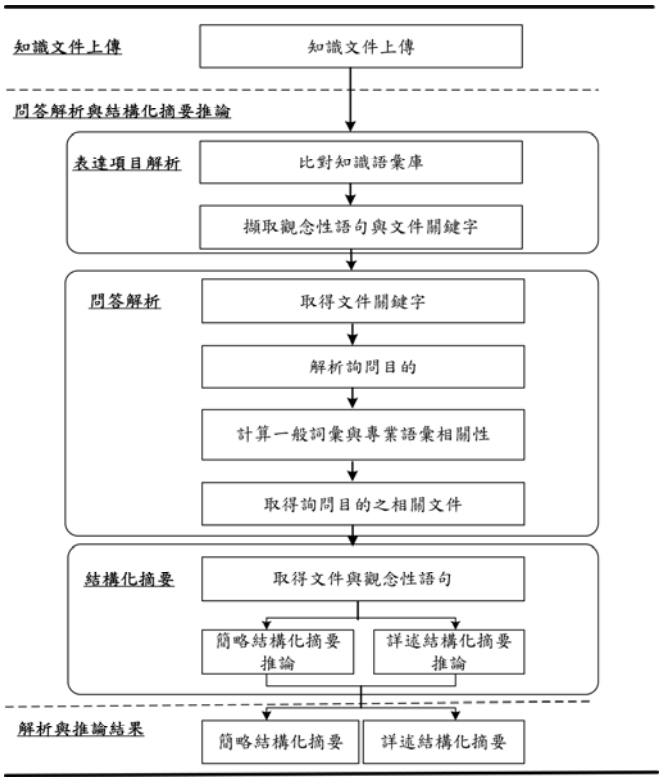

圖 4、整合知識問答解析與文件結構化摘要技術之知識文件檢索系統之流程架構

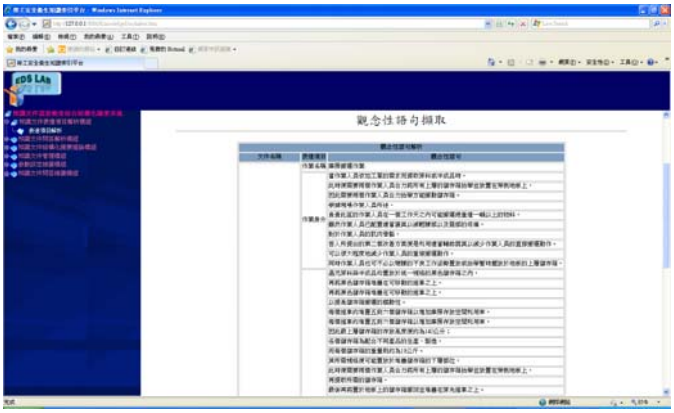

圖 5、知識文件表達項目解析結果(1)

圖 6、知識文件表達項目解析結果(2)

圖 7、知識文件問答解析結果(1)

圖 8、知識文件問答解析結果(2)

圖 9、簡略結構化摘要結果

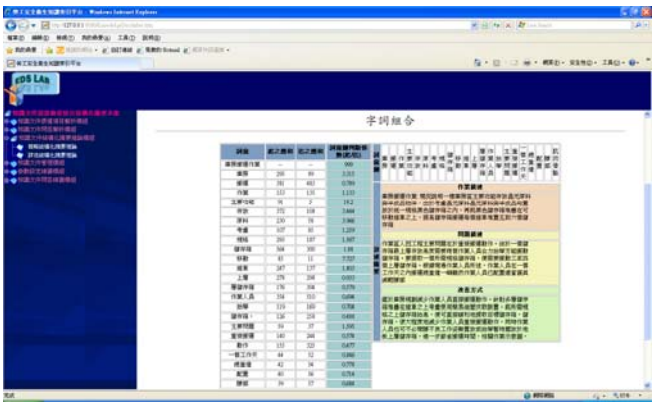

圖 10、詳述結構化摘要結果

5. 系統驗證與評估

本研究乃以「整合知識問答解析與文件結構化摘要技術之知識文件檢索推論」為基，開發一套「整合知識問答解析與文件結構化摘要技術之知識文件檢索系統」，並以勞工安全衛生知識網作為應用情境，將研究報告與技術叢書作為知識文件並以「知識文件問答解析」與「知識文件結構化摘要」等主題進行系統績效驗證，以下乃針對本系統各主題之驗證方式（即「驗證資料說明」、「驗證方式說明」與「驗證指標定義」）依序說明，以驗證「知識文件檢索系統」績效與準確性。

驗證資料取得

針對「知識文件檢索系統」之驗證，本研究乃先行勞工安全衛生研究所（勞研所）之網際論壇取得知識搜尋者之詢問句，並彙整成「知識搜尋者詢問句資料表」（如表5所示）。此外，勞研所亦針對專有名詞乃建立雙語詞彙，並說明該詞彙之解釋與意義，本研究即以勞研所所提供之雙語詞彙之解釋作為驗證實際檢索回覆詞（如表6所示）。針對知識文件結構化摘要之驗證資料，本研究乃先解析勞工安全衛生知識網研究報告或技術叢書中案例作業，藉由案例作業分析即可將各表達項目與相關語彙彙整成「知識文件表達項目資料表」（如表7所示）。

知識文件檢索系統整體驗證方式

針對「知識文件檢索系統」本研究乃提出二個課題以進行系統績效驗證，其課題分別為「(A)知識文件問答解析」與「(B)知識文件結構化摘要推論」等並分別設計指標以進行系統驗證分析，以下乃針對各課題分別細述與說明驗證設計、指標定義及驗證績效。

表 5、知識搜尋者詢問句資料表

題號	詢問句	知識搜尋者	詢問時間
1	作業作業環境測定項目	佛曰不可說...	2009/12/4
2	關於升降機	japanesesky	2007/10/17
3	請問有關緊急照明的問題	何志強	2010/10/11
4	請問有沖身洗眼器的檢查表嗎	loveven.tw	2010/9/9
5	想請問各位先進, 關於[惰性氣體]和[易燃氣體],	流星	2012/2/1

表 6、勞工安全衛生研究所定義之雙語詞彙資料表

類別	中文詞彙	English	簡稱	附註說明	出處	相關詞
A	擦傷	ABRASION	--	覆蓋於身體的表面組織（如皮膚或黏膜）被擦掉或刮掉	美國安全工程師協會	表面組織、皮膚、黏膜、擦掉、刮掉
A	容許負載	ALLOWABLE LOAD	--	請參考 Load Weight（承載重量）。	美國安全工程師協會	承載重量
M	機械防護	MACHINE GUARDING	--	在機械裝上設備或機具以…。參考 Barrier Guard（防護圍柵）。	美國安全工程師協會	機具、防護圍柵

表 7、知識文件表達項目資料表

表達細項		礦泉水運送作業	庫房搬運作業
作業領域		--	--
作業名稱		礦泉水運送作業	庫房搬運作業
作業身分	作業員性別	--	--
	作業員年齡	--	--
作業環境	設備佈置	輸送帶	--
		--	--
	工具介紹	拖車、木箱、水桶	黑色儲存箱、推車
作業行為	作業目的	主要工作是將輸送帶上已完成填灌作業…運至拖車上的木箱中	主要功能為存放晶元原料與半成品物件，以為後續製造、加工之用
	作業描述	礦泉水運送作業的主要工作是將輸送帶…貨車運送至各部門	將黑色儲存箱堆疊在可移動的推車之上，…增加庫房存放空間利用率
	作業姿勢	抬舉、搬運、高舉、彎腰	搬運、抬舉
	專業動詞	裝滿礦泉水的水桶重約為 20 公斤	儲存箱的重量約為 18 公斤…
作業時間	作業次數/天	一天必須至少搬運 200 桶、有時候則可能多達 500 桶	
	作業時間/次	--	--
	作業距離/次	輸送帶與拖車之間的最近距離為 1.5 公尺…	--
傷害成因	傷害因素	施力過大	重複搬運
	傷痛部位	手臂部分、背部	腰部、肩部
	改善目的	有效地舉高或放下礦泉水桶並改變其水平位置，…於拖車上的木箱中	大程度地減少作業人員的重複搬運動作，…，可以進一步節省搬運時間，增加工作效率
改善方法	改善工具	象鼻子真空省力裝置	油壓夾取裝置
	改善流程	只要以此裝置的吸物盤…整氣閥的流量	考慮使用簡易的油壓夾取裝置，…，完全避免搬動額外的儲存箱
	改善評估	負荷明顯降低部位	負荷明顯降低部位
	改善費用	--	--

