

結合語意解析與文件摘要技術之知識文件檢索模式

楊士霆

南華大學資訊管理學系
stingyang@mail.nhu.edu.tw

龔鈺婷

南華大學資訊管理學系
stingyang@mail.nhu.edu.tw

摘要

雖有關鍵字搜尋、文件分類等搜尋窗口以縮小搜尋範圍，但於特定領域一般搜尋者仍需不斷嘗試並取得回饋。因此，本研究針對特定領域知識文件發展一套「結合語意解析與文件摘要技術之知識文件檢索」模式，先行解析知識文件表達項目並擷取觀念性語句以建構知識語彙庫，透過「問答解析模組」進行知識文件之專有名詞與搜尋字詞語意關聯以取得文件，最後透過「結構化摘要模組」推論並擷取具有文件之代表性語句，以整合成制定摘要呈現方式，即避免個人閱讀偏好影響文件之篩選，進而加強勞工安全知識網搜尋技術以提高勞工安全知識網知識分享之成效。

關鍵詞：勞工安全知識網、知識管理、資料探勘、文件摘要技術、語意分析

第一章、研究動機與目的

為協助知識搜尋者方便取得資訊，大多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，即以直覺方式於該網站進行查詢。雖有關鍵字搜尋、文件分類等搜尋窗口以縮小搜尋範圍，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，文件摘要乃自由形式呈現且無字數之控管，導致搜尋者容易因個人瀏覽、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。綜合上述，其既有之運作模式如圖 1 所示。



圖 1 知識搜尋者取得資訊之既有模式

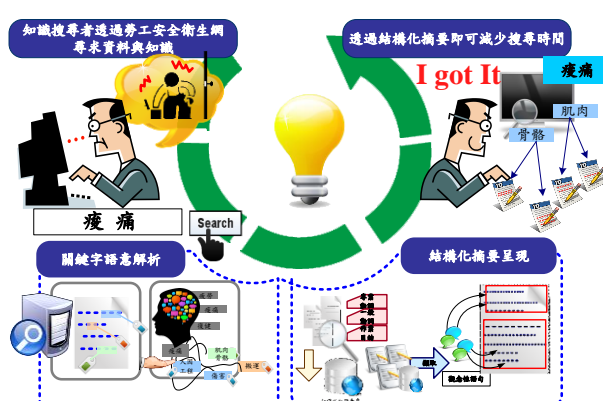


圖 2 知識搜尋者取得資訊之期望模式

如圖 2 所示，為掌握搜尋者需求並針對問題提供確實回饋，本研究所提出之「結合語意解析與文件摘要技術之知識文件檢索模式」乃針對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。本研究乃以勞工安全知識網為基，針對知識網之研究報告或技術叢書解析知識文件表達方式與相關性語彙，並建立知識語彙庫。於此即可與搜尋字串進行語意關聯，以加強關鍵字語意搜尋技術。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，以協助知識搜尋者加速判定文件，同時避免個人閱讀偏好影響文件之篩選，進而提高「勞工安全知識網」知識分享之效能。

第二章、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」與「摘要應用與技術分析」等兩大研究方向，以下即針對此兩大主題之相關研究進行文獻回顧及探討。

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。針對問答應用類型大多可分為問答系統與檢索系統，於問答系統中 Oh et al (2011) 提出提出一套複合式問答系統 (Compositional QA)，該研究先以準則判斷方式進行問題分析，針對問題格式 (單一或多重問項)、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。Cao et al (2010) 則針對醫學上臨床報告建立一套線上問答系統 (AskHERMES)，以擷取複雜、無格式化之臨床報告重點，並將擷取重點總結後形成答案。於資訊檢索中，Huang et al (2006) 提出一套混合式關係模式以擷取生物醫學文獻，透過淺層句法解析 (Shallow Parsing) 取得語法、語意之結構，再以貪婪法進行匹配以訓練模式擷取生物醫學文獻之主題。Teng et al (2010) 根據文件之關聯性與共同特徵，建立一套以使用者為導向之文件摘要檢索。對於問答技術部分，針對各種考量進行分類。以主題為主部分，Oh (2012) 等人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，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。**Han et al (2007)** 提出以先確定問題類型並建立各類型相關詞彙，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。以文件特徵為主部分，**Jones 與 Love (2007)** 提出於文件中若關係越具相似性則表示兩者之間存在一個共同角色，透過背景環境、關係為匹配準則以取得文件之共同關係。**Ko et al (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。於語意為主部分，**Dorr 與 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。**Dunlavy (2010)** 等人提出一套整合資訊查詢系統，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進行相關性查詢分析。

