圖書資訊學研究19：1 (December 2024)：37-84
DOI:10.30177/JLISR.202412_19(1).0002

圖書館參考諮詢服務機器人讀者問題意圖分析之研究

Library Reference Service Chatbots for Library FAQ Intent Analysis

陳舜德*

Shun-Der Chen

輔仁大學圖書資訊學系副教授
Associate Professor
Department of Library and Information Science
Fu Jen Catholic University

黃丰嘉

Feng-Chia Huang

輔仁大學圖書資訊學系研究生
Graduate Student
Department of Library and Information Science
Fu Jen Catholic University

【摘要 Abstract】

參考諮詢是圖書館提供讀者的重要服務，服務模式也從傳統的臨櫃諮詢，逐步轉變為線上同步或非同步模式，甚至發展到跨館合作線上聯合參考諮詢。隨著人工智慧、深度學習和自然語言技術的長足進步，對話機器人技術也漸臻成熟。許多即時互動軟體平臺紛紛釋出應用程式介面（application programming interface, API），方便開發者介接自家平臺提供對話諮詢服務。然而，在實際應用中，一些Chatbot由於無法準確辨識使用者的詢問意圖

*通訊作者：陳舜德 ryanchen@lins.fju.edu.tw
投稿日期：2023年12月27日；接受日期：2024年3月6日

（intent），因此藉由限制使用者提問的方式來輔助系統正確理解使用者的詢問目的，從而搜尋對應的回覆結果。本文希望從蒐錄參考諮詢問答集語料進行訓練學習，擷取詞彙特徵重新建立群集，藉此能更正確的分析識別出讀者詢問意圖，未來可以提供讀者一個更精準預測詢問意圖與智慧化的參考諮詢機器人。

Reference consultation is an important service provided by libraries to their readers, and the service mode has gradually changed from traditional counter consultation to online synchronous or non-synchronous mode, and even developed to cross-library cooperation for online joint reference consultation. With the great progress of artificial intelligence, deep learning and natural language technology, the conversation robot technology is also getting mature. Many real-time interactive software platforms have released Application Programming Interfaces (APIs) to facilitate developers to connect to their own platforms to provide conversational consultation services. However, in practice, some Chatbots are unable to accurately recognize the user's intent, so they restrict the user's questions to help the system understand the purpose of the user's query and search for the corresponding response results. This paper endeavors to address this challenge by amassing a corpus of reference consulting question and answer sets for comprehensive training and learning purposes. By extracting vocabulary features and re-establishing clusters, our aim is to refine the system's ability to analyze and recognize the intent behind reader queries accurately. Ultimately, this will empower us to furnish readers with a more precise prediction of their query intent, paving the way for the development of an intelligent reference consultation robot poised to cater adeptly to their informational needs in the future.

【關鍵詞 Keywords】

聊天機器人；基於密度的聚類演算法；意圖分析；自然語言處理；線上聯合參考諮詢

Chatbot; Density-Based Spatial Clustering of Applications with Noise (DBSCAN); Intent Analysis; Natural Language Processing; Online Joint Reference Consulting

壹、前言

參考諮詢服務一直是圖書館提供讀者的重要服務。從早期設置參考櫃臺等候服務讀者參考諮詢，但如此方式人力需求相當高。隨著網路的發展，讀者使用圖書館行為也逐漸改變，有些讀者也會透過網路線上進行問題諮詢。然而，要能快速且正確地回覆讀者提出的各類問題，對於參考館員的要求與訓練是一項巨大的挑戰。許多圖書館受限館員人力不足，只能提供非同步數位參考服務（或稱為虛擬數位服務），亦即讀者透過電子郵件或圖書館設置的參考諮詢網站提出問題，待參考館員有空閒時（甚至在下班後），再對讀者提出的問題進行解答（吳美美、龐宇珺，2011；柯皓仁，2004）。這樣可以讓讀者不受限於時空，透過網路進行參考諮詢。然而，這類服務仍然存在人力需求問題，因為館員無法24小時隨時支援解答讀者的問題（王愛珠，2007；蘇小鳳，2004）。因此，近年來圖書館也希望能利用對話機器人自動回覆功能，提供無時差的服務，協助讀者快速獲取所需資訊。