(A)知識文件問答解析驗證方式說明

於此課題本研究乃以勞工安全衛生研究所之網際論壇中，以知識搜尋者之詢問句作為測試資料，並以勞研所制定之雙語詞彙作為實際檢索回覆詞，以驗證知識文件問答解析之績效。

本研究乃規劃兩階段之系統驗證，於系統驗證第一階段乃於800筆詢問句與回覆句組合中，隨機選取200筆組合作為訓練資料，並利用訓練階段所取得之文件相關語意詞，再從中隨機挑選10句詢問句，以作為系統測試資料，藉以觀察系統問答解析是否符合實際相關字詞，以確認本研究所提方法論之正確性。於系統測試第二階段（即系統測試階段），於此階段乃分6個週期（每週期皆匯入100筆不重複之詢問句與回覆句配對組合，共計600筆組合），藉由持續匯入以分析系統於不同訓練測試資料下之解析效果，於各週期中乃利用前述10句詢問句重新進行知識文件問答解析，以分析系統之長期學期趨勢。

(A-1) 知識文件問答解析評估指標定義

藉由本系統所推論之回覆詞與實際檢索回覆詞進行驗證系統準確性，於此階段績效指標包含「回覆詞召回率」與「回覆詞準確率」。問答解析之回覆詞召回率即表示經由系統問答解析後能取得正確回覆詞之解析能力，而問答解析之回覆詞正確率即表示經由問答解析後能取得正確問答回覆詞之正確能力。

(A-2) 知識文件問答解析驗證結果與分析

於第一階段測試中，以 800 份文件作為訓練資料之基礎下，將其問答資料包含受測詢問句、透過詢問句取得之回覆句等資料匯入系統（如表 8），再隨機抽取 10 句詢問句作為測試資料，其第一階段測試結果乃彙整於表 9 所示。藉由統計結果以得知回覆詞召回率其平均為 31.63%，而回覆詞正確率之平均為 43.33%。整體而言，其回覆詞召回率與正確率績效，尚無法準確判斷知識文件之問答。

表 8、知識文件問答解析之詢問句與回覆句組合

詢問句	相關回覆句	回覆句數
台灣目前製造業意外事故調查步驟	近年來發生多起建築工程斜籬安裝時作業勞工墜落的意外事故，	5
	安裝時勞工常會站到斜籬上施作而無適當防護，墜落事故時有所聞。	
	提供適當之安全防護措施，並督促勞工佩帶個人安全防護具。	
	因腳踩踏於斜籬骨架，…罹災者雖繫安全帶但並未勾掛於母索上。	
	應設置防止物體飛落之設備，…，其中斜籬即為防止物體飛落之設備。	
疲勞程度規範	疲勞是當前…的職業安全與衛生議題之一，…，進而造成產能的損失…	3
	許多人經常認為司機在開車時入睡即表示其疲勞。	
	疲勞通常是用來描述類似「想睡」、…，自己疲倦的程度失去判斷能力	
酒精濃度含量多少會導致意外發生	當血液中酒精含量達 0.1%時，人的動作協調、視覺…出現中毒現象。	3
	當血液中酒精含量達 0.5%時，神經生理平衡會嚴重受損而且失去意識，	
	廚師罹患高比率的口腔癌與罹患高比率可能是由於高量飲酒所引起。	

表 9、知識文件問答解析第一階段之受測詢問句結果分析(1)

受測詢問句	相關詞 1	相關詞 2	相關詞 3	實際數	推論數	正確數	召回率	正確率
關於升降機	工作情況	工作場所	-	2	1	0	0.00%	0.00%
請問有關緊急照明的問題	安全法規	電扶梯	-	2	3	1	50.00%	33.33%
請問有沖身洗眼器…	照度	光通量	-	2	2	1	50.00%	50.00%
想請問…,關於[惰性氣體]和[易燃氣體],	沖洗水	沖洗軟管	-	2	2	1	50.00%	50.00%
為避免局部振動危害，在…有那些方法	不活性	氦、氖、氩、氪、氙	-	2	1	0	0.00%	0.00%
請問三公噸以下固定式起重機	壓路機	重型營建車輛	-	2	2	1	50.00%	50.00%
根本找不到母索...安全帶要何用...	鉸鏈	手搖桿	動力裝置	3	2	1	33.30%	100.00%
請問：事故調查之原因分析	固定	懸吊	碰撞	3	2	1	33.00%	50.00%
如果與鄰地有落差，該設置欄杆？	作業場所	實際原因	近似原因	3	1	0	0.00%	0.00%
如果與鄰地有落差，…該設置欄杆嗎？	護欄	扶手	-	2	1	1	50.00%	100.00%
平均值							31.63%	43.33%

於第二階段驗證中乃分 6 個週期，並以第一階段驗證時所選取之 10 筆詢問句與回覆句組合重新進行系統績效測試，其資料相關結果可整理如表 10，並可得知平均每週期知識文件問答解析召回率與正確率之整體平均成長率分別為 9.17%及 6.81%，故可知系統乃隨著不同週期及訓練組合句之增加下，各項數據呈現顯著之成長，且知識文件問答解析乃具學習能力。

表 10、知識文件問答解析績效彙整

知識文件問答解析		各週期知識文件問答解析驗證－詢問句與回覆句組合數量							平均
		第一階段	第二階段						
		第一週期	第二週期	第三週期	第四週期	第五週期	第六週期	第七週期	
		200 筆	300 筆	400 筆	500 筆	600 筆	700 筆	800 組	
召回率	平均值	31.63%	48.27%	61.63%	70.00%	78.33%	86.67%	86.67%	66.17%
	標準差	22.84%	9.58%	22.34%	21.94%	23.64%	21.94%	21.94%	20.60%
	成長率	-	16.64%	13.36%	8.37%	8.33%	8.34%	0%	9.17%
正確率	平均值	43.33%	50.00%	58.33%	66.67%	71.67%	84.17%	84.17%	65.56%
	標準差	37.02%	7.86%	16.20%	19.25%	20.86%	21.68%	21.68%	20.63%
	成長率	-	6.67%	8.88%	7.79%	5.00%	12.50%	0%	6.81%

最後，綜合兩階段之驗證成效，各項驗證指標之相關結果整理如表 11 所示。由表 11 整理結果可得知，各項驗證指標之「收斂前每週期平均成長率」及「整體每週期平均成長率」皆為正數成長，且各項驗證指標皆於六週期內呈現收斂狀態，因此，當系統使用約 700 筆訓練詢問句與回覆句組合時，可使系統各項推論指標績效提升至 85%之水準。故本研究所建置之「知識文件問答解析模組」確實能幫助知識搜尋者取得專業語彙與一般詞彙之關聯性。

表 11、知識文件問答解析綜合兩階段之驗證績效彙整

驗證指標	整體平均值	收斂週期	收斂前每週期 平均成長率	整體每週期 平均成長率
知識文件問答解析召回率	66.17%	第六週期	11.01%	9.17%
知識文件問答解析正確率	65.56%	第六週期	8.17%	6.81%

(B) 知識文件結構化摘要驗證方式說明

本研究乃以每份知識文件之案例作業作為一份文件，並以二十位知識搜尋者作為測試者，以了解結構化摘要呈現滿意度與表達準確性。針對測試者部分，本研究乃邀請二十位大專院校學生（十位女性與十位男性）作為知識搜尋者以進行測試。首先，本研究先行提供文件之原文內容以供受測者閱讀，藉由閱讀方式使測試者了解文件之內容與大綱。待閱讀完後，將本系統所推論之階段性成果給予測試者閱讀，根據測試者之閱讀主觀感受以進行滿意度評級。

對此，本研究乃規劃兩階段之系統驗證，於系統驗證第一階段乃於500份知識文件中，隨機選取100筆作為訓練資料，並利用訓練階段所取得之文件結構化摘要，再從中隨機挑選15份知識文件，以作為系統測試資料，藉以確認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並分5個週期（每週期皆匯入80筆不重複之知識文件，共計400筆組合），藉由持續匯入以分析系統於不同訓練測試資料下之解析效果，並從中了解知識文件結構化摘要之長期學習趨勢。

