以模糊限制為基礎之使用者描述紀錄學習方法設計

陳聖濤	余丁榮	左杰官
國防大學資訊管理學系	國防大學資訊管理學系	國防大學資訊管理學系
研究生	助理教授	教授
iiccanffly@gmail.com	cecyu.tw@gmail.com	brandttso@yahoo.com.tw

摘要

隨著網路科技快速發展，網路資料巨量的成長，使用者必須從龐大的回傳結果中自行過濾出感興趣的資訊，讓使用者面對資訊過載的窘境與資訊過濾的負擔，而使用者描述紀錄(User Profile)的建構，為協助使用者進行資訊過濾，降低資訊過載困擾的方法之一。

本研究導入模糊限制的方法來建構與學習使用者描述紀錄，在知識表達方面，以模糊限制式構連而成的模糊限制網路來呈現字詞語意的不明確性及概念間的多元關係，並以模糊限制網路表示多重興趣主題之單一使用者描述紀錄；在解題方面，將使用者描述紀錄的建構視為模糊限制滿足問題，藉由關係強度激發擴散模式萃取出多重興趣的主題，並依文件相關度分數作為資訊過濾的依據，呈現於回傳結果中，滿足使用者的檢索需求。

關鍵詞：使用者描述紀錄、模糊關係、模糊限制、激發擴散、資訊過濾。

1. 前言

隨著網路科技的快速發展，使用者往往需要透過搜尋引擎鍵入關鍵字來找尋需要的資訊，惟當今網路資料巨量的成長，使用者必須從龐大的回傳結果中自行過濾出自己感興趣的資訊，致使使用者不得不面對資訊過載(Information Overload)的窘境與資訊過濾(Information Filtering, IF)的負擔與困擾。

目前的搜尋引擎大部份採比對使用者輸入關鍵字的方式搜尋資料，亦即比對每個網頁/文件資料中是否含有使用者的關鍵字，若該網頁/文件資料含有此關鍵字就將此網頁/文件資料回傳給使用者。然而，使用者所使用同一概念(Concept)的關鍵字與網頁/文件資料內容所呈現的涵義不盡相同，致使資訊擷取的效果不盡理想。例如：當 A 使用者是個軍事迷，則關鍵字「阿帕契(Apache)」應該顯示美軍的阿帕契戰鬥直升機的相關網頁/文件資料；若使用者 B 是資訊領域愛好者，對於「阿帕契(Apache)」這個關鍵字應搜尋出開放原始碼之阿帕契網頁伺服器相關網頁/文件資料;若使用者 C 是從事文化工作者，對於「阿帕契(Apache)」這個關鍵字則應檢索出名為阿帕契的美洲印第安部落一族的相關網頁/文件資料。然而，因使用者輸入關鍵字的語言不同，搜尋引擎所回傳的搜尋結果與顯示優先順序也不同。以 Google 搜尋引擎為例，若使用者輸入「阿帕契」，搜尋結果會優先顯示以「阿帕契戰鬥直昇機」的相關網頁/文件資料；若輸入「Apache」關鍵字，則搜尋結果會將有關 Apache 網頁伺服器的網頁/文件資料優先列出。因此，不僅使用者 C 需要篩選數個頁面後，才能找到其感興趣的阿帕契印第安部落內容，對使用者 A 而言，倘若輸入英文的「Apache」關鍵字，或對使用者 B 而言，倘若輸入中文的「阿帕契」關鍵字，均須導致額外資訊過濾的負擔。

在許多資訊檢索的研究[1,4,7]中，利用知識本體來建構字詞(Term)語意或概念間的關係，目的在於搜尋出與使用者輸入關鍵字之語意或概念相符之資訊。然而，字詞的語意或概念常是模糊或不精確的，很難以同義或反義的二元描述方式來表達彼此間的關係，且相同的字詞或概念對每個使用者其感興趣的程度不同，使用者仍須從大量的檢索結果中，花費一定時間過濾出自己感興趣的內容。對此，亦有許多研究[18, 19, 27, 29]藉由使用者瀏覽紀錄的蒐整，學習使用者描述紀錄，篩選出使用者感興趣的資訊。在這些研究中，使用者描述紀錄的學習，往往是將使用者的多重興趣(Multiple Interests)視為彼此獨立不具相關性，進而分解成多個單一興趣主題的描述紀錄以方便學習或處理。雖然，Nanas et al. [19]將使用者多重興趣間的語彙與主題關聯性納入考量，提出以單一使用者描述紀錄來表現使用者

多重主題的興趣；然而，使用者描述紀錄學習的前提均需要將文件依主題予以事先分類，如此也增加了使用者的負擔。因此，如何在龐大的網頁資料中，建構能滿足使用者需求資訊的使用者描述紀錄，來降低使用者資訊過濾的負荷，便成了重要的研究課題之一。

據此，本研究以軍事新聞為使用者描述紀錄之學習資料集，並導入了模糊限制(Fuzzy Constraint)的方法來建構與學習使用者描述紀錄，在知識表達(Knowledge Representation)方面，以模糊限制式構連而成的模糊限制網路(Fuzzy Constraint Network)來呈現字詞語意或概念間的多元關係，並以模糊限制網路表示具多重興趣主題之單一使用者描述紀錄；在解題(Problem-solving)方面，將各使用者興趣主題的識別視為模糊限制滿足(Fuzzy Constraint Satisfaction)問題，再藉由關係強度激發擴散模式來萃取出多重興趣的主題，並按照文件相關度分數作為資訊過濾的依據，呈現於回傳結果之中，滿足使用者的檢索需求。

誠如前述所言，以使用者瀏覽紀錄為基礎建構之使用者描述紀錄，往往涵蓋使用者的多重興趣主題，若依多重的興趣主題建置多個使用者描述紀錄，將疏忽興趣主題間的相關性，而回歸傳統之某概念字詞屬於某一特定興趣主題的二元描述。故本研究旨在透過軍事新聞瀏覽紀錄的學習，建構一個涵括多興趣主題的單一使用者描述紀錄，並據此呈現興趣主題間多元關係的模塑及資訊過濾的進行，將使用者真正感興趣的內容呈現於回傳結果中，加速軍事新聞檢索的實質效率。茲將本研究重點及其目的概述如下：

一、詞彙處理：以國防部青年日報社(http://news.gpwb.gov.tw/)的軍事新聞網頁為資料來源，藉由記錄使用者的瀏覽紀錄，並將瀏覽紀錄中青年日報的網頁內文文字去除格式後擷取下來，經中研院研發之 CKIP 中文斷詞系統進行處理，從中擷取詞類為名詞與動詞之字詞，作為後續處理的輸入。

二、模糊邏輯：傳統擴展查詢主要係以使用者檢索之關鍵字詞與其同義詞為搜尋基礎，然而，這種以字詞「同義」與「不同義」二元描述的擴展查詢，經常無法滿足對字詞意義認知程度不同的使用者，而造成檢索的困擾。本研究導入模糊邏輯(Fuzzy Logic)的觀念，試圖將字詞的二元描述方式，擴展為具同義或不同義歸屬程度(Membership Degree)的多元描述，以呈現出「近義」字詞在檢索的重要性。

三、模糊限制：使用者描述紀錄的學習在資訊檢索上扮演著資訊過濾的角色。在使用者查詢的興趣主題中，往往彼此又具關聯性；因此，本研究利用模糊限制來呈現字詞或概念間的關係，以字詞或概念間關係為基礎所構建的模糊限制網路，用來表示為使用者描述紀錄，並藉由檢索字詞與使用者描述紀錄中各興趣主題的字詞進行比對，將檢索字詞與各興趣主題相關字詞的比對問題視為模糊限制滿足問題，以字詞比對結果所涉及之興趣主題表該模糊限制滿足問題所得之解，用於輔助使用者進行資訊過濾。

四、主題識別：使用者興趣主題的識別涉及使用者描述紀錄字詞間關係的構連，本研究首先利用加權反向文件頻率(Weighted Inverse Document Frequency, WIDF)計算出資料集中各字詞的權重；其次，以字詞共現的頻率(Co-occurrence Frequency)來決定字詞間一般性(Generality)與特殊性(Specificity)的從屬關係，以建構出具字詞階層關係之使用者描述紀錄(即模糊限制網路)；最後，再採本研究所提之關係強度激發擴散模式(Spreading Activation Model of Relationship Strength)來萃取出多重興趣的主題。

五、資訊過濾：藉由比對使用者輸入之搜尋關鍵字詞與字詞階層關係構建而成的使用者描述紀錄，關鍵字詞所屬之使用者興趣主題即可成為資訊過濾的依據，將使用者真正感興趣的內容按照文件相關度分數排序，並呈現於回傳結果之中。

本論文後續章節的安排上，首先在第二節中探討有關使用者描述紀錄學習之相關文獻；其次，在第三節中探討如何將使用者描述紀錄視為模糊限制網路及其建構程序；第四節則就本研究所提方法論就其知識表達的適切性與解題的時間複雜度進行評估；最後，提出結論與未來研究方向。