對於摘要應用與技術議題而言，本研究乃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。針對摘要應用領域而言，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 et al (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。**Zajic et al (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。**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。針對文件類型分析可分為「制式化文件」與「非制式化文件」。**Ko (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。**Moens (2007)** 針對具有高度結構化之法院判決紀錄文書進行電子化，透過判決紀錄文書以解析文件之概念序列並將判決文件劃分為：(1)基本要素（包含案由、投訴類型等）、(2)背景、(3)相關法律問題等。針對摘要建立技術部分可分為「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。於監督式學習部分，**Bollegala et al (2010)** 以支援向量機，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。**Li 與 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。**Jung et al (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，藉由雙向連詞結合相鄰語句以建立假設語句，並以標題與位置等準則計算文件之重要性，最後進行群聚分析取得相似性並形成摘要。於非監督式學習部分，**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。**Steinberger et al (2007)** 以潛在語意分析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。**Ko et al (2003)** 利用詞彙群聚提出一套以主題為基之文件摘要技術，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。

第三章、結合語意解析與文件摘要技術之知識文件檢索模式

本研究所提出之「結合語意解析與文件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、二十九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。

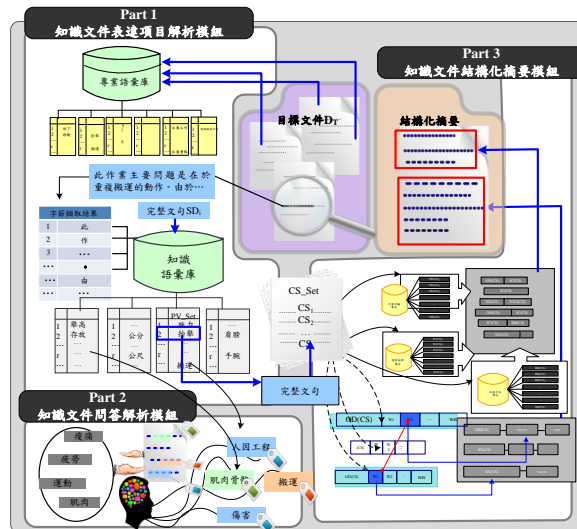


圖3 結合語意解析與文件摘要技術之知識文件檢索模式之流程架構

3.1 知識文件表達項目解析模組

本研究針對勞工安全知識網之技術叢書與研究報告作為知識文件基礎，並進行文件分析。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，本研究乃解析知識文件進行表達項目分析，為後續問答解析與觀念性語句擷取之用。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，本研究另增十個表達細項進行語句輔助，其二十九個語彙集合如表1與表2所示，並定義符號如下：

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」等。 |
| | 作業員年齡 | 形容該作業之作業員年齡範圍之語彙；其內容包含「中年」、「青年」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」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」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 | 為作業目標語彙之集合；其內容包含「主要工作」、「主要功能」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。 |
| | 施力程度 | 為此作業所需承受重量語彙之集合；其內容包含「重量約」、「重達」等。 |
| 作業時間 | 作業次數/天 | 為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 | 為此作業所花費時間集合；其內容包含「一次所花時間」、「一次需要」等。 |
| | 作業距離/次 | 為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 | 表示作業描述語彙之姿勢部位及傷痛部位之集合；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；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；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；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」等。 |
| 專業動詞 | 為作業描述語彙之姿勢集合；其內容包含「站姿」、「彎腰」、「低頭」、「施力」等。 |
| 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，並為後續問答解析與結構化摘要之應用。如公式(1)所示，首先以字節擷取方式取得完整語句 SD_i 。

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

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 。

$$IFSD_{i,j} \text{ exist in } OF(CS) \forall j \text{ Then } SD_i \in OF_Set \quad (2)$$

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 。

$$IFSD_{i,j} \text{ exist in } ON(CS) \forall j \text{ Then } SD_i \in ON_Set \quad (3)$$