(B-1) 知識文件結構化摘要評估指標定義

本研究乃設計「摘要滿意度指標」以及「摘要文字呈現指標」作為驗證指標以檢視本系統結構化摘要推論效果。針對「摘要滿意度指標」乃以使用者觀感進行績效驗證，本研究乃參考郭佩慧（2006），將文件重點（動機與目的明確）、文句結構（閱讀性）與字數（字數長度）作為閱讀摘要之評級標準。另外亦根據詹元智（2002）將評級（摘要字數呈現合宜、略多或略少與過多或過少；動機與目的非常明確、大致了解與不明確；內容順暢、有順序性、大致尚可與內容語意不清）記分採用1、3與5分以保留彈性空間，使測試者可於確定問項內容後再予以訂定給分。而「摘要文字呈現指標」乃驗證摘要之字數呈現，本研究參照Kuo等人（2002），將摘要文字最佳效果設定為原文百分之十至百分之十五，並作為簡略摘要之標準，而詳述摘要標準為原文百分之五十，若字數為標準範圍正負10字則為合宜（評分為5），若為正負30字或超過正負30字，將分數設為3與1。

(B-2) 知識文件結構化摘要驗證結果與分析

於第一階段之系統驗證中，隨機抽取15份知識文件作為測試資料（如表12所示），於第一階段測試結果其使用者滿意度約落於1.00與1.67，其平均滿意度為1.52。另外簡略結構化摘要文字呈現部分彙整於表13，其文字呈現效果平均為1.67，而詳述結構化摘要呈現效果平均為1.27。整體而言，根據滿意度指標可得知無論是使用者滿意度與文字呈現效果皆低於滿意度之平均值。

表 12、知識文件結構化摘要驗證之測試資料（其中 3 份）

文件名稱	作業名稱	文件內容	文件字數
人因工程現場不良工作姿勢改善績效評估研究-貝果運送作業	貝果運送作業	現況說明，此作業主要是將烘烤完成的貝果經由輸送帶運送到再加工區，…，此輸送帶並…，在爬升運送時，貝果…	905
工作現場人因工程危害預防效益研究-氣泡料攪拌作業	氣泡料攪拌作業	氣泡料攪拌作業:現況說明泡料室攪拌作業，通常是一人獨立完成所有工作。…，攪拌機圓桶高 94 公分，直徑約 89 公分，空…	913
人因工程工作場所改善方法-切塊作業	切塊作業	切塊作業分切雞塊作業是作業員將經由輸送帶運送到面前的盤裝雞肉拿起，…。首先，供料作…	847

表 13、第一階段知識文件之摘要呈現效果指標績效彙整

文件作業名稱	原文件字數	簡略摘要呈現效果指標績效				詳述摘要呈現效果指標績效		
		理想摘要字數範圍		結構化摘要字數	文字呈現效能	理想摘要字數	結構化摘要字數	文字呈現效能
貝果運送作業	905	91	136	206	1	453	513	1
氣泡料攪拌作業	913	91	137	184	1	457	517	1
切塊作業	847	85	127	125	5	424	517	1
馬達底座加工作業	639	64	96	191	1	320	470	1
模具加工作業	788	79	118	157	1	394	576	1
包裝重量檢驗作業	586	59	88	132	1	293	354	1
顏料桶上料作業	784	78	118	210	1	392	468	1
裁切作業	261	26	39	143	1	131	149	3
人工下料作業	528	53	79	126	1	264	374	1
鋼捲包裝作業	950	9	143	194	1	475	632	1
礦泉水運送作業	708	71	106	197	1	354	474	1
布捲吊掛作業	934	93	140	178	1	467	589	1
庫房搬運作業	930	93	140	126	5	465	492	1
座椅裝配作業	598	60	90	108	3	299	424	1
大理石補膠作業	816	82	122	196	1	408	431	3
平均值（簡略摘要）					**1.67**	**平均值（詳述摘要）**		**1.27**

於第二階段以每週期增加 80 份知識文件為單位，並以第一階段驗證時所選取之 15 份知識文件重新進行系統績效測試，而各驗證週期知識文件結構化摘要之「摘要滿意度」與「文字呈現效果」績效分佈趨勢如圖 11 與圖 12 所示。平均每週期使用者滿意度由 1.72 成長至 4.35，而文字呈現效果簡略部分由 1.93 成長至 3.40，詳述部分由 1.80 成長至 3.21，故可得知本系統結構化摘要推論良好，且具學習能力。

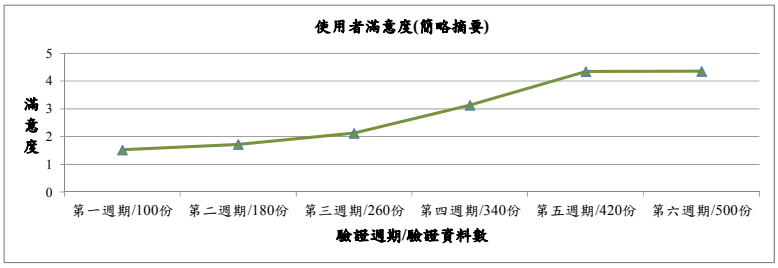

圖 11、各驗證週期之結構化摘要滿意度之分佈趨勢

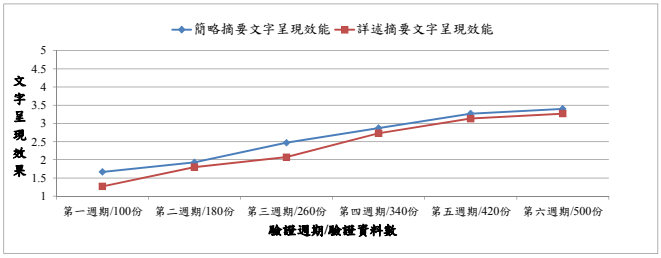

圖 12、各驗證週期之結構化摘要文字呈現效果之分佈趨勢

最後，綜合兩階段之驗證成效，各項驗證指標之相關結果整理如表14。各項驗證指標皆為正數成長，且皆於第五週期內呈現收斂狀態，因此，當系統使用約420份知識文件時，可使系統各項推論指標績效提升至滿意度3以上之水準。

表 14、知識文件結構化摘要綜合兩階段之驗證績效彙整

驗證指標	整體平均值	收斂週期	收斂前每週期 平均成長率	整體每週期 平均成長率
簡略摘要使用者滿意度	2.86	第五週期	0.71%	0.57%
簡略摘要文字呈現效果	2.60	第五週期	0.4%	0.35%
詳述摘要文字呈現效果	2.38	第五週期	0.47%	0.4%

6. 結論

由於特定知識領域過於專業化，導致一般使用者無法明確定義關鍵字詞，因而提高文件搜尋障礙與時間；此外，文件摘要乃自由形式呈現且無字數之控管，容易因個人瀏覽、閱讀喜好而影響文件篩選，進而降低知識網站分享知識功能。因此，本研究提出一套整合知識問答解析與文件結構化摘要技術之知識文件檢索方法論，同時亦建置系統以確認方法論與模式之可行性。藉由本系統運算判定後，即可進行知識文件語意問答，透過問答語意解析將一般詞彙與專業語彙進行語意關聯，知識搜尋者即可以一般詞彙進行搜尋，並取得相關知識文件，再以文件結構化摘要方式將自由形式知識文件轉為制式化之文件摘要，加強使用者文件閱讀與篩選，於此即可協助勞工安全衛生知識網等領域網站，使搜尋者省力且快速方式獲得資訊，進而提高知識領域搜尋效能。

二、參考文獻

1. Benedí, J. M. and Sánchez, J. A., 2005, “Estimation of stochastic context-free grammars and their use as language models” *Computer Speech & Language*, Vol. 19, No. 3, pp. 249-274.
2. Bollegala, D., Okazaki, N. and Ishizuka, M., 2010, “A bottom-up approach to sentence ordering for multi-document summarization,” *Information Processing and Management*, Vol. 46, No. 1, pp. 89-109.
3. Bouras, C., Poulopoulos, V. and Tsogkas, V., 2008, “PeRSSonal’s core functionality evaluation: enhancing text labeling through personalized summaries,” *Data&Knowledge Engineering*, Vol. 64, No. 1, pp. 330-345.
4. Cao, Y. G., Liu, F., Simpson, P., Antieau, L., Bennett, A. Cimino, J. J., Ely, J. and Yu, H., 2011, “AskHERMES: An online question answering system for complex clinical questions,” *Journal of Biomedical Informatics*, Vol. 44, No. 2, pp. 277-288.
5. 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.
6. Dalmas, D. and Webber, B., 2007, “Answer comparison in automated question answering,” *Journal of Applied Logic*, Vol. 5, No. 1, pp. 104-210.
7. Dorr, B. J. and Gaasterland, T., 2007, “Exploiting aspectual features and connecting words for summarization-inspired temporal-relation extraction,” *Information Processing and Management*, Vol. 43, No. 6, pp. 1681-1704.
8. 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*, Vol. 43, No. 6, pp. 1588-1605.
9. Elhadad, N., Kan, M. Y., Klavans, J. L. and McKeown, K. R., 2005, “Customization in a unified framework for summarizing medical literature,” *Artificial Intelligence in Medicine*, Vol. 33, No. 2, pp. 179-198.
10. Erdogan, H., Sarikaya, R., Chen, S. F., Gao, Y. and Picheny, M., 2005, “Using semantic analysis to improve speech recognition performance,” *Computer Speech & Language*, Vol. 19, No. 3, pp. 321-343.
11. Guo, Q. L. and Zhang, M., 2009, “Question answering based on pervasive agent ontology and semantic web,” *Knowledge-Based Systems*, Vol. 22, No.6, pp. 443-448.
12. Hahu, U., Romacker, M. and Schulz, S., 2002, “MEDSYNDIKATE-a natural language system for the extraction of medical information from findings reports,” *International Journal of Medical Informatics*, Vol. 67, No. 1-3, pp. 63-74.
13. Han, K. S., Song, Y. I., Kim, S. B. and Rim, H. C., 2007, “Answer extraction and ranking strategies for definitional question answering using linguistic features and definition terminology,” *Information Processing & Management*, Vol. 43, No. 2, pp. 353-364.
14. He, Y. H., Hui, S. C. and Quan, T. T., 2009, “Automatic summary assessment for intelligent tutoring systems,” *Computers & Education*, Vol. 53, No. 3, pp. 890-899.