近年來隨著對話機器人的發展，其應用在各領域範疇已相當普遍，例如在COVID-19期間有些醫院將其應用於病人照護中，有效建立與病人的溝通管道，提供正確即時的醫療資訊，並提供24小時不間斷的醫療諮詢（李金霎、高淑容、劉翠瑤、朱怡蓁，2022）。或有將人工智慧機器人（Artificial Intelligence Robot）應用於旅遊服務產業，隨時與顧客進行即時互動交流，提供遊客需要的訊息。更可結合地方政府推動相關觀光政策和活動，也幫助商家瞭解遊客需求（林妤蓁、高東慶、李祐丞、林紋如、洪名呈，2020）。因此，對於圖書館而言，導入對話機器人來服務讀者是一個具有潛力的選擇，在館員人力不足下仍能提供讀者不受限於時空的參考諮詢服務。

許多企業早已預見聊天機器人能夠以較低成本執行許多勞力密集型的任務，因此在這項新興技術上投入巨資（Nagarhalli, Vaze, & Rana, 2020）。而近十年隨著人工智慧、機器學習、自然語言處理等技術的長足進步，也帶動智慧客服機器人的發展。自2016年起，聊天機器人（Chatbot）的風潮漸起，應用更是遍地開花，除了各大通訊軟體開發商（如LINE、Facebook）相繼釋出應用程式介面（application programming interface, API），方便開發者將自家產品服務串接至大眾常用的通訊軟體上之外，IBM Watson Assistant、Amazon Lex、Google Dialogflow、Microsoft LUIS等對話平臺的出現，讓一般人能借由此類平臺輕易建置簡

單互動對談模式的Chatbot，更促使Chatbot逐漸成為與使用者溝通互動的重要工具。

隨著Chatbot建置的門檻降低，使其更容易拓展到各行各業上，特別是應用於即時客戶服務方面，舉凡股市理財投資（Line @微股力ScanTrader）、旅遊搜尋推薦（Facebook Messenger @Skyscanner）、披薩訂購（Google Dialogflow @達美樂Domino）等領域應用皆能看見它的身影。然而，儘管當今科技巨頭所開發的Chatbot背後具有機器學習引擎支撐，能夠藉由語料的擴充來不斷地學習和改善，但因為其自然語言理解（natural language understanding, NLU）核心引擎仍是以英文為基礎，套用至中文對話的處理仍不盡理想。

而上述方式建置的Chatbot往往依賴人工進行對話訓練語料的標註，以便學習理解每個問題的詢問意圖（intent），造成建置上既耗時又費工。另外，假如是開發者自行開發設計的話，礙於技術的限制以及對話設計的方便性，容易變成只是在通訊軟體介面中提供各種「點擊」功能的Clickbot，或是僅由「關鍵字觸發」固定回應的Keywordbot，使得使用者「諮詢」的彈性大大降低，也失去作為聊天機器人應有的「對話」的功能。因此，本研究從中文自然語言處理著手，自圖書館常見問答集的問答語料中透過對字詞、句型、句法結構的分析來理解使用者的詢問目的（意圖），以此基礎來建構一個能理解讀者詢問意圖、模擬與人類對話的意圖式聊天機器人（intent-based Chatbot）。

貳、文獻探討

一、聊天機器人

「聊天機器人」一詞源於「chatterbot」，該術語最初由Michael Mauldin於1997年提出（Deryugina, 2010）。而後隨著聊天機器人開始投入到各式的商業應用賦予其不同的名稱，如：對話系統（dialogue system）、對話代理人（conversational agent）、虛擬助理（virtual assistant）和個人助理（personal assistant）等（Altinok, 2018）。至於如何判斷聊天機器人是否具備智慧，端看其能否通過「圖靈測試」（Turing test），一旦受試者無法清楚分辨機器與人類回答之間的區別，如此便能得知機器在某種程度上已經能媲美人類智慧了（Turing, 1950）。