2. 文獻探討

在資訊檢索的領域中，隨著網際網路的發展，要利用搜尋引擎在這龐大的網際網路中找尋需要的資訊，已變得愈來愈困難，且不同使用者鍵入相同的關鍵字來查詢時，搜尋引擎並不判別使用者是否不同，僅會回傳相同的結果。在一般情況之下，不同的使用者通常會有不同的資訊需求，所以資訊檢索應該適應不同使用者的資訊需求。

為了預測這樣的資訊需求，許多研究[9, 11, 28, 31]提出以資料探勘的方法於網頁紀錄(Web Log)中萃取習慣的樣式，但這些習慣資料中發掘的樣式並不能完全實現個人化的工作，故亦另有研究[19, 20, 33]提出從瀏覽網頁紀錄中，利用計算字詞權重的方法來建構使用者描述紀錄，但都需要使用者事先進行文件的類型指定。傳統為了要實現個人化資訊或提供各相關資訊給使用者，可歸納為「使用具相關性回饋的系統」、「使用者登錄其興趣或具有統計資訊(Demographic Information)的系統」、「基於使用者評價來推薦資訊的系統」三種方式來提供上述資訊[29]。在相關性回饋的系統中，使用者必須事先登錄像興趣、年齡等個人化的資訊，或是提供尺度從 1(不相關)到 5(非常相關)的相關性判斷的回饋，而推薦系統則收集使用者這種評價形式的回饋給特定領域的項目，並且從許多使用者中相似卻不相同的使用者描述紀錄來決定如何推薦項目；如果使用者願意提供評價則能提供額外的評價資訊；然而，實際上因為這些類型的登錄、回饋或是評價是耗時的，所以通常使用者不太願意提供評價，也造成推薦的準確率降低。

Sugiyama et al. [29]提出了基於瀏覽歷史的方式來建構使用者描述紀錄，以呈現使用者瀏覽網頁的喜好。在這方法中將使用者瀏覽的歷史分為隨著時間持續發展的持續喜好(Persistent Preferences)，及只收集當天瀏覽活動的短暫喜好(Ephemeral Preferences)，並以內隱地方式從瀏覽歷史中得到的使用者描述紀錄定義為上述兩種喜好的加總，使用者描述紀錄表示如下：

$$P = a\,P^{per} + b\,P^{today};$$

其中 P 代表使用者描述紀錄；P^{per} 代表持續的喜好，P^{today} 代表短暫的喜好，式子中給予一個調整比重的常數 a 及 b，且 $a + b = 1$。

雖然，這種使用者描述紀錄雖可以透過內隱的學習來獲得使用者的喜好，但這種資訊過濾的方法卻是利用比較使用者與相鄰近使用者對於特定字詞賦予權重的關係，來協同地將資訊過濾出，這種參考鄰近使用者喜好的方法並無法完整呈現使用者個人的喜好。

Nanas et al. [19]提出透過使用者事前指定與興趣相關的文件中萃取出字詞，利用這些字詞的集合代表使用者興趣，並計算字詞及字詞間連結的權重的方法來建構使用者描述紀錄，再以字詞階層網路的形式來呈現，最後再利用激發擴散的方式過濾出使用者真正感興趣的主題及該主題所屬字詞，來描述使用者的興趣。為促使使用者描述紀錄可用來描述多主題的興趣，必須將字詞間關聯度納入考量，包含了字詞間詞彙和主題的關聯度，這方法改善了從資訊檢索及文字分類(Text Categorization)中延伸出來的資訊過濾所忽略字詞相關性的缺點；但計算字詞的權重時，需要由使用者判斷字詞與文件是否與全部的字詞與文件相關，並且依主題予以事先分類，因此該研究在實際運用上著實降低了實用性。

Solskinnsbakk and Gulla [27]提出從特定領域的文件集中，用句子、段落及文件的集合來建構與每篇文件的索引，在這三類的索引中，單一句子中的字詞有最高的語意一致性，其次是段落，文件則表具最低的語意一致性，完成索引後賦予文件知識本體的概念並計算權重來建構知識本體的使用者描述紀錄，其概念如圖 2-1 所示。

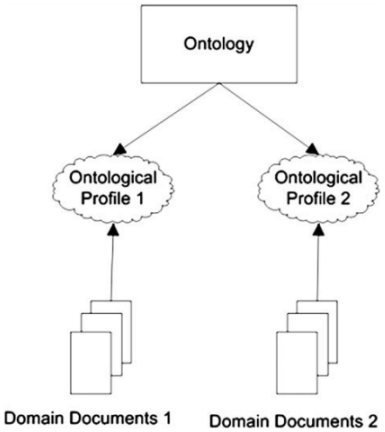


圖 2-1 本體論的使用者描述紀錄建構概念 (資料來源：[27])

Mezghani et al. [18]提出，使用者描述紀錄建構方法可依照使用者資訊修改的頻率區分為，收集像姓名、年齡及住址等很少改變的資訊的靜態法，收集頻繁改變資訊的動態法，而使用者的資訊可以由使用者自己外顯地

(Explicitly)獲得，或內隱地(Implicitly)觀察使用者的瀏覽行為(瀏覽歷史、點擊及瀏覽的頁面等)，並可利用社群網路(Social Networks)以標籤(Tag)為基礎，作為使用者描述紀錄的建構方式。這方法雖可解決因社群網路持續發展造成使用者、資源及互動數量持續增加產生資訊遺失及無法找到有用的資訊，但標籤註解(Annotation)的方式可能因標記人的不同而有不同的意思，因而造成語意模糊的標籤，另外因透過社群網路，很多使用者都可對某一個資源進行註解，造成這些大量的標籤就像垃圾郵件般，變成無效的資訊；另外，這種群眾分類法(Folksonomy)是非常多變的，缺乏分類而無脈絡可循，因此這種分類法的標籤對於相同的字可能有不同的格式(例如 walk、walking 及 walked 等)，故無法處理同義或同音字等缺點。

綜合以上文獻，各研究方法仍有以下潛在的問題：

一、建構使用者描述紀錄時，透過使用者註冊或回饋方式，來得到使用者的相關資訊與喜好，礙於使用者配合程度的不同，使用者描述紀錄的學習效果將會很有限。

二、利用社群網路以標籤為基礎的使用者描述紀錄建構方式是一個新的概念，但其中涉及到標籤賦予的語意與標籤氾濫的隱憂，如何判斷標籤的代表性與適用性，仍然沒有明確的答案。

三、雖已將字詞與字詞、字詞與興趣主題的關係納入考量，提升了資訊過濾的效果，但這種使用者描述紀錄僅能表達出字詞間同義或反義的二元關係，若出現字詞的語意或概念間的模糊或不清確的情況，就無法把字詞間的多元關係表達出來。

四、涵蓋使用者多重興趣主題之使用者描述紀錄，將使用者的多重興趣視為彼此獨立不具相關性，進而分解成多個單一興趣主題的描述紀錄，將忽略興趣主題間的相關性，而回歸傳統之某概念字詞只屬於某一特定興趣主題的二元描述，無法真實描述概念字詞的關係。

針對上述問題，本研究擬以下不同方法，試圖解決使用者描述紀錄學習的困境，內容摘述如后：

一、以使用者瀏覽紀錄為基礎建構用以描述使用者習慣或喜好的使用者描述紀錄，採內隱的方式來觀察使用者的瀏覽紀錄，避免耗時的使用者資料登錄、回饋或評價。

二、導入模糊邏輯的方式，將字詞的二元關係擴展為具同義或不同義歸屬程度的多元關係，呈現「近義」字詞在檢索的重要性。

三、利用模糊限制來呈現字詞或概念間的關係，以字詞或概念間關係為基礎所構建的模糊限制網路，用來表示為具多重興趣主題的單一使用者描述紀錄，並藉由檢索字詞與使用者描述紀錄中各興趣主題的字詞進行比對，將檢索字詞與各興趣主題相關字詞的比對問題視為模糊限制滿足問題，以字詞比對結果所涉及之興趣主題表該模糊限制滿足問題所得之解，用於輔助使用者進行資訊過濾。

3. 以模糊限制網路模塑使用者描述紀錄

3.1 使用者描述紀錄視為模糊限制網路

本研究旨在以模糊限制的概念作為使用者描述紀錄的知識表達，並利用模糊限制的解題方法，來學習出使用者的興趣主題及其相關概念字詞集合。由於特定興趣主題為特定領域知識的表達，故可視為是特定領域之知識本體，而知識本體誠如 Gruber [12]的定義為對概念的規範，即概念是由特定屬性所描述；因此，將特定興趣主題對應到知識本體的定義，可視特定興趣主題為概念，用以描述興趣主題內容的字詞可被視為是描述概念的屬性；此外，特定字詞又可同時用以描述不同的興趣主題，故興趣主題間仍具相關性，不盡然是獨立不相關。故本研究將使用者描述紀錄定義為使用者習慣或喜好瀏覽的興趣主題資料集及其關聯，並以模糊限制網路來模塑使用者描述紀錄，意即將模糊限制式用以對應使用者描述紀錄中各興趣主題間的關聯，模糊限制式中之變數及其值域對應使用者描述紀錄中之興趣主題及其描述的字詞，詳如圖 3-1 所示。