15. Huang, M., Zhu, X. and Li, M., 2006, "A hybrid method for relation extraction from biomedical literature," *International Journal of Medical Informatics*, Vol. 75, No. 6, pp. 443-455.
16. Jones, M., Love, B. C., 2007, "Beyond common features: The role of roles in determining similarity," *Cognitive Psychology*, Vol. 55, No. 3, pp. 196-231.
17. Jung, W., Ko, Y. and Seo, J., 2005, "Automatic text summarization using two-step sentence extraction," *Lecture Notes in Computer Science*, Vol. 3411, pp. 71-81.
18. Ko, Y., Kim, K. and Seo, J., 2003, "Topic keyword identification for text summarization using lexical clustering," *IEICE transactions on information and systems*,Vol. E86-D, No. 9, pp. 1695-1701
19. Ko, Y., Park, J. and Seo, J., 2004, "Improving text categorization using the importance of sentences," *Information Processing and Management*, Vol. 44, No. 1, pp. 65-79.
20. 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.
21. Legara, E. F., Monterola, C. and Abundo, C., 2011, "Ranking of predictor variables based on effect size criterion provides an accurate means of automatically classifying opinion column articles," *Physica A: Statistical Mechanics and its Applications*, Vol. 390, No. 1, pp. 110-119.
22. Li, Q. and Chen, Y. P., 2010, "Personalized text snippet extraction using statistical language models," *Pattern Recognition*, Vol. 43, No. 1, pp. 378-386.
23. Lin, F. R. and Liang, C. H., 2008, "Storyline-based summarization for news topic retrospection," *Data & Knowledge Engineering*, Vol. 45, No. 3, pp. 473-490.
24. Ling, X., Jiang, J., He, X., Mei, Q., Zhai, C. and Schatz, B., 2007, "Generating gene summaries from biomedical literature: A study of semi-structured summarization," *Information Processing and Management*, Vol. 43, No. 6, pp. 1777-1791.
25. Lorch, R. F., Lorch, E. P., Ritchey, K., McGovern, L. and Coleman, D., 2001, "Effects of headings on text summarization," *Contemporary Educational Psychology*, Vol. 26, No. 2, pp. 171-191.
26. Mingxin, W., 2011, "An architecture of information retrieval towards the semantic Web," *Energy Procedia*, Vol. 11, No. 10, pp. 4857-4861.
27. Moens, M. F., 2007, "Summarizing court decisions," *Information Processing and Management*, Vol. 43, No. 6, pp. 1748-1764.
28. Moens, M. F., Angheluta, R. and Dumortier, J., 2005, "Generic technologies for single- and multi-document summarization," *Information Processing and Management*, Vol. 41, No. 3, pp. 569-586.
29. Moreda, P., Llorens, H., Saquete, E. and Palomar, M., 2011, "Combining semantic information in question answering systems," *Data & Knowledge Engineering*, Vol. 47, No. 6, pp. 870-885.
30. Nomoto, T. and Matsumoto, Y., 2001, "A new approach to unsupervised text summarization," *Proceedings of the 24th International Conference on Research in Information Retrieval*, pp. 26-34.
31. Oh, H. J., Myaeng, S. H. and Jang, M. G., 2012, "Effects of answer weight boosting in strategy-driven question answering," *Information Processing and Management*, Vol. 48, No. 1, pp. 83-93.
32. Oh, H. J., Sung, K. Y., Jang, M. G. and Myaeng, S. H., 2011, "Compositional question answering: A divide and conquer approach," *Information Processing and Management*, Vol. 47, No. 6, pp. 808-824.
33. Oliva, J., Serrano, J. I. Castillo, M. D. and Iglesias, A., 2011, "SyMSS: A syntax-based measure for short-text semantic similarity," *Data&Knowledge Engineering*, Vol. 70, No. 4, pp. 390-405.
34. Ouyang, Y., Li, W., Li, S. and Lu, Q., 2011, "Applying regression models to query-focused multi-document summarization," *International Journal of Medical Informatics*. Vol. 47, No. 2, pp. 227-237.
35. Rindflesch, T. C. and Fiszman, M., 2003, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text," *Biomedical Informatics*, Vol. 36, No. 6, pp. 462-477.
36. Ruiz-Casado, M., Alfonseca, E. and Castells, P., 2007, "Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia," *Data & Knowledge Engineering*, Vol. 61, No. 3, pp. 484-499.
37. Scheffer, T., 2004, "Email answering assistance by semi-supervised text classification," *Intelligent Data Analysis*, Vol. 8, No. 5, pp, 481-493.
38. Steinberger, J., Poesio, M., Kabadjov, M. A. and Jezek, K., 2007, "Two uses of anaphora resolution in summarization," *Information Processing and Management*, Vol. 43, No. 6, pp. 1663-1680.

39. Sweeney, S., Crestani, F. and Losada, D. E., 2008, "'Show me more': Incremental length summarisation using novelty detection," *Information Processing and Management*, Vol. 44, No. 2, pp. 663-686.
40. Teng, C., Xiong, N., He, Y., Yang, L. T. and Liu, D., 2010, "A behavioural mode research on user-focus summarization," *Mathematical and Computer Modelling*, Vol. 51, No. 7-8, pp. 985-994.
41. Terol, R. M., Martínez-Barco, P. and Palomar, M., 2007, "A knowledge based method for the medical question answering problem," *Computers in Biology and Medicine*, Vol. 37, No. 10, pp. 1511-1521.
42. Uzuner, O., Mailoa, J., Ryan, Y. and Sibanda, T., 2010, "Semantic relations for problem-oriented medical records," *Artificial Intelligence in Medicine*, Vol. 50, No. 2, pp. 63-73.
43. Vanderwende, L., Suzuki, H., Brockett, C. and Nenkova, A., 2007, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Information Processing and Management*, Vol. 43, No. 6, pp. 1606-1618.
44. Xie, S., Liu, Y., 2010, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, Vol. 24, No. 3, pp. 495-514.
45. Yang, C. C. and Wang, F. L., 2007, "An information delivery system with automatic summarization for mobile commerce," *Decision Support Systems*, Vol. 43, No. 1, pp. 46-61.
46. Yangarber, R., Grishman, R., Tapanainen, P. and Huttunen, S., 2000, "Automatic acquisition of domain knowledge for information extraction," *Association for Computational Linguistics*, Vol. 2, pp. 940-946.
47. Yangarber, R., 2003, "Counter-training in discovery of semantic patterns," *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 343-350.
48. Ye, S., Chua, T. S., Kan, M. Y. and Qiu, L., 2007, "Document concept lattice for text understanding and summarization," *Information Processing and Management*, Vol. 43, No. 6, pp. 1643-1662.
49. Yeh, J. Y., Ke, H. R., Yang, W. P. and Meng, I.H., 2005, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management*, Vol. 41, No. 1, pp. 79-95.
50. Yin, X., Huang, J. X. and Li, Z., 2011, "Mining and modeling linkage information from citation context for improving biomedical literature retrieval," *Information Processing and Management*, Vol. 47, No. 1, pp.53-67.
51. Zajic, D., Dorr, B. J., Lin, J. and Schwartz, R., 2007, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks," *Information Processing and Management*. Vol. 43, No. 6, pp. 1549-1570.
52. Zhou, L., Tao, Y., Cimino, J. J., Chen, E. S., Liu, H., Lussier, Y. A., Hripcsak, G. and Friedman, C., 2006, "Terminology model discovery using natural language processing and visualization techniques," *Journal of Biomedical Informatics*, Vol. 39, No. 6, pp. 626-636.