✓ 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。

$$IFSD_{i,j} \text{ exist in } ORT(CS) \forall j \text{ Then } SD_i \in OR_Set \quad (4)$$

$$IFSD_{i,j} \text{ exist in } (ORT(CS) \text{ and } ORA(CS)) \forall j \text{ Then } SD_i \in OR_Set \quad (5)$$

$$IFSD_{i,j} \text{ exist in } (ORT(CS) \text{ and } ORS(CS)) \forall j \text{ Then } SD_i \in OR_Set \quad (6)$$

$$IFSD_{i,j} \text{ exist in } (ORT(CS) \text{ and } N(CS) \text{ and } AU(CS)) \forall j \text{ Then } SD_i \in OR_Set \quad (7)$$

3. 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。

$$\text{IFSD}_{i,j} \text{ exist in } \left(\begin{array}{l} \text{F(CS) and FL(CS)} \\ \text{and N(CS) and LU(CS)} \end{array} \right) \forall j \text{ Then } \text{SD}_i \in \text{OE_Set} \quad (8)$$

$$\text{IFSD}_{i,j} \text{ exist in } (\text{OT(CS) and N(CS) and LU(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{OE_Set} \quad (9)$$

4. 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)。

$$\text{IFSD}_{i,j} \text{ exist in } (\text{OG(CS) and GV(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{OV_Set} \quad (10)$$

$$\text{IFSD}_{i,j} \text{ exist in } (\text{OG(CS) and GV(CS) and PV(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{OV_Set} \quad (11)$$

$$\text{IF } \text{SD}_{i,j} \text{ exist in } \left(\begin{array}{l} \text{ON(CS) and GV(CS)} \\ \text{and PV(CS) and OT(CS)} \end{array} \right) \forall j \text{ Then } \text{SD}_i \in \text{OV_Set} \quad (12)$$

5. 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。

$$\text{IFSD}_{i,j} \text{ exist in } (\text{OFQ(CS) and PV(CS) and N(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{OH_Set} \quad (13)$$

$$\text{IFSD}_{i,j} \text{ exist in } (\text{OH(CS) and N(CS) and FU(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{OH_Set} \quad (14)$$

$$\text{IF } \text{SD}_{i,j} \text{ exist in } \left(\begin{array}{l} \text{OT(CS) and ODT(CS)} \\ \text{and N(CS) and LU(CS)} \end{array} \right) \forall j \text{ Then } \text{SD}_i \in \text{OH_Set} \quad (15)$$

6. 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。

$$\text{IFSD}_{i,j} \text{ exist in } (\text{IC(CS) and B(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{IC_Set} \quad (16)$$

7. 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。

$$\text{IFSD}_{i,j} \text{ exist in } (\text{IG(CS) and GV(CS) and PV(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{IM_Set} \quad (17)$$

$$\text{IFSD}_{i,j} \text{ exist in } (\text{IR(CS) and GV(CS) and OT(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{IM_Set} \quad (18)$$

$$\text{IFSD}_{i,j} \text{ exist in } \text{R(CS)} \forall j \text{ Then } \text{SD}_i \in \text{IM_Set} \quad (19)$$

$$\text{IFSD}_{i,j} \text{ exist in } (\text{RV(CS) and ORT(CS) and PV(CS)}) \forall j \text{ Then } \text{SD}_i \in \text{IM_Set} \quad (20)$$

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。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。於此階段所使用符號定義如下：

符號定義

| | |
|------------------------|---|
| QW_i | 第 i 個詢問句 |
| $QW_{i,j}$ | 第 i 個詢問句中第 j 個詢問詞 |
| AW_k | 第 i 個回覆句 |
| $AW_{k,m}$ | 第 k 個回覆句中第 m 個回覆詞 |
| $C(QW_{i,j})$ | 第 i 個詢問句中出現第 j 個詢問詞之次數 |
| $C(AW_{k,m})$ | 第 k 個回覆句中出現第 m 個回覆詞之次數 |
| $P(AW_{k,m} QW_{i,j})$ | 第 k 個回覆句之第 m 個回覆詞出現於第 i 個詢問句第 j 個詢問詞之機率值 |
| $QWAW_k$ | 第 k 個詢問詞與回覆句配對組合，即回覆集合 |
| $D_{i,q}$ | 第 i 份文件包含 q 個關鍵字詞 |
| $D\omega_q$ | 第 q 個文件關鍵字集合向量 |
| $QWAW_{\omega_k}$ | 第 k 個回覆集合向量 |
| $Sim(D_q QWAW_k)$ | 第 q 個文件與第 k 個回覆集合之相似值 |
| $\omega(D_q, QWAW_k)$ | 文件與回覆集合之篩選門檻值（若回覆集合與文件相似值 $Sim(D_q QWAW_k)$ 大於門檻值即為候選文件 |
| $Sum(D_i)$ | 第 i 份文件與專業語彙之比對總和 |
| Q | 整體語意類別之第三四分位數指標 |
| $CandidateDoc_Set_d$ | 第 d 個具有專業語彙連結之候選文件集合 |