根據Al-Zubaide與Issa（2011）對聊天機器人的分類，可分為Rule-Based Chatbot和AI-based Chatbot兩種類型。基於規則（rule-based）的聊

天機器人是藉由事先編寫的條件或規則來作為聊天機器人判斷應執行何種回覆動作的依據；而基於人工智慧（AI-based）的聊天機器人，亦可稱為基於意圖（intent-based）識別的聊天機器人，則是透過大量對話語料的訓練，以及機器學習、自然語言處理技術的輔助來達到模擬人類自我學習並獲得有效訊息的能力，從而識別使用者的詢問意圖並自動判斷應該回覆何種訊息（Rosruen & Samanchuen, 2018）。

二、基於意圖識別聊天機器人架構

一般而言，建置一個基於意圖識別的聊天機器人，整體架構如圖1所示。大致上，依據使用者訊息傳遞的處理順序需要自動語音辨識（automatic speech recognition, ASR）、NLU和自然語言生成（natural language generation, NLG）這三項核心元件。建置聊天機器人系統，首先需經過ASR將使用者的語音訊號轉換為文字輸入，使得使用者無須操作鍵盤輸入，便可立即對聊天機器人透過語音下達指令（諮詢需求）。接著，藉由NLU或自然語言處理技術處理將非結構化的文字資料轉換為系統可以理解和處理的形式，同時從中提取有意義的訊息以及需要進一步查詢的細節知識，方便後續藉由決策引擎（decision engine）判斷使用者的詢問意圖。而意圖判斷的結果，將被作為選擇查找何種「外部知識庫」（相當於圖1中Data Layer區塊）來源的依據，透過採取相應的檢索動作，求得回覆使用者的訊息內容。最後，在接收來自外部知識庫傳回