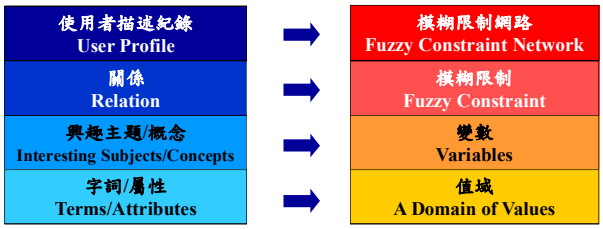


圖 3-1 使用者描述紀錄與模糊限制網路對應圖 (資料來源：本研究整理)

模糊限制兼具模糊邏輯與限制式知識表達與解題方法的特性，這也是為何本研究以模糊限制為基礎來進行知識表達與解題的主要原因。以下茲就模糊邏輯與模糊限制分別予以說明。

3.1.1 模糊邏輯

傳統僅以「屬於或不屬於」或「符合或不符合」明確的布林邏輯(Boolean Logic)二元方式來描述事實，在實際應用上確有不足之處。而這種僅以「屬於或不屬於」或「符合或不符合」的二元知識表達方式，若將「屬於」或「符合」以 1 或 True 表示，「不屬於」或「不符合」以 0 或 False 代表，則可將明確集合(Crisp Set)歸屬函數定義為 $\mu_{\mathcal{F}} : X \rightarrow \{0,1\}$。

Zadeh [32]提出了模糊集合(Fuzzy Set)的觀念，強調以模糊邏輯來描述生活周遭中事物性質的等級，並建立一套語言分析的數學模式，將人類思維及語言上之模糊性與不確定性具體的呈現。模糊理論的觀念在強調以模糊邏輯來描述現實生活中事物的等級，以彌補傳統布林邏輯無法描述對事物具不明確邊界定義的缺點。故模糊理論將模糊概念，以模糊集合對所描述的事物進行定義，藉由模糊量化得到一歸屬程度，來處理各種問題。為使後續模糊領域知識的表達更清楚，本研究將模糊領域相關觀念定義如下：

定義 1：模糊集合－模糊集合 F 是由一特定領域(Domain) X 中的物件所組成，在 F 中的每個物件 $x_i \in X$ 其歸屬程度是由歸屬函數 $\mu_{\mathcal{F}} : X \rightarrow [0,1]$ 所定義。

從其歸屬函數定義而言，明確集合為模糊集合的一種特例。而歸屬函數通常是根據操作經驗或統計加以確定，亦即根據使用者的主觀意識來做判別，然後透過學習及實踐經驗逐步修正與調整使歸屬函數更加完善及客觀。

定義 2：模糊關係(Fuzzy Relation)－模糊關係 $R(X,Y)$ 是由作用於領域 $X \times Y$ 之模糊集合 R 所定義，其中，X 與 Y 為兩個明確集合，在模糊集合 R 中每個物件 (x_i, y_i) 的歸屬程度由歸屬函數 $\mu_{\mathcal{R}} : X \times Y \rightarrow [0,1]$ 所定義；其中，所有的 $x_i \in X$ 且 $y_i \in Y$。

除了二元的模糊關係外，模糊關係亦可擴充為多元關係，以三元模糊關係 $\widetilde{R}(X,Y,Z)$ 為例，X, Y 與 Z 為明確集合，在模糊集合 R 中每個物件 (x_i, y_i, z_i) 的歸屬程度由歸屬函數 $\mu_{\mathcal{R}} : X \times Y \times Z \rightarrow [0,1]$ 所定義。

本研究導入模糊邏輯觀念，旨在將字詞傳統「同義」與「不同義」之二元關係，擴展為具同義或不同義歸屬程度的多元關係，以呈現出「近義」字詞在檢索的重要性。

3.1.2 模糊限制

限制式(Constraint)是描述一個或多個物件/變數間彼此相互關係的概念；限制滿足問題(Constraint Satisfaction Problem; CSP)就是要找出所有物件/變數皆符合特定關係的解答，亦即求出滿足所有物件/變數限制式的解答，而其結果則可能有解，也可能無解。然而，許多應用中，變數可能具有不確定性或是具有雜訊。亦有情形，是限制滿足問題會發生限制過多或是限制不足的狀況。此時，需要增加或去除部分的限制來解決問題，而非增加或去除整個限制。亦即實際上可能只需要修改限制式的一部份。另外，Davis [10]與 Mackworth [17]提出限制滿足問題的求解過程是一個 NP-hard 的問題，因此如何有效地找出同時滿足一組限制的解，就成為本研究的研究議題之一。無論上述哪種情形，都有必要將模糊邏輯的概念整合到限制滿足問題中，而 Bowen et al. [8]、Lai [13]、Lai et al. [14]與 Lin et al. [16]等人均分別提出模糊限制與模糊限制網路來表示具不精確性及不確定性資訊物件間的關係。

模糊限制可看成是一個從 $\mathbf{U} = \mathbf{U}_1 \times ... \times \mathbf{U}_k$ 定義域(限制式中變數的有限定義域卡笛生乘積)映射到[0, 1]的區間，對模糊限制 C 而言，$C(x_1,...,x_k)$ 代表值組$(x_1,..,x_k)$如何滿足限制式 C 的程度。模糊限制對應到模糊集合的歸屬函數，以$(x_1,..,x_k)$來說，

$\mu_C(x_1,...,x_k)$ 是藉由實例(x_1,..,x_k)滿足限制式 C 的滿意程度。如果 $\mu_C(x_1,...,x_k)=1$ 可以說(x_1,..,x_k)完全滿足 C，而當 $\mu_C(x_1,...,x_k)=0$ 則可以說(x_1,..,x_k)完全不滿足 C。$\mu_C(x_1,...,x_k)>0$ 中的(x_1,..,x_k)值組即所謂的支持集合(Support Set)以 S 表示。令 α (1≥α>0)表滿足限制式 C 的程度，則 α-支持集合(α-support Set)表示為 S_α 是所有(x_1,..,x_k)∈ $\mathbf{U}_1\times...\times\mathbf{U}_k$ 的集合，其中，$\mu_C(x_1,...,x_k)\geq\alpha$ [24]。

現實環境中的問題，往往不能由單一模糊限制式來呈現，而需要多個或一群模糊限制式才能完成知識的表達；因此，本研究採用 Lin et al. [16]所描述的模糊限制網路來代表模糊限制式的集合體，其定義如下：

定義 3：模糊限制網路－模糊限制網路 $\widetilde{\Omega}$=(**U**, **X**, **C**)代表模糊限制所成集合；其中，$\mathbf{U}=(\bigcup_{k=1}^{m}U_k)$ 為用以描述所有變數值域所成的宇集合(Universal Set)，$\mathbf{X}=(\bigcup_{k=1}^{m}X_k)$ 表由非重複出現的變數所成的集合，$\mathbf{C}=(\mathrm{C}^i\bigcup\mathrm{C}^e)$ 則代表用以表達變數關係之內部(Internal)與外部(External)模糊限制所成集合；其中，內部模糊限制集合 C^i 為變數內因值域的規範所成模糊限制的集合，而外部模糊限制集合 C^e 為變數間關係所成模糊限制的集合。

令模糊限制網路(**U**, **X**, **C**)所求得的解，寫成 $\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}$，為在模糊限制網路中的變數集合 **X** 符合限制式所成集合 **C** 的每個限制式之(n-ary)可能性分配(Possibility Distribution)，意即

$$\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}=\overline{C}_1(T_1)\cap...\cap\overline{C}_m(T_m); \quad (1)$$

而每個限制式 $C_j(T_j)\in\mathbf{C}$，$\overline{C}_j(T_j)$ 為在變數空間 $\mathbf{X}=(X_1,...,X_m)$ 中各個對應的圓柱狀延伸(Cylindrical Extension)。

同時，${}_\alpha\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}$ 視為 $\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}$ 在門檻值(Threshold)或滿意度 α 以上的解所成的集合。在模糊限制網路之限制總滿意度被定義為在模糊限制網路中之最低限制滿意度，意即

$$\mu_{\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}}(\mathbf{u})=\min_{j=1...n}(\mu_{C_j}(\mathbf{u})); \quad (2)$$

其中，**u** 表變數集合 **X** 中之任一值組(Tuple)。

定義 4：模糊限制滿足－給定一個模糊限制網路(**U**, **X**, **C**)，門檻值 α 為[0, 1]區間的實數，模糊限制滿足問題為決定是否，

$${}_\alpha\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}\neq\{\}; \quad (3)$$