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 ($QW_{i,j}$)，根據詢問詞所取得之回覆句 (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，最後如 **公式(23)** 所示，藉由詢問詞相對於回覆詞語意門檻值 $\omega(AW, QW)$ 篩選與該回覆詞句有相關性之詢問詞以成集合。

$$\begin{aligned} QW_i &= \{QW_{i,1}, QW_{i,2}, QW_{i,3}, \dots, QW_{i,j}, \dots\} \\ AW_k &= \{AW_{k,1}, AW_{k,2}, AW_{k,3}, \dots, AW_{k,m}, \dots\} \end{aligned} \quad (21)$$

$$\begin{aligned} &P(AW_{k,m} | QW_{i,j}) \\ &= \frac{C(AW_{k,1} | QW_{i,1}) + C(AW_{k,1} | QW_{i,2}) \dots + C(AW_{k,1} | QW_{i,j})}{C(QW_{i,j})} \cdot C(AW_{k,1}) \end{aligned} \quad (22)$$

$$\text{IF } P(AW_{k,m}, QW_{i,j}) > \omega(AW, QW) \text{ Then } AW_{k,m}, QW_{i,j} \in QWAW_{\omega_k} \quad (23)$$

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。

$$D_i = \{D_{i,1}, D_{i,2}, D_{i,3}, \dots, D_{i,q}, \dots\} \quad (24)$$

$$D_q^\sigma = [w_1, w_2, \dots, w_q]^T, \quad QWAW_k^\sigma = [w_1, w_2, \dots, w_k]^T$$

$$\text{Sim}(D_q | QWAW_k) = \frac{D_q^\sigma \cdot QWAW_k^\sigma}{|D_q^\sigma| \cdot |QWAW_k^\sigma|} \quad (25)$$

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 $\omega(D_q, QWAW_k)$ 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_Set_d)。

$$\omega(CT, D_i) = \frac{\sum_{all n} \text{Sum}(D_i)}{N(CT_n)}, \text{ IF } \text{Sum}(D_i) \geq \omega(CT, D_i) \text{ Then } D_i \in \text{CandidateDoc_Set}_d \quad (26)$$

$$\varpi(CT, D_i) = \begin{cases} \frac{\text{Sum}(D_{\frac{i+1}{2}})}{2} & \text{IF } N(CT_n) \text{ is odd} \\ \frac{1}{2} \times \left(\text{Sum}(D_{\frac{i+1}{2}}) + \text{Sum}(D_{\frac{i+1}{2}+1}) \right) & \text{IF } N(CT_n) \text{ is even} \end{cases} \quad (27)$$

and IF $\text{Sum}(D_i) \geq \varpi(CT, D_i)$ Then $D_i \in \text{CandidateDoc_Set}_d$
 $Q = N(CT_n) \times 75\%$

$$\varpi(CT, D_i) = \begin{cases} \text{Sum}(D_{i \times 75\%}) & \text{IF } Q \notin \{X : |X| \in N\} \\ \frac{1}{2} \times (\text{Sum}(D_{i \times 75\%}) + \text{Sum}(D_{i \times 75\%+1})) & \text{IF } Q \in \{X : |X| \in N\} \end{cases}$$

and IF $\text{Sum}(D_i) \geq \varpi(CT, D_i)$ Then $D_i \in \text{CandidateDoc_Set}_d$ (28)

IF $\text{Sum}(D_i) \geq \varpi(CT, D_i)$ Then $D_i \in \text{CandidateDoc_Set}_d$ (29)