出席國際學術會議心得報告

103 年　06 月　01 日

計畫編號	NSC 102-2221-E-343 -003
計畫名稱	整合知識問答解析與文件結構化摘要技術之知識文件檢索模式(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	2014/3/28~2014/3/30, Japan (Tokyo)
會議名稱	The 2nd Annual Conference on Engineering & Information Technology (ACEAIT 2014)
發表論文題目	A Semantic Analysis Technology between Domain Vocabulary and Colloquial Query String for Knowledge Document

一、參加會議經過與心得

此次 The 2nd Annual Conference on Engineering & Information Technology (ACEAIT 2014) 研討會乃安排於日本（Japan）之東京（Tokyo）舉辦，配合研討會主辦單位之行程規劃與可行機位安排，個人於 03/27 上午七點，即出發前往台北松山機場，進行登記作業且於上午 09：00 由台北松山機場起飛，並於中午 12：40 抵達東京羽田 Tokyo Haneda Airport 國際機場，辦理入境手續後即搭乘「東京單軌電車」至濱松町駅，再轉搭「JR 山手線」抵達新橋駅，最後搭乘「東京 METRO 銀座線」至赤坂見附駅，並於當地下午 14：00 左右抵達住宿飯店「赤坂東急大飯店」(Akasaka Excel Hotel)(此飯店離研討會舉辦飯店 Toshi Center Hotel 僅需 15 分鐘步行路程)，台北與東京時差為 1 小時；此程搭機、轉乘地鐵時間總計 5 小時左右，故辦理 Check in 手續並稍做休息後，於本日及 03/28 上、下午即參觀東京中城、六本木 Hill、晴空塔、淺草雷門、台場等周遭景點，以瞭解日本東京都之交通、飲食習慣、友善的風俗民情及東京都之建築、古蹟等風貌。

由於個人住宿於東京之 Akasaka Excel Hotel，離研討會會場 Toshi Center Hotel 僅需 15 分鐘步行路程，個人於 03/28 下午 16：30 隨即前往報到，完成報到手續（如圖 1 及圖 2 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中型規模，研討會總計發表篇數約為 450 篇左右，議程數為 66 個左右，當中 Oral 議程約個 50 場次、Poster 約 17 個場次左右。大會正式行程日期為 3/28 至 3/30 三日，3/28 乃為提前報到日期，正式發表日期亦為 3/29 至 3/30 兩日。本次研討會內容乃安排與此次會議主題相關之工程與資訊科技（Engineering and Information Technology）專題演講與論文發表，再依不同論文主題每天分至 5 個時段 6 個左右平行 Session 進行發表。個人的論文被安排於發表日 3/29 的上午（10:30~11:30）場次（編號 Poster Sessions 707）「Information

Engineering and Technology」發表，由於此研討會主題乃著重於工程與資訊科技，個人研究（資料探勘、知識管理、系統開發）與其他學者甚為相似，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 ACEAIT 2014 會場並註冊(1)

圖 2、抵達 ACEAIT 2014 會場並註冊(2)

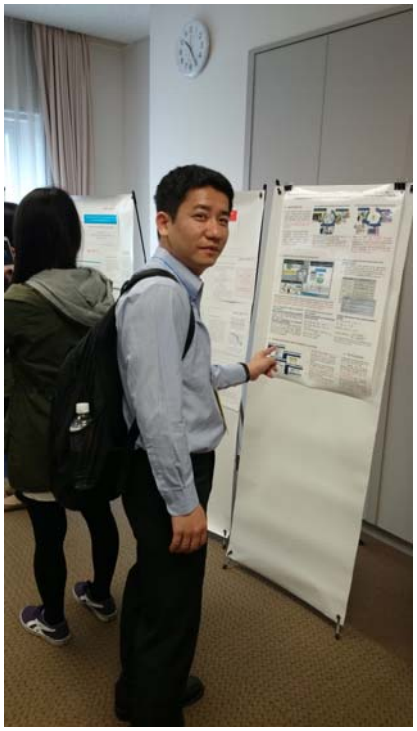

圖 3、論文發表與研討(1)

圖 4、論文發表與研討(2)

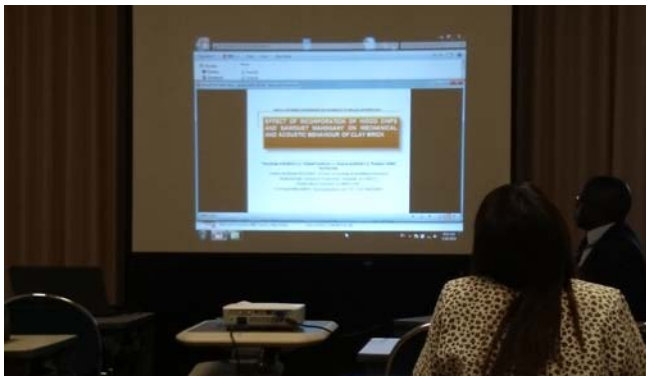

圖 5、論文發表與研討(3)

圖 6、論文發表與研討(4)

圖 7、論文發表與研討(5)

圖 8、論文發表與研討(6)

待研討會圓滿結束後，個人於隔日（3/31）即搭車前往東京羽田 Tokyo Haneda Airport 國際機場，並搭機回台北松山機場，結束此次 ACEAIT 2014 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中型之規模，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 Toshi Center Hotel 之 7 樓，故需受限於飯店之場地限制（如各議程場地較為狹小及休息區之規劃皆不甚完美），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 本及論文摘要集 1 本。
2. 國內外學者學術交流名片。

ACEAIT-2806
A Semantic Analysis Technology between Domain Vocabulary and Colloquial Query String for Knowledge Document

Shih-Ting Yang* and Yu-Ting Gong
Department of Information Management, Nanhua University, No. 55, Sec. 1, Nanhua Rd., Dalin Township, Chiayi County 62249, Taiwan

*Corresponding Author: stingyang@mail.nhu.edu.tw

Abstract

Although websites at present allow users to obtain information by topic or keywords, the search may not be successfully if users lack domain knowledge in the specified field. For example, in the domain website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw), the most demanders cannot obtain the needed knowledge documents using the colloquial query string without domain keywords respect to knowledge documents and thus reducing the knowledge sharing efficient of this domain website. To solve above problem, 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 domain vocabulary and colloquial query string can be obtained. It is expected that knowledge demanders can directly read the desired parts according to problems to ensure they can find document they want within a short time. In order to demonstrate applicability of the proposed methodology, a web-based knowledge document retrieval system is also established based on the proposed model. Furthermore, the knowledge documents (i.e., ergonomic technology reports) from the website of "Institute of Occupational Safety and Health" are applied as examples to demonstrate the proposed model and system. As a whole, this research provides an approach for knowledge demanders to efficiently and accurately acquire the domain knowledge documents.

Keyword: Institute of Occupational Safety and Health, Knowledge Management, Data Mining, Semantic Analysis

1. Introduction

As the Internet becomes a popular source of information acquisition, many papers 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 domain fields; these websites are representative in the specific domain fields. In other words, when users are searching for information, they conduct searches on the website intuitively. Although websites at present allow users to obtain information by topic or keywords, the search may not be successfully if the users lack domain knowledge in the specified field. For example, in the domain website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw), the most demander cannot obtain the needed knowledge documents using the colloquial search phrases without domain keywords respect to knowledge documents and thus reducing the knowledge sharing efficient of this domain website.

To solve above problem, this paper proposes a semantic analysis technology between domain vocabulary and colloquial query string for knowledge document to help users to rapidly and efficiently obtain the documents needed. Based on the domain website of IOSH, this paper analyzes the knowledge document expressions and related domain vocabulary regarding the research reports or technical books on the knowledge websites, and establishes the knowledge vocabulary library. In this way, semantic association with the search phrase can thus be established to enhance the keyword semantic search technology. The proposed technology can strengthen the search phrase in semantic determination by using the representative vocabulary of the document and thus enhancing the knowledge sharing effectiveness of the website of IOSH. To sum up, this paper develops a semantic analysis technology between domain vocabulary and colloquial search phrase for domain knowledge document.

2. Literature Review

Regarding the topic of Q&A (Question and Answer) application and technology, in this study, this study conducted literature review relating to Q&A application types and Q&A 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 mode. According to the document association and common features' BE (Basic Element), Teng et al. (2010) established a user-oriented document abstract retrieval system.