$\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}$ 視為滿足所有限制之一組解。

在模糊限制滿足問題中，$\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}$ 中變數組合 $(x_1,...,x_m)$ 的滿意度則寫成 $\mu_{\Pi_{\mathbf{U},\mathbf{X},\mathbf{C}}}(x_1,...,x_m)$，簡化為 $\mu_C(x_1,...,x_m)$。

3.1.3 模糊限制模塑理由

本研究以模糊限制網路來模塑使用者描述紀錄之理由如下：

一、不確定性或不精確性的知識表達：傳統僅以「屬於或不屬於」或「符合或不符合」明確的布林邏輯二元方式來描述事實，在實際應用上卻有無法描述對事物具不明確邊界定義的缺點。模糊限制可援用模糊邏輯將模糊概念，以模糊集合對所描述的事物進行定義，藉由模糊量化得到一歸屬程度，即在「屬於或不屬於」或「符合或不符合」間多一些等級的區分，如：「屬於的程度」或「符合的程度」，來處理各種問題。本研究導入模糊邏輯觀念，旨在將字詞傳統「同義」與「不同義」之二元描述方式，擴展為具同義或不同義歸屬程度的等級式描述，以呈現出「近義」字詞在檢索的重要性。

二、概念字詞間多元關係的模塑：概念字詞間的關係不全然是二元關係，雖然也可以二元關係反覆進行關係的描述，但模糊限制具限制式多元關係表達的特性，不僅可以更簡潔地進行知識表達，也更具關係的可讀性。雖然，現實環境中的問題，往往不能由單一模糊限制來呈現，而需要多個或一群模糊限制式才能完成知識的表達；然而，本研究以模糊限制網路來代表模糊限制的集合體，除可模塑多元關係外，亦易於視覺化的關係呈現。

三、漸進式解題效率的提昇：模糊限制的解題就是個模糊限制滿足問題，即在求得滿足所有模糊限制的解。由於每一模糊限制均

具模糊量化的歸屬程度，解題時是由各模糊限制歸屬程度最高的值開始進行，當無法得滿足所有模糊限制的解時，各模糊限制再以歸屬程度次高的值重複進行解題，直到找出符合特定滿意程度的解為止。

在本研究中是利用其漸進式的解題方式來計算各概念字詞的階層關係，以建構出使用者描述紀錄，意指各則文件中各概念字詞均有其出現的頻率，每則文件先釋出概念字詞出現頻率最高者，用來計算各概念字詞間共同出現於新聞中的頻率，以決定概念字詞的階層關係，計算完成後，再以各則文件概念字詞出現頻率次高者繼續計算，依此類推，直到概念字詞出現頻率不符特定滿意程度時即停止。由於使用者描述紀錄的建構每個概念字詞須與其他字詞逐一比較其階層關係，故識別出 n 個概念字詞間階層關係的計算次數為 $n(n\text{-}1)$，故其最差、最佳及平均的時間複雜度均為 $\mathbf{O}(n^2)$。

而本研究所提模糊限制特有的漸進式解題方式，此演算法的時間複雜度雖與傳統限制式的解題方法同為 $\mathbf{O}(n^2)$，但利用解模糊限制的方式，先將字詞滿意度由高至低的遞減方式排序後，透過使用者自訂的滿意度門檻值，將門檻值以上的概念字詞由高至低地代入計算字詞之階層關係，最差狀況即當滿意度門檻值為 0 時，計算次數始與傳統限制式之計算次數相同，最佳狀況只需計算 2 次即可；由此可知，導入的模糊限制式能降低運算的次數，提升解題的效率，減少運算資源的浪費。

四、 知識表達與解題方法的統一：本研究以模糊限制作為知識表達之基礎，也利用模糊限制本身的解題特性，提供漸進式之解題方法，減少計算的複雜度，而無須援用其他理論或方法論，有效降低學習的複雜度。

3.2 研究架構

本研究以電子報的軍事新聞為研究資料的來源，並依使用者所瀏覽的軍事新聞內文透過中文字詞處理，將軍事新聞內文的字詞萃取出來，做為使用者習慣性瀏覽主題使用者描述紀錄的學習資料，以建構出可代表字詞間多元關係的模糊限制網路，再利用本研究所提關係強度激發擴散的能量傳遞概念，以解模糊限制滿足問題的方法產生使用者對於軍事新聞的興趣主題模糊限制網路，並且使用此適用於使用者的興趣主題模糊限制網路來過濾資訊，最後再將回傳結果進行排序，將相關度較高的網頁優先呈現給使用者，以減低使用者過濾資訊的負載問題。

3.3 研究步驟

承上所述，本研究區分為四個階段，第一階段為「資料蒐集」，第二階段為「中文斷詞處理」，第三階段為「同義字詞編修」，第四階段為「模糊概念關係識別」，第五階段為「模糊限制網路建構」，第六階段為「模糊限制網路裁剪」，第七階段為「興趣主題網路萃取」，第八階段為「資訊過濾」，研究架構詳如圖 3-2 所示，各階段說明一一分述如下。

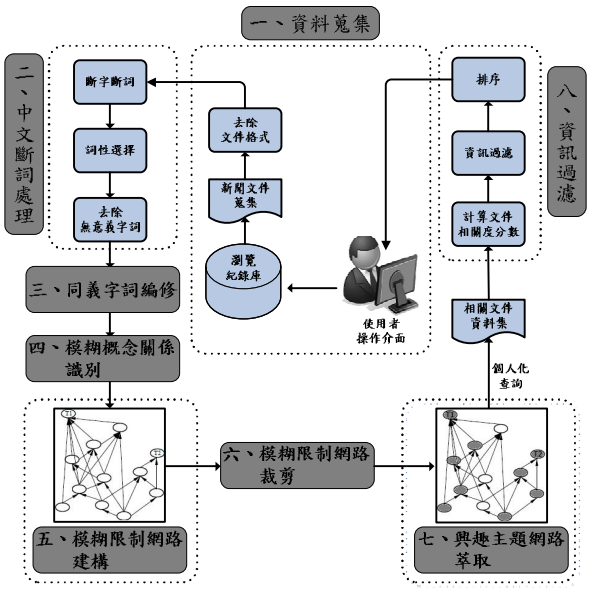


圖 3-2 研究架構圖
(資料來源：本研究整理)

3.3.1 資料蒐集

一、 新聞文件蒐集：

本研究以使用者瀏覽國防部青年日報社(http://news.gpwb.gov.tw/)的軍事新聞網頁(如圖 3-3 所示)為資料來源，實驗對象係挑選資訊相關科系及工作的使用者，依個人喜好選擇感興趣的軍事新聞閱覽，並利用 IEHistoryView 軟體取得一個月的使用者瀏覽紀錄(如圖 3-4 所示)，從中擷取青年日報的網頁內容，以進行

必要的字詞處理。

圖 3-3 國防部青年日報社
(資料來源：青年日報網站)

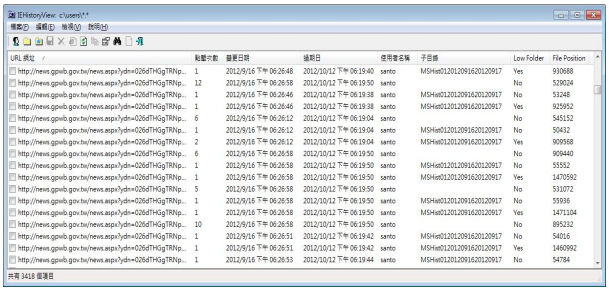


圖 3-4 IEHistoryView 瀏覽紀錄軟體畫面
(資料來源：IEHistoryView 軟體)

二、去除文件格式：

將蒐集到的資料去除其資料格式，例如版面格式、圖片及影像等，本階段產出為純文字組成的資料格式，並建置瀏覽紀錄資料庫儲存。

3.3.2 中文斷詞處理

一、斷字斷詞：

目前針對中文斷詞處理，國內研究多利用中央研究院跨所成立的詞庫小組(Chinese Knowledge Information Processing group, CKIP)所開發之 CKIP 中文斷詞系統，進行字詞的辨識與詞類標記，本研究亦採該系統，將擷取之文件內容實施中文字詞的辨識與詞類標記。表 3-1 為一軍事新聞經中文斷詞系統後之分詞與詞性標記範例。。

表 3-1 中文斷詞與詞性標記(範例)

斷詞前	海軍司令董翔龍上將昨日主持羅正東上校、郭天勇上校等二員屆退同仁頒獎典禮，表彰他們於服務軍旅期間，對海軍的卓越貢獻。
斷詞後	海軍司令(N) 董翔龍(N) 上將(N) 昨日(N) 主持(Vt) 羅正東(N) 上校(N)、(PAUSECATEGORY) 郭天勇(N) 上校(N) 等(POST) 二(DET) 員(M) 屆退同仁(N) 頒獎典禮(N) ，(COMMACATEGORY) 表彰(Vt) 他們(N) 於(P) 服務(Nv) 軍旅(N) 期間(N) ，(COMMACATEGORY) 對(P) 海軍(N) 的(T) 卓越(Vi) 貢獻(N)。