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。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。如圖5所示，簡略摘要乃計算觀念性語句之向心性以及結構性，並以總分數排序與擷取以簡略說明該份文件之動機與目的；詳述摘要乃根據詞頻取出相關詞彙，再透過詞彙之間起迄總和為基形成詞彙鏈，透過詞彙鏈以擷取相關語句已說明該份文件作業行為、傷害成因與改善方式等描述。

3.3.1 簡略部分建立

結構化摘要之簡略部分主要乃計算語句之向心性程度，再進行語句之結構完整性之篩選。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。語句之結構強度計算乃考量摘要閱讀之順暢性，因而計算該語句之結構強度（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

步驟(A1) 具有觀念性語句之向心性計算

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拆解成二至六字詞。各表達項目之向心性判斷方式如公式(30)至公式(32)所示，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)、Score IC(CS)與 Score TM(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。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 (W₂)、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 (W₃)、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 (W₄)、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 (W₅)、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 (W₆) 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 (W₇)。

$$\begin{aligned} \text{ON}(\text{CS}_i) &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF ON}(\text{CS}_{i,j}) \text{ exist in } \text{OG}(\text{CS}) \forall j &\text{ Then } W_1 = 1, \text{ Otherwise } W_1 = 0 \\ \text{IF ON}(\text{CS}_{i,j}) \text{ exist in } \text{OT}(\text{CS}) \forall j &\text{ Then } W_2 = 1, \text{ Otherwise } W_2 = 0 \\ \text{Score ON}(\text{CS}_i) &= W_1 + W_2 \end{aligned} \quad (30)$$

$$\begin{aligned} \text{IF } 1 < \text{ScoreON}(\text{CS}_i) \leq 2 &\text{ Then } \text{ON}(\text{CS}_i) \in \text{CandidateON} \\ \text{IC}(\text{CS}_i) &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF IC}(\text{CS}_{i,j}) \text{ exist in } \text{PV}(\text{CS}) \forall j &\text{ Then } W_3 = 1, \text{ Otherwise } W_3 = 0 \\ \text{IF IC}(\text{CS}_{i,j}) \text{ exist in } \text{B}(\text{CS}) \forall j &\text{ Then } W_4 = 1, \text{ Otherwise } W_4 = 0 \\ \text{Score IC}(\text{CS}_i) &= W_3 + W_4 \end{aligned} \quad (31)$$

$$\begin{aligned} \text{IF } 1 < \text{ScoreIC}(\text{CS}_i) \leq 2 &\text{ Then } \text{IC}(\text{CS}_i) \in \text{CandidateIC} \\ \text{IM}(\text{CS}_i) &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF IM}(\text{CS}_{i,j}) \text{ exist in } \text{IG}(\text{CS}) \forall j &\text{ Then } W_5 = 1, \text{ Otherwise } W_5 = 0 \\ \text{IF IM}(\text{CS}_{i,j}) \text{ exist in } \text{OT}(\text{CS}) \forall j &\text{ Then } W_6 = 1, \text{ Otherwise } W_6 = 0 \\ \text{IF IM}(\text{CS}_{i,j}) \text{ exist in } \text{RV}(\text{CS}) \forall j &\text{ Then } W_7 = 1, \text{ Otherwise } W_7 = 0 \\ \text{Score IM}(\text{CS}_i) &= W_5 + W_6 + W_7 \\ \text{IF } 2 < \text{ScoreIM}(\text{CS}_i) \leq 3 &\text{ Then } \text{IM}(\text{CS}_i) \in \text{CandidateIM} \end{aligned} \quad (32)$$

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 (作業名稱、傷害成因與改善方法) 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。若該語句具三種結構元素之結合, 其結構之分數為 2。若有動詞但結合受詞或是主詞, 即結構之分數為 1。若完全不符合則為 0。各表達項目之結構強度計算方式如公式(33)至公式(41)。

$$\begin{aligned} \text{ONCS}_i &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF ON}(\text{CS}_{i,j}) \text{ exist in } (\text{PV}(\text{CS})) \text{ OR } (\text{GV}(\text{CS})) \forall j & \\ \text{And ON}(\text{CS}_{i,j}) \text{ exist in } (\text{OT}(\text{CS})) \forall j &\text{ And IF ON}(\text{CS}_{i,j}) \text{ exist in } (\text{ON}(\text{CS})) \forall j \end{aligned} \quad (33)$$