In terms of Q&A technology, this study categorized various considerations. The factors for consideration may be based on subject, the document characteristics or the semantics for analysis. In the case of using subject as the main reference, Oh et al. (2012) proposed a Q&A system 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 technology. Using semantics as the main reference, 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.

3. A Semantic Analysis Technology between Domain Vocabulary and Colloquial Query String for Knowledge Document

The proposed Semantic Analysis Technology between Domain Vocabulary and Colloquial Query String for Knowledge Document used the technical books and research reports on the website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw) as the basis for analysis. The corresponding knowledge document can be found to enhance retrieval accuracy. Therefore, the main research procedures can be divided into the following parts including Part1 "Knowledge Document Expression Item Analysis Module", Part2 "Conceptual Sentence Acquisition (CSA) Module", and Part3 "Question and Answer Analysis (QAA) module".

3.1 Knowledge Document Expression Item Analysis Module

This paper consulted ergonomics staffs and summarized 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 as below.

3.1.1 Establishment of the expression items of the knowledge documents

The analysis of the content of the ergonomics workplace research reports can be divided into 8 expression items, and 19 detailed expression items. To strengthen the sentence smoothness, nine detailed expression items are added for sentence assistance. Hence, knowledge vocabulary database consists of 28 vocabulary sets can be shown in Table 1 and Table 2.

3.2 Conceptual Sentence Acquisition (CSA) Module

Since the ergonomic technology reports (target document) are written by experts in the domain field, the expression methods are not consistent with each other. The CSA module acquires the complete sentence SD_i by segmenting the target document D_T, and conducts vocabulary comparison rules on the basis of domain vocabulary set created by domain experts. Then, this module compares the complete sentence SD_i and vocabulary comparison rules to extract the conceptual sentences and attribute them to corresponding sets.

Step (A1): Target Document Sentence Acquisition

This step first builds the punctuation marks set (for example:. !,;, etc) to obtain the sentences of the target document D_T.

(A1.1): Subsection of Target Document:

According to the table of punctuation symbols (for example:. !,;), sub-sections of the target document are worked out. After this step, the complete sentences of the target document D_T including SD_1, SD_2, SD_3, …, SD_i, $SD_{N(DT)}$ can be obtained.

(A1.2): Word Dismantling of the Complete Sentences:

After getting the complete sentence SD_i, the word series are dismantled into word groups ranging from 2 to 6 words to form the vocabulary set. $SD_{i,j}$ represents the j‘th word of the i‘th sentence after dismantling, consisting of a number of words as shown in Equation (1).

$$SD_i = \{ SD_{i,1}, SD_{i,2}, SD_{i,3} \cdots, SD_{i,j}, \cdots \} \quad (1)$$

Table 1: Detailed expression items and corresponding contents and means of expression

Expression Items	Detailed Expression Items	Descriptions of Expression
Operation Field	Operation field	Set of industrial category with contents including "agriculture, forestry, fishery and animal husbandry", "mining and quarrying", "food manufacturing", and "textiles and clothing industry".
Operation Name	Operation name	Set of name of the operation in this industry with contents including "mode replacement operation", "packaging operation", and "transportation operation".
Operation Identity	Operator gender	Set of vocabulary describing the gender of the operations with contents including "male" "female" or "male or female".
	Operator age	Set of vocabulary describing the operator age, for example "20~30 years old"; with contents including "the middle aged", "the youth", and "no age limit".
	Operator title	Set of vocabulary describing the operator title with contents including "nurse", "technician", and "operator".
Operation Environment	Equipment vocabulary	Set of the equipment vocabulary for the operation such as "blood bed", "thermotank", and "thermal forging machine".
	Facilities layout	Set of the vocabulary of facilities layout and placement such as "the height of the transmission belt is 75 cm".
	Tool introduction	Set of the vocabulary describing tools used in operation or improvement process with contents including "butterfly cage", "arm support rest", "lift vehicle", and "cart".
Operation Behavior	Operation goal	Set of the vocabulary describing the name of the operation corresponding to the operation target with contents including "major job", "main function" and "main points".
	Operation description	Set of vocabularies corresponding to the operation name, operation goal and operation tool.
	Professional verbs	Set of vocabulary describing the corresponding operations of the operator with contents including "standing posture", "bending", "force application" and "lifting".
Operation Hour	Operation hours/day	To describe the times of repetitive actions of the operator including "daily requirements", and "daily necessity".
	Operation hour/times	Set of vocabulary describing the time for the job of the operator with contents including "time for one times", "one times requirement", and "one times necessity".
	Operation distance/times	Set of vocabulary describing the operation distance of the operator with contents including "distance", and "shortest distance".
Causes of Injury	Factors of injury	Set of vocabulary describing causes of injury with contents including "excessive force", "highly repetitive actions", "vibration", "low temperature", and "poor working posture".
	Pain parts	Set of vocabulary describing the posture and pain parts of the operator with contents including "neck", "torso", "hand", "wrist", and "leg".
Improvement Method	Improvement goal	Set of vocabulary describing the improvement goal of the causes of injury with contents including "main improvement", "effective improvement", "considerably", and "significant reduce".
	Improvement procedure	Set of vocabulary describing procedural improvement with contents including "consider", "use", "suggestion", and "as long as".
	Improvement review	Set of vocabulary describing the review after operation assessment with contents including "action level", "grading points", "total inspection score", "risk level", and "significant reduction of load".

Table 2: Auxiliary detailed expression items and corresponding contents

Detailed Expression Items	Descriptions of Expression
Linking vocabulary	Including “and”, “but”, “as well as”.
General verbs	Set of the vocabulary of general verbs including “is”, “mainly is”, “as”, “raise”, “put down”, “move”, and “store”.
Numerical vocabulary	Vocabulary recording numbers from “0” to “9” and their combinations
Monetary unit vocabulary	Including “RMB”, “NTD”, etc.
Age unit vocabulary	Including “years old”.
Length unit vocabulary	including “distance”, “cm”, “meter”, “length”, etc.
Time unit vocabulary	including “time”, “minute”, “hour”, etc.
Wight unit vocabulary	Including “kg”, “g”, “ton”, etc.
Frequency vocabulary	including “times”, etc.

Step (A2): Establishment of Structured Vocabulary Comparison Rules

After the formation of the complete sentences SD_1, SD_2, SD_3, …, SD_i, …, $SD_{N(DT)}$, the conceptual sentences can be judged. This paper establishes eight selection rules regarding the vocabulary comparison rules to obtain the representative sentences of the vocabularies.

1. Operation Field Vocabulary Rule (R_OF): This rule is to express the industrial classification, and the rule is as shown in Equation (2). If the complete sentence S is in the operation field conceptual vocabulary, then the complete sentence SD_i is the operation field conceptual sentence OF_Set.

$$\text{IF } SD_{i,j} \text{ exist in } OF(CS)\ \forall j \text{ Then } SD_i \in OF\_Set \quad (2)$$

2. Operation Name Vocabulary Rule (R_ON): This rule is to express the name of the action, and hence the rule is as shown in Equation (3). If the complete sentence is in the operation name conceptual vocabulary, then the complete sentence SD_i is the operation name conceptual sentence ON_Set.

$$\text{IF } SD_{i,j} \text{ exist in } ON(CS)\ \forall j \text{ Then } SD_i \in ON\_Set \quad (3)$$

3. Operator Identity Rule (R_OR): This rule's expressions include the operator's gender vocabulary, age vocabulary, and title vocabulary. To strengthen the accuracy of the judgment rules in this paper, at this step, a strict rule to ensure accurate acquisition is built. The method is to acquire the set of words of the complete sentence by Equation (1) and select the vocabulary by rules. The loose and strict rules are defined as follows:

✓ The loose rule: This rule is to express the concept of operation by one to two words, for example, as shown in Equation (4), using operator title ORT (CS) to represent the operator identity, or using the operator title ORT(CS) coupled with operator age ORA(CS) to represent the age of the operator, using the combination of the operator age, operator title ORT(CS) and operator gender ORS(CS) to express the gender of the operator (Equations (5) and (6)).

$$\text{IF } SD_{i,j} \text{ exist in ORT(CS) } \forall j \text{ Then } SD_i \in OR\_Set \tag{4}$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{ORT(CS)} \\ \text{and ORA(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in OR\_Set \tag{5}$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{ORT(CS)} \\ \text{and ORS(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in OR\_Set \tag{6}$$

✓ The strict rule: This rule uses a couple of words to form the strict structure for the expression of the concept relating to the operator identity, uses the numerical vocabulary N(CS) and age unit vocabulary AU(CS) to expressly represent the operator's age range (Equation (7)).