資料來源：本研究整理

二、詞性選擇：

文件特徵為文件之代表，文件特徵建構常見方法為關鍵字選取。藉由數個關鍵字代表文件，使用者可比對關鍵字檢索文件。由於每個新聞事件可包含數十到數百個詞彙，為了保持良好的系統效能並兼顧良好的分類效能，不可能把全部的內容都拿來做為特徵值。因此，必須在數量繁多的字詞中，選出鑑別力較高者來做為特徵值。

一般而言，文件中並非所有關鍵詞都具有相同重要性，通常名詞和動詞其重要性比起冠詞、介系詞等要重要得多。因此，本研究在詞性選擇便是以名詞和動詞為主。

三、去除無意義字詞：

依據 Ricardo and Berthier [23]研究所示，在一份文件中，有超過 80%的字詞是不具任何意義的。本研究保留前述詞性類別為名詞和動詞之字詞，餘字詞則予以刪除，以減少儲存空間及降低計算的複雜度。

3.3.3 同義字詞編修

本研究整合陳道輝 [2]、陳良駒、陳日鑫[3]與鄭百勝 [5]等字串篩選方法與「同義關鍵詞」、「相似關鍵詞」等修整規則，建立關鍵詞彙整及編修準則，其規則簡單說明如下：

- 規則一：意涵相同但字義表現不同者，以出現頻率較高者(或以統一的詞彙)為分析關鍵詞。例如：「孫中山」、「中華民國國父」、「孫文」、「中山樵」等，均以「中華民國國父」為替代關鍵詞。
- 規則二：中英文關鍵詞予以統一律定。例如：「IDF」與「經國號戰機」、「UAV」與「無人飛行載具」、「E-2T」與「空中預警機」等均將以中文字代稱。
- 規則三：部分過於普通且與新聞關連性較弱

之名詞將予以刪除不列入分析。如「每次」、「同時」、「因為」、「尤其」、「直接」等。

此外，本研究參考詞庫斷詞法，經由關鍵詞彙整及編修準則，擷取出有意義的名詞和動詞，利用中國大陸梅家駒等編著之「同義詞林」乙書、教育部辭典、中研院知網等彙整同義字詞成為同義字詞詞庫。

3.3.4 模糊概念關係識別

誠如前述，字詞可視為是使用者描述紀錄的基礎，在完成對使用者瀏覽紀錄的文字處理後，必須進一步表現出概念間的關係，而概念間的關係可由概念本身的一般性與特殊性決定，Sanderson and Croft [26]提出以字詞間共現的條件機率來建構概念的階層關係，本研究亦以字詞共同出現於新聞/文件集(以下以文件集統稱)的頻率，作為識別一般性與特殊性關係的基礎，即如果出現字詞 y 的文件集為出現字詞 x 文件集的子集，則稱 x 比 y 更具一般性或 y 比 x 更具特殊性。然而，為避免經常出現於各件中的連接詞、語助詞、介詞、……等等，較不具代表特定概念的字詞影響一般性與特殊性關係的判斷，在此僅於中研院研發的 CKIP 中文斷詞系統識別出的詞類標註中，擷取名詞、動詞作為概念字詞。

概念字詞常涉及同義或近義字詞，本研究將概念字詞 y 與 x 語意的相似程度定義如下：

$$\mu_{C_x}(y)=\frac{\left|A_{C_x}\cap A_{C_y}\right|}{\left|A_{C_x}\right|};\qquad (4)$$

其中，C_x 與 C_y 分別表示用以描述 x 與 y 語意的所有限制式集合；A_{C_x} 與 A_{C_y} 則各別表示符合限制式 C_x 與 C_y 的所有特徵集合；$\left|A_{C_x}\cap A_{C_y}\right|$ 為 C_x 與 C_y 共同的特徵數，而 $\left|A_{C_x}\right|$ 為 C_x 的所有特徵數量。

當 $\mu_{C_x}(y)=1$ 時，表字詞 y 與 x 為同義詞；當 $\mu_{C_x}(y)\geq\beta$ 且 $\beta\in[0.75,1)$ 時，則表字詞 y 與 x 為近義詞且其相似度為 β。

描述特定概念之代表性字詞萃取係由該概念同義字詞中詞頻最高者選出，意即取同義字詞中於所有文件出現頻率最高者為該概念的代表字詞，除概念代表字詞外之同義字詞與其近義詞則被視為該概念字詞的屬性，而這種概念與屬性之從屬關係，對應於模糊限制的知識表達即為變數與值域的關係。

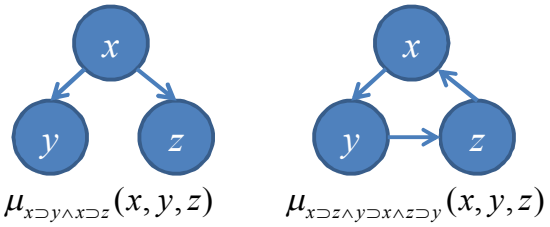
使用者描述紀錄涉及概念字詞及其關係的構成，由於自然語言在表達上的不確定與不精確性，概念字詞間的關係就不適以二元關係來表示，故本研究定義概念字詞間模糊概念關係如下：

定義 5：模糊概念關係(Fuzzy Concept Relation)
－給定概念字詞集 **W**，令任二個概念字詞 $x,y\in\mathbf{W}$，若符合以下模糊限制式即稱 x 比 y 更具一般性，以 $\mu_{x\supset y}(x,y)$ 或 $\mu_{y\subset x}(x,y)$ 表示之。

$$\mu_{x|y}(x,y)\geq\alpha \quad\text{and}\quad \mu_{y|x}(x,y)<\alpha;\qquad (5)$$

其中，$\mu_{x|y}(x,y)$ 代表一模糊關係，即字詞 y 出現情況下，字詞 x 出現的可能性；$\mu_{y|x}(x,y)$ 則表示字詞 x 出現情況下，字詞 y 出現的可能性；α 為一常數且 $\alpha\in[0.5,1]$。

若以三元的模糊概念關係圖為例(如圖 3-5 所示)，圖 3-5 (a)及 3-5 (b)可分別以模糊概念關係 $\mu_{x\supset y\wedge x\supset z}(x,y,z)$ 與 $\mu_{x\supset z\wedge y\supset x\wedge z\supset y}(x,y,z)$ 表示之。



$\mu_{x\supset y\wedge x\supset z}(x,y,z)$ $\mu_{x\supset z\wedge y\supset x\wedge z\supset y}(x,y,z)$

(a) (b)

圖 3-5 三元模糊概念關係圖
(資料來源：本研究整理)

依方程式(5)賡續計算各概念字詞間的模糊關係，即可構建出一具階層性之使用者描述紀錄。圖 3-6 顯示概念字詞間模糊概念關係，位於關係圖愈上方位置的字詞愈具一般性；反之，則愈具特殊性。

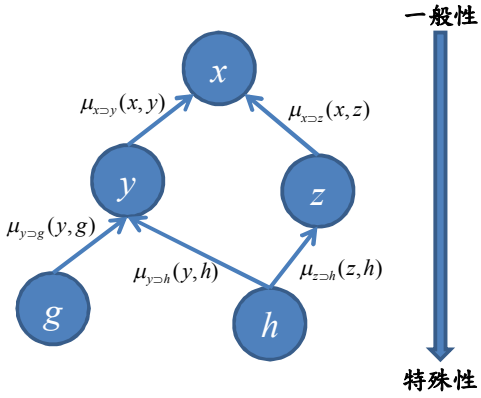


圖 3-6 模糊概念關係示意圖
(資料來源：本研究整理)

3.3.5 模糊限制網路建構

在本研究中，將模糊概念關係的計算即為一模糊限制滿足問題，模糊概念關係比對係以各文件經中文斷詞處理與同義字詞編修後的字詞為基礎，以 Tokunaga and Iwayyama [30] 提出的字詞權重計算方式，針對字詞進行加權式反向文件頻率的運算，以凸顯字詞於各文件的關連性。雖然，有部份研究[15, 25]採用字頻/反向文件頻率(Term Frequency/ Inverse Document Frequency, TF/IDF)的字詞權重計算方式，但依據反向文件頻率的定義，只在乎字詞出現的文件總數，不論文件出現該字詞次數，會出現特徵分佈不合理問題。因為不管字詞出現在文件中幾次，結果所得到 IDF 值都是 1，顯然並不合理。因此，本研究採用 WIDF 的字詞權重計算方式來改善這個問題，其公式如下：

$$WIDF(d,t)=\frac{TF(d,t)}{\sum_{i\in \mathbf{D}} TF(i,t)}; \qquad (6)$$

其中，$TF(d,t)$ 為字詞 t 在 d 文件中出現的頻率，而 i 表 **D** 文件集合範圍內的所有文件。

完成 WIDF 權重值的計算後，將各字詞之 WIDF 權重值視為該字詞歸屬於該文件之歸屬程度值，隨即從各文件先行挑選出 WIDF 權重值(或稱歸屬程度值)最高之字詞與其他文件中 WIDF 權重值最高之字詞，進行模糊概念關係的計算，關係計算完成後，各文件依續挑選出 WIDF 權重值次高之字詞與其他文件中 WIDF 權重值次高之字詞，再進行模糊概念關係的計算，依此類推，藉由模糊概念關係的計算建構出具階層性之模糊限制網路(或稱使用者描述紀錄)。惟並非所有在文件中的字詞均要參予模糊概念關係的計算，使用者可以自訂一結束關係計算的門檻值 δ，只要字詞的 WIDF 權重值低於 δ 值即不納入模糊概念關係的計算。