$$\begin{aligned} \text{Then (Score(TS, TV, TO))} &= 2 \\ \text{ONCS}_i &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF ON}(\text{CS}_{i,j}) \text{ exist in } (\text{PV}(\text{CS})) \text{ OR } (\text{GV}(\text{CS})) \forall j & \\ \text{And ON}(\text{CS}_{i,j}) \text{ exist in } (\text{OT}(\text{CS})) \forall j &\text{ Or IF ON}(\text{CS}_{i,j}) \text{ exist in } (\text{ON}(\text{CS})) \forall j \end{aligned} \quad (34)$$

$$\begin{aligned} \text{Then (Score(TS, TV, TO))} &= 1 \\ \text{ONCS}_i &= \{ \text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \dots, \text{CS}_{i,j}, \dots \} \\ \text{IF ON}(\text{CS}_{i,j}) \text{ not exist in } (\text{PV}(\text{CS})) \text{ OR } (\text{GV}(\text{CS})) \forall j & \\ \text{Then (Score(TS, TV, TO))} &= 0 \end{aligned} \quad (35)$$

$$\begin{aligned}
& ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIC(CS_{i,j}) \text{ exist in } (PV(CS)) \text{ OR } (GV(CS)) \forall j \\
& \text{And } IC(CS_{i,j}) \text{ exist in } (IC(CS)) \forall j \text{ And } IFIC(CS_{i,j}) \text{ exist in } (IC(CS)) \forall j
\end{aligned} \tag{36}$$

$$\begin{aligned}
& \text{Then } (Score(TS, TV, TO)) = 2 \\
& ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIC(CS_{i,j}) \text{ exist in } (PV(CS)) \text{ OR } (GV(CS)) \forall j \\
& \text{And } IC(CS_{i,j}) \text{ exist in } (IC(CS)) \forall j \text{ OR } IFIC(CS_{i,j}) \text{ exist in } (IC(CS)) \forall j
\end{aligned} \tag{37}$$

$$\begin{aligned}
& \text{Then } (Score(TS, TV, TO)) = 1 \\
& ICCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIC(CS_{i,j}) \text{ not exist in } (PV(CS)) \text{ OR } (GV(CS)) \forall j \text{ Then } (Score(TS, TV, TO)) = 0
\end{aligned} \tag{38}$$

$$\begin{aligned}
& IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIM(CS_{i,j}) \text{ exist in } PV(CS) \text{ OR } GV(CS) \text{ OR } RV(CS) \forall j \\
& \text{And } IM(CS_{i,j}) \text{ exist in } (OT(CS)) \forall j \text{ And } IFIC(CS_{i,j}) \text{ exist in } (ORT(CS)) \forall j \\
& \text{Then } (Score(TS, TV, TO)) = 2
\end{aligned} \tag{39}$$

$$\begin{aligned}
& IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIM(CS_{i,j}) \text{ exist in } PV(CS) \text{ OR } GV(CS) \text{ OR } RV(CS) \forall j \\
& \text{And } IM(CS_{i,j}) \text{ exist in } (OT(CS)) \forall j \text{ And } IFIC(CS_{i,j}) \text{ exist in } (ORT(CS)) \forall j \\
& \text{Then } (Score(TS, TV, TO)) = 1 \\
& IMCS_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \dots, CS_{i,j}, \dots\} \\
& IFIM(CS_{i,j}) \text{ not exist in } PV(CS) \text{ OR } GV(CS) \text{ or } RV(CS) \forall j \\
& \text{Then } (Score(TS, TV, TO)) = 0
\end{aligned} \tag{40}$$

3.3.2 詳述部分建立

於詳述部分之內容乃為描述部分以及評估部分，其內容多為細項描述。根據上述，針對描述部分所擷取之語句集合包含：作業行為集合以及作業環境集合，而評估部分所擷取集合乃以改善方法集合為主。本階段擷取語句方式乃以下三項為主要步驟，分別為語彙相依程度計算、連結詞之連接語句以及標點符號以判斷結束語句。