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{ORT(CS) and N(CS)} \\ \text{and AU(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in OR\_Set \tag{7}$$

4. Operation environment vocabulary rule (R_OE): This rule is to express the facilities and tools of the operation environment with descriptions including the descriptions of length, width, height and other specifications. As shown in Equation (8), the description of the operation environment is realized by facility vocabulary F(CS), facility layout vocabulary FL(CS), numerical vocabulary N(CS) and length unit vocabulary LU(CS) for definite expression of the operation facility's specifications. The description of the operation tools is as shown in Equation (9), the operation tool vocabulary OT(CS) is combined with the numerical vocabulary N(CS) and length unit vocabulary LU(CS) to definitely express the specifications of the operation tools.

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{F(CS) and FL(CS)} \\ \text{and N(CS) and LU(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in OE\_Set \tag{8}$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{OT(CS) and N(CS)} \\ \text{and LU(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in OE\_Set \tag{9}$$

5. Operation behavior vocabulary rule (R_OV): This rule is to express the description of the operation goals. The expressions include the including operation goal vocabulary,

operation tool vocabulary and the domain verbs to express the operations and postures. According to Equation (10), the description of operation goal should be integrated with the operation goal OG (CS) and the general verb vocabulary GV (CS); or as shown in Equation (11), the operation goal vocabulary OG(CS) can be integrated with the general verb vocabulary GV(CS) and domain verb vocabulary PV(CS) to more strictly express the concepts. The expression for the operation definition vocabulary is as shown in Equation (12), the operation definition rule is to combine the operation name vocabulary ON(CS) with the general verb vocabulary GV(CS), domain verb vocabulary PV(CS) and operation tool vocabulary OT(CS).

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} OG(CS) \\ \text{and } GV(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OV\_Set \quad (10)$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} OG(CS) \text{ and } GV(CS) \\ \text{and } PV(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OV\_Set \quad (11)$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} ON(CS) \text{ and } GV(CS) \\ \text{and } PV(CS) \text{ and } OT(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OV\_Set \quad (12)$$

6. Operation hour vocabulary rule (R_OH): This rule is to represent the operation frequency and operation time. The descriptions include operation frequency (operation times/day) vocabulary OFQ (CS), operation hour (operation hour/times) vocabulary OH(CS), operation distance (operation distance/times) vocabulary ODT(CS). By the selection of Equations (13), (14), and (15), the sentences are listed in line with the standards as the set of the operation time vocabulary conceptual sentences OH_Set.

$$\text{IF } SD_{i,j} \text{ exist in} \begin{pmatrix} OFQ(CS) \text{ and } PV(CS) \\ \text{and } N(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OH\_Set \quad (13)$$

$$\text{IF } SD_{i,j} \text{ exist in} \begin{pmatrix} OH(CS) \text{ and } N(CS) \\ \text{and } FU(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OH\_Set \quad (14)$$

$$\text{IF } SD_{i,j} \text{ exist in} \begin{pmatrix} OT(CS) \text{ and } ODT(CS) \\ \text{and } N(CS) \text{ and } LU(CS) \end{pmatrix} \forall j \text{ Then } SD_i \in OH\_Set \quad (15)$$

7. Injury cause vocabulary rule (R_IC): This rule is to express the injuries caused by the operations. The expressions include injury cause vocabulary and body part vocabulary. As shown in Equation (16), expressions of injury cause can be realized by integrating the injury cause vocabulary IC(CS) with the operation body part vocabulary B(CS).

$$\text{IF } SD_{i,j} \text{ exist in} \left(IC(CS) \text{ and } B(CS)\right) \forall j \text{ Then } SD_i \in IC\_Set \quad (16)$$

8. Improvement method vocabulary rule (R_IM): This rule includes improvement purpose, improvement process, and improvement review. As shown in Equation (17), the expression and description of the improvement purpose should be combined the improvement purpose vocabulary IG(CS) and the general verb vocabulary GV(CS) and domain verb vocabulary PV(CS). The expression forms of the improvement process

vocabulary are as shown in Equation (18). The description of the improvement process is expressed by the combination of the improvement process vocabulary IR(CS), the general verb vocabulary GV(CS) and operation tool vocabulary OT(CS). Regarding the expression of the review vocabulary is as shown in Equation (19) by improvement review vocabulary R(CS) directly or as shown in Equation (20) by the combination of the review verb vocabulary RV(CS), operation title vocabulary ORT(CS) and domain verb vocabulary PV(CS) in a strict way.

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{IG(CS) and GV(CS)} \\ \text{and PV(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in IM\_Set \quad (17)$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{IR(CS) and GV(CS)} \\ \text{and OT(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in IM\_Set \quad (18)$$

$$\text{IF } SD_{i,j} \text{ exist in R(CS) } \forall j \text{ Then } SD_i \in IM\_Set \quad (19)$$

$$\text{IF } SD_{i,j} \text{ exist in } \begin{pmatrix} \text{RV(CS) and ORT(CS)} \\ \text{and PV(CS)} \end{pmatrix} \forall j \text{ Then } SD_i \in IM\_Set \quad (20)$$

Finally, this module can obtain the sets of eight conceptual sentences including operation field, operation name, operation title, operation environment, operation, operation time, injury cause and improvement method. At the stage of the conceptual sentence acquisition module, the free-form documents are converted into structured expressions containing conceptual sentences for the **question and answer analysis**.

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.

Step (B1): Determination of the implied goals of the question words

In this step, Q&A can be used to obtain the implied goals and relevant words of the question words to for the training of the similarity probability of question words and relevant semantic words.

As shown in Equation (21), when the users inputs the question sentence (QW_i), after the segmentation of the sentence, the meaningful question word ($QW_{i,j}$) can be captured. In accordance with the bilingual vocabulary developed by the Institute of Occupational Health and Safety, the relevant semantic words can be obtained. According to the domain term "bruise", the relevant semantic words "skin" or "membrane" can be captured from the notes "the superficial tissues (e.g., skin or membrane) covering the body has been scratched or torn off". By using the semantic words as the question words, the answer sentence (AW_k) composing of multiple answer words ($AW_{k,m}$) can be obtained. Equation (22) computes the probability of the words in the question sentence ($QW_{i,j}$) and answer words ($AW_{k,m}$) to obtain relevant vocabulary to establish the set of question words and the corresponding sets of answer words to various question words. Finally, as shown in Equation (23), by using the semantic threshold value of the question word against answer words ω(AW,QW), the question words associated with the answer sentence can be filtered and selected to form the set.

$$\begin{aligned} QW_i &= \left\{ QW_{i,1}, QW_{i,2}, QW_{i,3}, \cdots, QW_{i,j}, \cdots \right\} \\ AW_k &= \left\{ AW_{k,1}, AW_{k,2}, AW_{k,3}, \cdots, AW_{k,m}, \cdots \right\} \end{aligned} \tag{21}$$

$$\begin{aligned} &P(AW_{k,m} \mid QW_{i,j}) \\ &= \frac{C(AW_{k,1} \mid QW_{i,1}) + C(AW_{k,1} \mid QW_{i,2}) \cdots + C(AW_{k,1} \mid QW_{i,j})}{C(QW_{i,j})} \bullet C(AW_{k,1}) \end{aligned} \tag{22}$$

$$IF\ P(AW_{k,m}, QW_{i,j}) > \varpi(AW, QW)\ Then\ AW_{k,m}, QW_{i,j} \in QWAW^{\varpi}{}_k \tag{23}$$

Step (B2): Determination of the vocabulary category similarity

Through the expression items of the target document analysis, this paper establishes the document keyword set ($D_{i,q}$). Moreover, as shown in Equation (24), the expression vocabulary of each expression item is defined as a domain vocabulary. In this step, the VSM (Vector Space Model) Cosine is used to compute the similarity of the document and the set of the answer words and the level of similarity of the answer word set and the document. $Sim(D_q|QWAW_k)$, can be determined by Equation (25). If the similarity level is above the threshold value ω and is closer to 1, it means the set has more parsing meanings of the document.

$$D_i = \left\{ D_{i,1}, D_{i,2}, D_{i,3}, \cdots, D_{i,q}, \cdots \right\} \tag{24}$$

$$D_q^{\varpi} = [w_1, w_2, \cdots, w_q]^T$$
$$QWAW^{\varpi}{}_k = [w_1, w_2, \cdots, w_k]^T \quad (25)$$
$$Sim(D_q \mid QWAW_k) = \frac{D^{\varpi}{}_q \cdot QWAW_k{}^{\varpi}}{|D^{\varpi}{}_q| \cdot |QWAW_k{}^{\varpi}|}$$

In addition, four ways of setting the threshold values are proposed for users in the selection of documents. The threshold value can be set as the average, the median, the quartile or the direct definition. If it is above the threshold value $\omega(D_q,QWAW_k)$, it means that the domain vocabulary is connected with the document and it is placed in the reserve document set ($ReserveDoc\_Set_d$).