舉例而言，假設在文件 1 中所涵蓋的字詞集合為 $Doc_1=\{w_{11}@0.9, w_{12}@0.7,...,w_{1i}@0.1\}$，表字詞 w_{11} 的 WIDF 權重值為 0.9，w_{12} 的 WIDF 權重值為 0.7，w_{1i} 的 WIDF 權重值為 0.1，而文件 2 與文件 3 所涵蓋的字詞集合分別為 $Doc_2=\{w_{21}@0.92,w_{22}@0.8,...,w_{2j}@0.15\}$ 與 $Doc_3=\{w_{31}@0.85, w_{32}@0.74,...,w_{3k}@0.12\}$，則各回合字詞模糊概念關係計算分述如后：

第一回合

- 各文件先挑選出權重值大於 0.9 之字詞：僅文件 1 之 $w_{11}@0.9$ 與文件 2 之 $w_{21}@0.92$ 符合條件。
- 模糊概念關係計算：若字詞 w_{21} 出現的情況下，字詞 w_{11} 也出現的頻率高於 0.85；反之，字詞 w_{11} 出現的情況下，字詞 w_{21} 也出現的頻率小於 0.85，即 $\mu_{w_{11}|w_{21}}(w_{11},w_{21})\geq 0.85$ 且 $\mu_{w_{21}|w_{11}}(w_{11},w_{21})=0.62<0.85$，即字詞 w_{11} 較字詞 w_{21} 具一般性，字詞 w_{21} 較字詞 w_{11} 具特殊性，其階層關係如圖 3-7 所示。

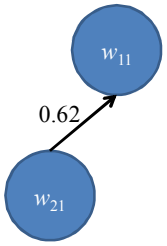


圖 3-7 第一回合模糊概念關係圖
(資料來源：本研究整理)

第二回合

- 各文件挑選出權重值大於 0.8 之字詞：由於文件 1 之 $w_{11}@0.9$ 與文件 2 之 $w_{21}@0.92$ 已於第一回合計算過，本回合僅挑選出文件 2 之 $w_{22}@0.8$ 與文件 3 之 $w_{31}@0.85$ 進行計算。
- 模糊概念關係計算：此回合必須計算的字詞關係包括：(w_{11}, w_{22}), (w_{11}, w_{31}), (w_{21}, w_{22}), (w_{21}, w_{31}), (w_{22}, w_{31})，假設字詞 w_{31} 僅與字詞 w_{11} 有關聯，字詞 w_{22} 分別與字詞 w_{11}、字詞 w_{21} 有關聯，計算結果如圖 3-8 所示。

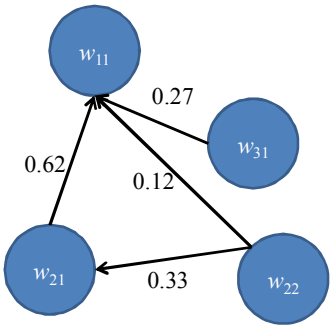


圖 3-8 第二回合模糊概念關係圖
(資料來源：本研究整理)

第三回合

- 若用以計算字詞間模糊概念關係的最低滿意的權重值為 0.8，則停止字詞間模糊概念關係的計算。
- 若用以計算字詞間模糊概念關係的最低滿意的權重值為δ，則繼續字詞間模糊概念關係的計算，直至字詞權重值δ(含)以上字詞之模糊概念關係計算完後停止，最後所得之模糊概念關係圖即為一模糊限制網路圖或稱使用者描述紀錄。

在計算過程中權重值由高而低進行模糊概念關係的計算，即為在模糊限制滿足問題中，將模糊限制的條件逐次放鬆(意指模糊限制式的滿意度由高逐漸降低)求解，當所得之解已足以滿足使用者需求(在此意指模糊概念關係建構而成的模糊限制網路已足以表達當前之使用者描述紀錄)，即停此模糊限制的條件放鬆；而因為模糊限制條件係採滿意度由高而低的方式求解，因此，在停此模糊限制條件的放鬆後，所得之解即為符合所有模糊限制式且滿意度最高的解。這種漸進式的解題方式較傳統將所有概念字詞權重值視為一致，並進行所有字詞關係識別的限制解題方式更有效率。

3.3.6 模糊限制網路裁剪

由概念字詞及其關係建構而成的模糊限制網路，仍需予以適當的裁剪以辨識出概念類別(即本研究中的興趣主題)，本研究以激發擴散(Spreading Activation)理論 Quillian [21, 22]為基礎，提出一關係強度激發擴散模式，作為模糊限制網路裁剪之依據。關係強度激發擴散模式主要係賦予在模糊限制網路中各字詞節點(Node)一激發值(Activation Value) $v \in [0,1]$，字詞間的連結(Link)則以模糊概念關係值為權重，進而計算該字詞節點所接收與傳送的關係強度，其定義如下：

定義 6：關係強度激發擴散－給定一模糊限制網路，令模糊限制網路圖中任意字詞節點 n 之當前激發值 R_n^c 均為 1，字詞節點對外之連結以字詞與連結字詞之模糊概念關係值為權重，則除終端字詞節點外之字詞節點其所接收與傳送的關係強度值如下：

$$R_k^r = 1 + \sum_{j \in N_k^i} R_{jk}^s ; \quad (7)$$

$$R_{kj}^s = \sum_{j \in N^o} R_k^c * \left(\frac{w_{kj}}{\sum_{j \in N_k^o} w_{kj}} \right); \quad (8)$$

R_k^c 代表字詞節點 k 的當前關係強度；w_{kj} 代表字詞節點 k 與 j 之間的模糊概念關係值；N_k^i 與 N_k^o 分別表示被包含於字詞節點 k (即輸入連結 Inlink)與包含字詞節點 k (即輸出連結 Outlink)之字詞節點集合；R_k^r 與 R_{kj}^s 則代表字詞節點 k 所接收與傳送到字詞節點 j 的關係強度值。

在關係強度激發擴散模式中，關係強度係依照模糊概念關係值大小的順序傳送，先從模糊概念關係值最小的被激發字詞節點開始，再依序往模糊概念關係值大的字詞節點處理。假設被激發字詞節點 i 與另一個模糊概念關係值較大的被激發字詞節點 j 連接，關係強度 R_{ij}^s 就會透過兩字詞節點彼此之間的連結從字詞節點 i 往字詞節點 j 傳送。

關係強度激發擴散的計算，以圖 3-9 為例，令終端字詞節點 g 與 h 其激發值為 1，因字詞節點 h 的輸出連結分別指向字詞節點 y 及 z，其激發值會依與字詞節點 y 及 z 模糊概念關係的權重值 w_{hy} 與 w_{hz} 予以分配其激發值，故字詞節點 h 傳送至字詞節點 y 的關係強度激發值為 $1 * \frac{w_{hy}}{w_{hy} + w_{hz}}$，而字詞節點 y 收到字詞節點 g 與 h 的關係強度激發值後，其所得之當前激發值即為 $1 + \left(1 * \frac{w_{hy}}{w_{hy} + w_{hz}} \right) + \left(1 * w_{gy} \right)$。

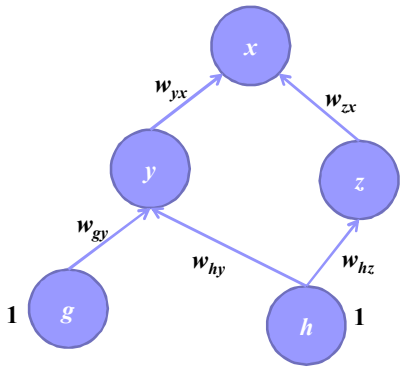


圖 3-9 模糊限制網路示意圖
(資料來源：本研究整理)

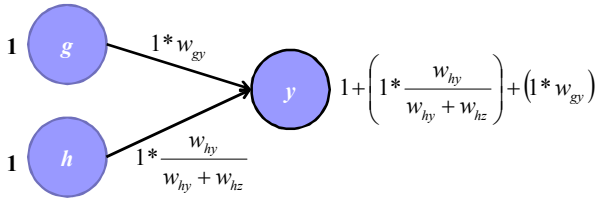


圖 3-10 關係強度激發擴散值計算示意圖
(資料來源：本研究整理)