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字詞權重計算之方法如公式(42)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(43))。

$$WOD_{i,j} = TFOD_{i,j} \cdot \log \frac{NumSet}{ODF_i} \tag{42}$$

$$Res(S_i) = \{S_{i,1}, S_{i,2}, S_{i,3}, \dots, S_{i,j}, \dots\} \tag{43}$$

步驟(B2) 計算具有觀念性語彙之相依程度

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(44)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 $F[CW_k, CW_{k+1}]$ 表示之，根據公式(45)即可計算語彙 *k* 至語彙 *k+1* 之發生頻率，並如公式(46)至公式(47)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）。

$$I[i,k] = \begin{cases} 1, & \text{IF } CW_k[P_i] \in A(CS), CS_{k+1}[P_i] \in B(CS) \\ 0, & \text{Otherwise} \end{cases} \quad (44)$$

$$F[CW_k, CW_{k+1}] = \sum_{all i, all k} I[i,k] \quad (45)$$

$$P^{From}(k) = \sum_{all k+n} F[CW_k, CW_{k+n}], \quad P^{To}(k) = \sum_{all k} F[CW_{k+n}, CW_k] \quad (46)$$

$$R[CW_k] = \frac{P^{From}(k)}{P^{To}(k)}, \quad SCSW = CW'_{k-1} \rightarrow CW'_k \rightarrow \dots \rightarrow CW'_{k+n} \rightarrow \dots \quad (47)$$

$$\text{Where } R[CW'_{k-1}] \geq R[CW'_k] \geq \dots \geq R[CW'_{k+n}] \geq \dots$$

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(48)至公式(53)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 $Mk_{i,4}=1$ ；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 $Mk_{i,6}=1$ 。

$$MK_{i,1} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in SCSW } \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (48)$$

$$MK_{i,2} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in } C(CS) \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (49)$$

$$MK_{i,3} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in } LinkC(CS) \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (50)$$

$$MK_{i,4} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in } CW_k \text{ and } R[CW_k] = MaxR[CW_k] \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (51)$$

$$MK_{i,5} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in } CW_k \text{ and } R[CW_k] = MinR[CW_k] \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (52)$$

$$MK_{i,6} = \begin{cases} 1, & \text{IF } Res(S_{i,j}) \text{ exit in } EndMark_Set \forall j \\ 0, & \text{Otherwise} \end{cases} \quad (53)$$

如公式(54)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(55)與公式(56)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(57)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。

$$MK = \begin{bmatrix} MK_{1,1} & MK_{2,1} & \dots & MK_{i,1} \\ MK_{1,2} & MK_{2,2} & \dots & MK_{i,2} \\ MK_{1,3} & MK_{2,3} & \dots & MK_{i,3} \\ MK_{1,4} & MK_{2,4} & \dots & \dots \\ MK_{1,5} & MK_{2,5} & \dots & MK_{i,5} \\ MK_{1,6} & MK_{2,6} & \dots & MK_{i,6} \end{bmatrix} \quad (54)$$

$$Sum1(Res(S_i)) = \sum_{j=2}^3 MK_{i,j} \quad (55)$$

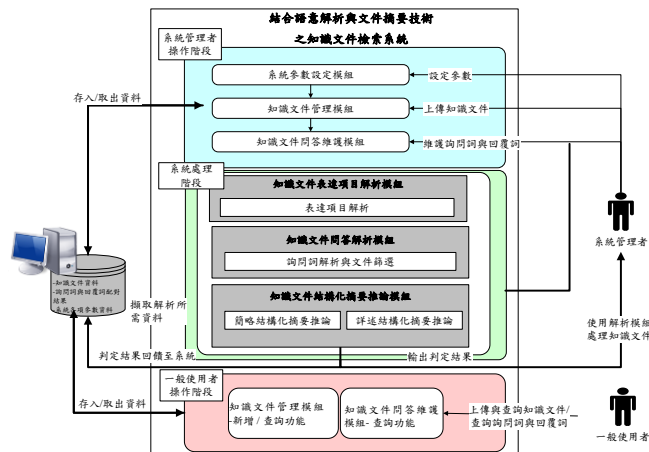
$$Sum2(Res(S_i)) = \sum_{j=5}^6 MK_{i,j} \quad (56)$$

$$SCSW = CW'_{k-1} \rightarrow CW'_k \rightarrow \dots \rightarrow CW'_{k+n} \rightarrow \dots$$

$$\text{Where } R[CW'_{k-1}] \geq R[CW'_k] \geq \dots \geq R[CW'_{k+n}] \geq \dots \quad (57)$$