4. Knowledge Document Retrieval System

In this web-based system, common users can upload the knowledge document; then, the system administrator can set the system parameters and add new question word and answer words, and thus implement the analysis of the expression items and Q&A analysis of the knowledge document.

In order to verify the feasibility of the knowledge document retrieval system in the practical application, this study used ergonomic technology reports (i.e., knowledge documents) from Institute of Occupational Safety and Health website for verification and applies the kernel modules of the system (including "Q&A analysis") to demonstrate feasibility of the proposed methodology and the developed system. The common users may realize the function of "knowledge document uploading" through "knowledge document management module".

After the knowledge document uploading, the system administrator can perform "knowledge document expression item parsing module" to capture the conceptual sentences of various expression items of the knowledge document (as shown in Figure 1 and Figure 2). 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 system can obtain the keywords of the document such as "storage box" and "handling", based on which the system can analyze the question goal and the relevant answer word combinations (as shown in Figure 3). According to the analysis of the question word, answer sentences and answer words matching, the system can obtain the relevant answer words of the question word "pain" such as "construction industry" and "age" (as shown in Figure 4).

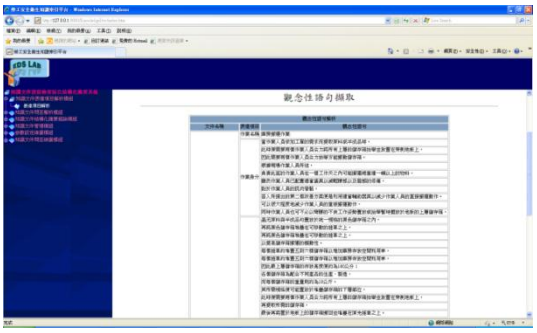

Figure 1: Document expression items analysis result(1)

Figure 2: Document expression items analysis result(1)

Figure 3: Q&A analysis result (1)

Figure 4: Q&A analysis result (2)

5. Conclusions

As the specific knowledge fields are too professional, common users can hardly know and define the key words. As a result, the document search will take more time and have more obstacles. Therefore, it can easily affect the selection of documents due to personal browsing and reading preferences, and thus reducing the knowledge sharing effectiveness of the knowledge websites. Hence, this study proposes a semantic analysis technology between domain vocabulary and colloquial query string for knowledge document and establishes a web-based system to confirm the feasibility of the methodology and model. As the verification results have suggested, the system can process knowledge document semantic Q&A and realize the semantic association of general vocabulary and professional vocabulary through Q&A semantic analysis. Hence, common users can search domain knowledge documents by colloquial query string and get the relevant knowledge documents to enhance the reading and selection of users. In this way, it can help users to access to the information on the website of IOSH, and thus enhancing the domain knowledge document search effectiveness.

References

[1] Cao, Y. G., Liu, F., Simpson, P., Antieau, L., Bennett, A. Cimino, J. J., Ely, J. and Yu, H. 2011. AskHERMES: An Online Question Answering System for Complex Clinical Questions, *Journal of Biomedical Informatics*, 44(2), 277-288.

[2] Dorr, B. J. and Gaasterland, T. 2007. Exploiting Aspectual Features and Connecting Words for Summarization-inspired Temporal-relation Extraction, *Information Processing and Management*, 43(6), 1681-1704.

[3] 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.

[4] Han, K. S., Song, Y. I., Kim, S. B. and 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.

[5] Huang, M., Zhu, X. and Li, M., 2006, "A Hybrid Method for Relation Extraction from Biomedical Literature, *International Journal of Medical Informatics*, 75(6), 443-455.

[6] Jones, M. and Love, B. C. 2007. Beyond Common Features: The Role of Roles in Determining Similarity, *Cognitive Psychology*, 55(3), 196-231.

[7] Ko, Y., Park, J. and Seo, J. 2004. Improving Text Categorization Using the Importance of Sentences, *Information Processing and Management*, 44(1), 65-79.

[8] Oh, H. J., Myaeng, S. H. and Jang, M. G. 2012. Effects of Answer Weight Boosting in Strategy-driven Question Answering, *Information Processing and Management*, 48(1), 83-93.

[9] Oh, H. J., Sung, K. Y., Jang, M. G. and Myaeng, S. H. 2011. Compositional Question Answering: A Divide and Conquer Approach, *Information Processing and Management*, 47(6), 808-824.

[10] Teng, C., Xiong, N., He, Y., Yang, L. T. and Liu, D. 2010. A Behavioural Mode Research on User-focus Summarization, *Mathematical and Computer Modelling*, 51(7-8), 985-994.

Acknowledgment

This research is partially supported by the National Science Council under project No. NSC 102-2221-E-343 -003

科技部補助計畫衍生研發成果推廣資料表

日期:2014/10/28

科技部補助計畫	計畫名稱: 整合知識問答解析與文件結構化摘要技術之知識文件檢索模式(I)
	計畫主持人: 楊士霆
	計畫編號: 102-2221-E-343-003- 學門領域: 資訊系統

無研發成果推廣資料

102 年度專題研究計畫研究成果彙整表

計畫主持人：楊士霆	計畫編號：102-2221-E-343-003-
計畫名稱：整合知識問答解析與文件結構化摘要技術之知識文件檢索模式(I)	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	2	2	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	4	4	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	1	2	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

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。)	此計畫之部分議題「A semantic analysis technology between domain vocabulary and colloquial query string for knowledge document」乃發表於「The 2nd Annual Conference on Engineering & Information Technology, Tokyo, Japan, March 28-30.」研討會中，並獲「ACEAIT 2014- Best Manuscript」之殊榮。

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

科技部補助專題研究計畫成果報告自評表

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：

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 100 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)本計畫乃以特定知識領域之知識文件為基礎，發展一套「整合知識問答解析與文件結構化摘要技術之知識文件檢索」模式，以解析知識文件表達項目並進行專業語彙語意關聯以及結構化摘要呈現。此方法論之相關重點成效乃歸納如下：

(A.2)本計畫乃解析並取得知識文件之觀念性語句以及具有代表性之專業語彙，並將一般詞彙與專業語彙進行語意關聯，知識搜尋者即可以一般詞彙進行搜尋，並取得相關知識文件。

(A.3)本計畫提出之結構化摘要即可將自由形式知識文件轉為制式化之文件摘要，知識搜尋者能以制式呈現方式更容易審視知識文件，同時亦可避免個人閱讀偏好影響文件之篩選。

(B)技術開發層面

(B.1)本計畫乃以「結合語意解析與文件摘要技術之知識文件檢索」方法論為依據，並以 JSP（Java Server Pages）以及 SQL Server 2008 等系統開發工具建置「整合知識問答解析與文件結構化摘要技術之知識文件檢索」系統，其系

統具體成效乃歸納如下：
(B.2)本系統所開發之「整合知識問答解析與文件結構化摘要技術之知識文件檢索」乃建置於網際網路中，並可執行「知識文件表達項目解析」、「知識文件問答解析」、「知識文件結構化摘要推論」以取得知識文件觀念性語句、關鍵字詞與相關語意詞，以及知識文件簡略結構化摘要與詳述結構化摘要呈現。
(B.3)由驗證結果得知，本系統之各項驗證指標（即「知識文件問答解析」之回覆詞召回率、回覆詞正確率與「知識文件結構化摘要推論」使用者滿意度與文件字數呈現）皆具學習能力；當使用至一定訓練資料數量系統即可有效解析知識文件之觀念性語句、專業語彙與一般詞彙之語意關聯與建立結構化摘要。

(C)實務應用層面
(C.1)本計畫所建置之「整合知識問答解析與文件結構化摘要技術之知識文件檢索」系統可有效分析知識文件，故此系統可應用於線上特定領域網站以解決知識文件搜尋等任務。其具體貢獻與成效乃歸納如下：
(C.2)本系統可應用於勞工安全衛生知識網等相關領域網站，以有效表達知識文件之描述、並協助知識搜尋者尋求問題解答等搜尋任務。
(C.3)本系統可應用於勞工安全衛生知識網等相關領域網站，以結構化摘要方式呈現知識文件，以協助知識搜尋者省力且快速方式獲得資訊，進而提高知識領域搜尋效能。