模糊限制網路裁剪之步驟分述如下：

- 步驟一：針對模糊限制網路進行字詞間關係強度激發擴散的關係強度激發值傳遞。
- 步驟二：當模糊限制網路之關係強度激發值達停止傳遞的穩態時，設定門檻值 $\gamma \in R^+$，將小於此門檻值 γ 的字詞從模糊限制網路裁剪。

3.3.7 興趣主題網路萃取

代表使用者描述紀錄之模糊限制網路，經關係強度激發擴散及將關係強度激發值低於門檻值 γ 之字詞裁剪後，所識別出之概念類別字詞及其包含之字詞集合，因均屬同一概念類別，故由該概念類別字詞所構成之模糊限制網路，在此，稱之為興趣主題網路，定義如下：

定義 7：興趣主題網路(Interesting Subject Network; ISN)－給定一個裁剪過之模糊限制網路 $\widetilde{\Omega}=(\mathbf{U}, \mathbf{X}, \mathbf{C})$ 與概念 s' 語意屬性集合 $\mathbf{U}_{s'} \in \mathbf{U}$，則概念 s' 之興趣主題網路定義為 $\widetilde{\Omega}_{s'} = (\mathbf{U}_{s'}, \mathbf{X}_{s'}, \mathbf{C}_{s'})$ 且 $\arg_{i=1..m} \mathrm{U}_i \bigcap \mathrm{U}_{s'} \neq \phi$，意即興趣主題網路為與概念 s' 語意屬性集合 $\mathrm{U}_{s'}$ 連結之關係網路，亦為模糊限制網路的子網路。

興趣主題網路可以從模糊限制網路中，以關係強度激發值傳遞終止的概念字詞節點為根節點，以中序搜尋法找出其所屬之概念子節點及其模糊概念關係組成的子網路即得；而停止關係強度激發值傳遞的概念字詞節點即為興趣主題。以圖 3-11 為例，概念字詞節點 *Concept*1 與 *Concept*2 即可分別萃取出成為該興趣主題網路之興趣主題。

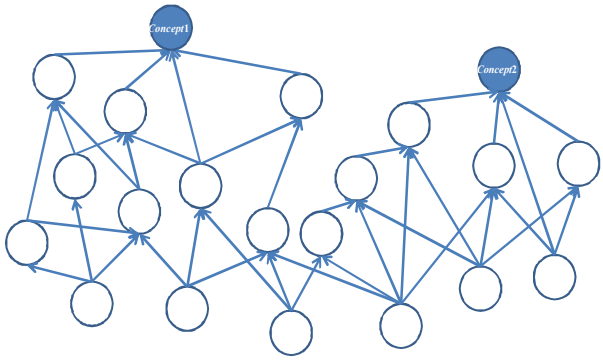


圖 3-11 興趣主題網路示意圖
(資料來源：本研究整理)

3.3.8 資訊過濾

完成興趣主題之識別後，利用下列步驟來進行資訊過濾，增加使用者感興趣的回傳結果，步驟如下：

- 步驟一：根據使用者輸入的關鍵字，搜尋使用者描述紀錄中是否存在相同的字詞，如果是，進行步驟二；反之，則直接利用關鍵字進行搜尋。
- 步驟二：將興趣主題網路中的字詞與使用者所輸入關鍵字進行比對，擷取興趣主題網路中相關主題及主題內的字詞。
- 步驟三：將使用者輸入的關鍵字與興趣主題網路擷取出的字詞，從資料集中找出該字詞所屬的新聞文件集。

在完成依個人喜好進行的搜尋後，得到的文件會導致回傳大量的相關文件，仍無法有效解決資訊過濾的問題，本研究針對回傳結果暫依照文件相關度分數[6]來進行排序，將使用者感興趣的文件排序在前面，減少使用者篩選文件的負擔，後續視成效再另擇其他研究方法針

對文件與予過濾及排序。文件相關度分數計算步驟如下：

●步驟一：針對查詢後得到的文件計算其文件相關度分數，S_D 的計算方法是以被激發字詞的最後能量為基礎來進行計算，公式如下：

$$S_D = \frac{\sum_{i \in A} w_i * E_i^f}{\log(NT)} * \log\left(1 + \frac{b+d}{b}\right); \quad (9)$$

其中 A 代表被激發字詞的集合;NT 代表文件中所有字詞的總數；w_i 代表字詞 t_i 的權重值。那些沒有傳送任何能量的字詞，稱為主導字詞(Dominate Term)。主導字詞的數量 b 用來衡量文件的廣度；文件的大小 d 為有傳送能量的被激發字詞數量，即除了主導字詞以外的其他被激發字詞，也就是 b 與 d 的總合會等於被激發字詞的數量。因子 $\log\left(1+\frac{b+d}{b}\right)$ 用在有傳送能量的被激發字詞數量多且主導字詞數量少的文件，由於本研究以新聞網頁做實驗，字詞數量多且主題較明確，故符合此特性。

●步驟二：設定一個文件相關度分數的門檻值，進行資訊過濾。

●步驟三：最後依文件相關度分數將搜尋結果由高至低重新排序後，推薦給使用者，使用者越有興趣的文件，將會排序在越前面。

4. 方法評估

本研究利用模糊限制網路來建構具多興趣主題的單一使用者描述紀錄，以取代傳統僅能描述二元關係的使用者描述紀錄，並透過本研究所提關係強度激發擴散模式分別建構成字詞間的階層關係與找出使用者紀錄檔中足以代表使用者感興趣的主題，最後再透過學習後的使用者描述紀錄來過濾及重新排序龐大的回傳結果，以下就方法論進行定性評估：

一、人類思考模式的符合性

透過模糊理論將使用者的興趣由傳統的相關與不相關、0 與 1 的觀念泛化到相關與不相關之間、0 與 1 之間，並透過歸屬函數將抽象的喜好予以量化，藉由歸屬函數值來呈現使用者的喜好程度，以期更符合人類的思考模式。

二、知識表達的簡潔性

許多研究定義的概念間模糊關係仍僅限於二元關係的表示，若文件集有 n 個概念字詞，則最多需要 $n(n\text{-}1)$個關係或限制式來表達字詞間的關係。

具多元關係表示特性的限制，最少僅需 1 條限制式(或單一限制網路)即可表達字詞關係；而限制係模糊限制的特例，故模糊限制亦具有多元關係的表示特性，其最少亦僅需 1 條模糊限制式(或單一模糊限制網路)即可表達字詞關係。

三、解題方法的效率性

由於使用者描述紀錄的建構時，文件集中每個概念字詞須與其他字詞逐一比較其階層關係，在傳統的限制式解題方法中，欲識別出 n 個概念字詞間階層關係的計算次數為 $n(n\text{-}1)$，而其最差、最佳及平均的時間複雜度均為 $\mathbf{O}(n^2)$ 。

而本研究所提模糊限制特有的漸進式解題方式，此演算法最差、最佳及平均的時間複雜度雖與傳統限制式的解題方法同為 $\mathbf{O}(n^2)$，但利用解模糊限制的方式，先將字詞滿意度由高至低的遞減方式排序後，透過使用者自訂的滿意度門檻值來界定運算範圍，將門檻值以上的概念字詞由高至低地代入計算字詞之階層關係，故令大於及等於門檻值的概念字詞數為 j $(2 \leqq j \leqq n)$，則 j 個概念字詞間階層關係的計算次數為

$$\sum_{i=2}^{j} i\,(i-1);$$

此方法最佳情況只需計算 2 次即可得到階層關係，最差則同為 $n(n\text{-}1)$次(即當滿意度門檻值為 0 時)；而平均計算次數為

$$\frac{1}{j-1} \times \sum_{i=2}^{j} i\,(i-1);$$

因 $2 \leqq j \leqq n$，故

$$\frac{1}{j-1} \times \sum_{i=2}^{j} i\,(i-1) \leq \frac{1}{n-1} \times \sum_{i=2}^{n} n\,(n-1);$$

即導入的模糊限制式其平均計算次數會小於傳統限制式的平均計算次數。由此可知，

本研究所導入的模糊限制式能降低運算的次數，提升解題的效率，減少運算資源的浪費。

5. 結論

因應使用者在巨量網路資源中的資訊過濾需求，本研究將模糊限制式的方法導入使用者描述紀錄的建構與學習中，在知識表達方面，以模糊限制式構連而成的模糊限制網路來呈現字詞語意的不明確性及概念間的多元關係，以及表示具多重興趣主題之單一使用者描述紀錄；在解題方面，將使用者描述紀錄的建構視為模糊限制滿足問題，再藉由關係強度激發擴散模式來萃取出多重興趣的主題，並按照文件相關度分數作為資訊過濾的依據，呈現於回傳結果中，滿足使用者的檢索需求。相較於其他文獻，本研究提供以下重要觀點：