第四章、系統架構與規劃

依據本研究所發展之結合語意解析與文件摘要技術之知識文件檢索方法論，本研究乃於網際網路環境下建構一套結合語意解析與文件摘要技術之知識文件檢索系統，以供使用者透過本系統進行問答語意解析以及透過結構化摘要取得相關文件。其系統運作架構如圖 4 所示。當使用者上傳未推論之知識文件後，系統管理者乃需完成參數設定以及新增詢問詞與回覆詞詞彙，待設定與新增完畢後，系統管理者即可執行知識文件表達項目解析模組以取得該文件之各表達項目之觀念性語句與文件之關鍵字詞。系統管理者再藉由知識文件問答解析模組判定搜尋字串之詢問目的進而篩選相關文件，亦可透過知識文件結構化摘要模組將文件進行簡略結構化摘要與詳述結構化摘要，最後使用者即由知識文件管理模組之查詢功能檢視系統之知識文件與判定結果，亦可透過知識文件問答維護模組檢視詢問詞與回覆詞配對組合結果。



- summarization-inspired temporal-relation extraction,” *Information Processing and Management* (43:6), 2007, pp. 1681-1704.
5. Dunlavy, D. M., O’Leary, D. P., Conroy, J. M. and Schlesinger, J. D. “QCS: A system for querying, clustering and summarizing documents,” *Information Processing and Management* (43:6), 2007, pp. 1588-1605.
 6. Elhadad, N., Kan, M. Y., Klavans, J. L. and McKeown, K. R. “Customization in a unified framework for summarizing medical literature,” *Artificial Intelligence in Medicine* (33:2), 2005, pp. 179-198.
 7. Han, K. S., Song, Y. I., Kim, S. B. and Rim, H. C. “Answer extraction and ranking strategies for definitional question answering using linguistic features and definition terminology,” *Information Processing and Management* (43:2), 2007, pp. 353-364.
 8. Huang, M., Zhu, X. and Li, M. “A hybrid method for relation extraction from biomedical literature,” *International Journal of Medical Informatics* (75:6), 2006, pp. 443-455.
 9. Jones, M., Love, B. C. “Beyond common features: The role of roles in determining similarity,” *Cognitive Psychology* (5:3), 2007, pp. 196-231.
 10. Jung, W., Ko, Y. and Seo, J. “Automatic text summarization using two-step sentence extraction,” *Lecture Notes in Computer Science* (3411:71-8), 2005.
 11. Ko, Y., Kim, K. and Seo, J. “Topic keyword identification for text summarization using lexical clustering,” *IEICE transactions on information and systems* (E86-D:9), 2003, pp. 1695-1701.
 12. Ko, Y., Park, J. and Seo, J., “Improving text categorization using the importance of sentences,” *Information Processing and Management* (44:1), 2004, pp. 65-79.
 13. Ko, Y., Seo, J., “An effective sentence-extraction technique using contextual information and statistical approaches for text summarization,” *Pattern Recognition Letters* (29:9), 2008, pp. 1366-1371.
 14. Li, Q., Chen, Y. P. “Personalized text snippet extraction using statistical language models,” *Pattern Recognition* (43:1), 2010, pp. 378-386.
 15. Moens, M. F. Summarizing court decisions,” *Information Processing and Management* (43:6), 2007, pp. 1748-1764.
 16. Oh, H. J., Myaeng, S. H. and Jang, M. G. “Effects of answer weight boosting in strategy-driven question answering,” *Information Processing and Management* (48:1), 2012, pp. 83-93.
 17. Oh, H. J., Sung, K. Y., Jang, M. G. and Myaeng, S. H. “Compositional question answering: A divide and conquer approach,” *Information Processing and Management* (47:6), 2011, pp. 808-824.
 18. Steinberger, J., Poesio, M., Kabadjov, M. A. and Jezek, K., “Two uses of anaphora resolution in summarization,” *Information Processing and Management* (43:6), 2007, pp. 1663-1680.
 19. Teng, C., Xiong, N., He, Y., Yang, L. T. and Liu, D., “A behavioural mode research on user-focus summarization,” *Mathematical and Computer Modelling* (51:7-8), 2010, pp. 985-994.
 20. Zajic, D., Dorr, B. J., Lin, J. and Schwartz, R., “Multi-candidate reduction: Sentence compression as a tool for document summarization tasks,” *Information Processing and Management* (43:6), 2007, pp. 1549-1570.