- 不確定性或不精確性的知識表達：導入模糊邏輯的觀念，以模糊邏輯來描述現實生活中事物的等級，以彌補傳統布林邏輯無法描述對事物具不明確邊界定義的缺點。

- 概念字詞間多元關係的模塑：概念字詞間的關係不儘然全是二元關係，雖然也可以二元關係進行關係的描述，但以模糊限制為基礎的多元關係表達，不僅可以更簡潔地進行知識表達，也更具關係的可讀性。

- 內隱式使用者描述紀錄的學習：傳統以使用者回饋為基礎的研究中，因為進行個人化資訊登錄、回饋或是評價是耗時的，所以通常使用者不太願意提供評價，也造成推薦的準確率降低。本研究以內隱式的方式進行使用者描述紀錄的學習，減少使用者涉入學習過程，避免耗時的使用者資料登錄、回饋或評價。

- 漸進式解題效率的提昇：本研究將使用者描述紀錄視為模糊限制網路，以解模糊限制問題的方式進行解題，提供一時間複雜度最差、最佳及平均情況均為 $\mathbf{O}(n^2)$，但在計算次數上優於傳統限制式的解題方法，當概念字詞 n 值愈大時，將讓使用者描述紀錄建構的時間更有效率。

- 檢索效益的提昇：針對回傳結果依照文件相關度分數來進行排序，將使用者感興趣的網頁，優先呈現給使用者瀏覽，減少使用者篩選文件的負擔，有效提昇使用者的檢索效益。

- 知識表達與解題方法的統一：本研究以模糊限制作為知識表達之基礎，也利用模糊限制本身的解題特性，提供漸進式之解題方法，減少計算的複雜度，而無須援用其他理論或方法論，有效降低學習的複雜度。

檢視以上定性的方法評估，本研究所提出導入模糊限制為基礎的使用者描述紀錄建構方法，確能改善傳統使用者描述紀錄學習與知識表達的不足，提高檢索效率與效益，協助使用者過濾資訊，滿足使用者的資訊需求；惟目前在解題效率上仍有強化及改善空間，將持續修訂並擴大研究的應用面與實用性。

參考文獻

[1] 許孟淵，以本體論為基礎之新聞事件檢索與瀏覽，***國立雲林科技大學資訊管理系碩士論文***，2006。

[2] 陳道輝，“利用學位論文資訊萃取資訊領域相關研究主題關聯性”，***國立中山大學資管所碩士論文***，2003。

[3] 陳良駒、陳日鑫，“植基於詞彙數量關係探討軍事新聞主題-以青年日報為例”，***資訊管理展望***，第十二券．第一期，2010。

[4] 陳志銘、張美華、邱偉嘉，“基於自動查詢語句擴展之主題地圖智慧型新聞搜尋引擎”，***圖書館學與資訊科學***，第三十四券．第二期，pp.19-41，2008。

[5] 鄭百勝，“應用關聯規則建構具方向性之領域知識結構圖－以資訊管理領域為例”，***國立中山大學資管所碩士論文***，2006。

[6] 楊宗翰，“以單一使用者興趣檔為基礎的查詢擴展與文件重排序系統”，***國立中央大學資訊管理系所碩士論文***，2010。

[7] 吳典恩，“結合本體論以及關聯法則於查詢擴展之研究”，***國立成功大學資訊管理研究所碩士論文***，2006。

[8] Bowen, J., Lai, R., and Bahler, D., “Lexical Imprecision in Fuzzy Constraint Networks,” ***Proc. AAAI-92,*** San Jose, Calif, pp. 616-621, 1992.

[9] Cooley, R., Mobasher, B., and Srivastava, J.,

"Data Preparation for Mining World Wide Web Browsing Patterns," ***Knowledge and Information Systems***, Vol. 1, No. 1, pp. 5-32, 1999.

[10] Davis, E., "Constraint Propagation with Interval Labels," ***Artificial Intelligence***, Vol. 32, No. 3, pp. 281-331, 1987.

[11] Fu, X., Budzik, J., and Hammond, K. J., "Mining Navigation History for Recommendation," ***Proc. of the 5th International Conference on Intelligent User Interfaces (IUI 2000)***, pp. 106-112, 2000.

[12] Gruber, T. R., "A Translation Approach to Portable Ontology Specifications," ***Knowledge Acquisition***, Vol. 5, No. 2, pp. 199-220, 1993.

[13] Lai, Kuo-Robert., ***Fuzzy Constraint Processing***, Ph.D. thesis. NCSU, Raleigh, N. C, 1992.

[14] Lai, Kuo-Robert., Lin, Menq-Wen., and Yu, Ting-Jung., "Learning Opponent's Beliefs via Fuzzy Constraint-Directed Approach to Make Effective Agent Negotiation," ***Applied Intelligence***, Vol. 33, No. 2, pp. 232-246, 2010.

[15] Lang, K., "NEWSWEEDER: Learning to Filter Netnews," ***Proc. of 12th International Conference on Machine Learning*** (Lake Tahoe, CA, 1995), pp. 331-339, 1995.

[16] Lin, Menq-Wen., Lai, Kuo-Robert., and Yu, Ting-Jung., "Fuzzy Constraint-based Agent Negotiation," ***Journal of Computer Science and Technology***, Vol. 20, No. 3, pp. 319-330, 2005.

[17] Mackworth, A. K., "Consistency in networks of relations," ***Artificial Intelligence***, Vol. 8, No. 1, pp. 99-118, 1977.

[18] Mezghani, Manel., Zayani, Corinne Amel., Amous, Ikram. and Gargouri,Faïez., "A user profile modelling using social annotations: a survey," ***WWW Companion***, Vol ACM, pp. 969-976, 2012.

[19] Nanas, N., Uren, V., de Roeck, A., and Domingue, J., "Multi-topic Information Filtering with a Single User Profile," ***Methods and Applications of Artificial Intelligence***, pp. 400-409, 2004.

[20] Pretschner, A., and Gauch, S., "Ontology Based Personalized Search," Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, ***IEEE Computer Society***, p. 391, 1999.

[21] Quillian, M. R., "A Revised Design for an Understanding Machine," ***Mechanical Translation***, Vol.7, No. 1, pp. 17-29, 1962.

[22] Quillian, M. R., "Semantic Memory," ***Unpublished Doctoral Dissertation, Carnegie Institute of Technology***, 1966. (Reprinted in part in M. Minsky [Ed.], Semantic information processing. Cambridge, Mass.: M.I.T. Press, 1968.)

[23] Ricardo, Baeza-Yates., and Berthier, Ribeiro-Neto., "Modern Information Retrieval," ***ACM Press***, 1999.

[24] Ruttkay, Z., "Fuzzy Constraint Satisfaction," ***Proceedngs of 3rd IEEE Intern. Conf. on Fuzzy Systems***, Vol. 3, pp. 1263-1268, 1994.

[25] Salton, G., and McGill, M. J., ***Introduction to Modern Information Retrieval***, New York: McGraw-Hill Co., 1983.

[26] Sanderson, Mark., and Croft, Bruce., "Deriving Concept Hierarchies from Text," ***In SIGIR***, pp. 206-213, 1999.

[27] Solskinnsbakk, Geir., and Gulla, Jon Atle., "Combining Ontological Profiles with Context in Information Retrieval," ***Data & Knowledge Engineering***, Vol. 69, No. 3, pp. 251-260, 2010.

[28] Spiliopoulou, M., and Faulstich, L., "WUM–A Tool for WWW Utilization Analysis," ***Proc. of the International Workshop on the World Wide Web and Databases (WebDB'98)***, pp. 184-203, 1998.

[29] Sugiyama, K., Hatano, K., and Yoshikawa, M., "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users," ***Proceedings of the 13th international conference on World Wide Web***, pp. 675 - 684, 2004.

[30] Tokunaga, Takenobu., and Iwayama, Makoto., "Text Categorization based on Weighted Inverse Document Frequency," ***Technical Report of Tokyo Institute of Technology***, March, 1994.

[31] Wang, J., Chen, Z., Tao, L., Ma, W. Y., and Wenyin, L., "Ranking User's Relevance to a Topic through Link Analysis on Web Logs," ***Proc. of the 4th ACM CIKM International Workshop on Web Information and Data Management (WIDM'02)***, pp. 49-54, 2002.

[32] Zadeh, L. A., "Fuzzy Sets," ***Information and Control***, Vol. 8, pp. 338-353, 1965.

[33] Zhu, Z., Xu, J., Ren, X., Tian, Y., and Li, L.,

"Query Expansion Based on a Personalized Web Search Model," Proceedings of the ***Third International Conference on Semantics***, Knowledge and Grid, IEEE Computer Society, pp. 128-133, 2007.

[34] 國防部，青年日報社，參見國防部網站 http://news.gpwb.gov.tw/ [visited in 2012/09/12]。