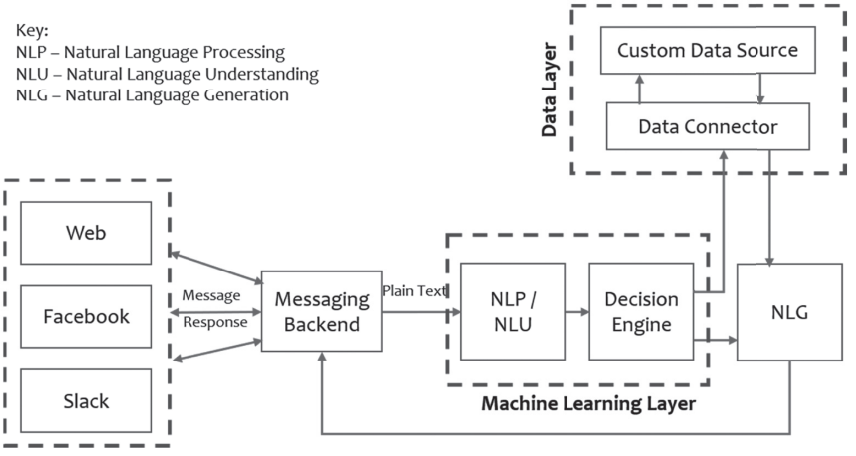

圖1　智慧型聊天機器人的整體架構

資料來源：Ayanouz, Abdelhakim, & Benhmed（2020）。

的結構化數據（即解答）後，NLG會進一步規劃內容，並生成最終給使用者的回應訊息。

三、圖書館參考諮詢機器人的今昔發展

在圖書館領域上，線上自動參考諮詢系統在圖書館應用上很早便被提出並嘗試建置（王明玲、杜立中、曾彩娥，2011；王愛珠，2007），希望透過線上系統對於讀者經常性、固定性的館務問題諮詢進行自動回覆，如此可降低參考館員於此非專業回覆問題的工作量，能更專注於回覆讀者專業參考諮詢問題。早期所建置的線上參考諮詢系統，作法類似專家知識系統，大多藉由關鍵語彙對應問答資料庫（Q&A Database）進行檢索查詢，再將預存於問答資料庫中對應的答案回覆給讀者，此種模式實作上較為簡單，但卻有著極為重大的缺點——僅能處理預設制式的問題。然而伴隨自然語言處理、語音辨識技術發展漸趨成熟，以及深度學習技術的突破，如同真人般對談的聊天機器人應運而生。近年來，隨著圖書館數位化的腳步，實際進館讀者人數雖有下降，但數位化讀者使用圖書館資源的人數與頻率並未減少。現今利用圖書館資源的行為與需求有著大幅改變，年輕世代讀者更習慣透過網路服務尋求「問題」解答，也觸發線上參考諮詢需求逐年提高（林靖文、謝寶煖、莊馥瑄，2013）；再者圖書館亦與各產業相同，均面臨營運經費與服務人力日益緊縮的問題，特別是在非公營或大學圖書館這樣的情形尤為嚴峻，在此現況與困境上更凸顯聊天機器人應用於圖書館參考諮詢服務的迫切性。

四、中文聊天機器人在圖書館應用與挑戰

隨著自然語言處理技術與相關輔助科技應用發展，國內、外大學圖書館陸續推出聊天機器人的參考諮詢服務應用。例如：2010年中國清華大學以A.L.I.C.E.（Artificial Linguistic Internet Computer Entity）開源軟體為基礎、XML為語料庫儲存格式，搭配搜尋引擎技術建構的「小圖」（姚飛、紀磊、張成昱、陳武，2011）；2014年美國加州大學爾灣分校（University of California, Irvine）使用開源軟體Program-O開發網頁版聊天機器人「ANTswers」，其藉由超文字預處理器（hypertext preprocessor, PHP）編寫的人工智慧標記語言（artificial intelligence markup language, AIML）直譯器透過模式匹配來回覆讀者（Kane, 2016）；2019年美國奧克拉荷馬大學（University of Oklahoma）則利用Ivy.ai對話平臺構建網頁版聊天機器人「Bizzy」（Young, 2019）等，皆獲得廣大且熱烈的迴響。反觀臺灣圖書館在聊天機器人研究成果也逐漸顯露，陳宜琳（2019）

開發一個使用「模式匹配」方式建構的國立臺灣師範大學圖書館（以下簡稱臺師大圖書館）參考諮詢機器人，范蔚敏（2020）則在Google DialogFlow對話平臺的基礎上建置一個國立臺灣大學圖書館（以下簡稱臺大圖書館）虛擬參考諮詢服務機器人。

針對中文聊天機器人在圖書館的應用，具代表性的有三。

（一）中國清華大學圖書館——「小圖」

2010年設計並以開源軟體A.L.I.C.E.為基礎建置的「小圖」，提供參考諮詢、圖書檢索、自我學習……等多種服務，其使用的關鍵技術包含中文的自然語言處理（如：語料庫建置分析、中文分詞、詞性標註），協助加快檢索效率與輸出相關回覆的搜尋引擎技術（如：倒排索引〔inverted index〕、TF-IDF（term frequency-inverse document frequency）演算法、相似度匹配演算法……等），還有客製化「Q：問題A：答案」教學命令的教學系統，以及利用AIML知識庫推理返回最適答案的知識推理機制。

「小圖」的使用流程如圖2所示：讀者輸入「問題」指令後，小圖會對「問題」進行中文斷詞，而後識別「問題」類別，並導引到相應的服務功能，最後才到各別獨立的參考紀錄語料庫、搜尋引擎、教學問答學習庫和AIML知識庫進行處理和檢索。例如：當使用者想要詢問一些常見問題時，系統會先進行中文分詞，並以搜尋引擎的檢索方式查詢；當使用者想要搜尋圖書時，則會導引至INNOPAC系統檢索；當使用者輸入「Q：問題A：答案」的句型時，系統將判斷使用者想要教育「小圖」，並將教學指令添加到學習庫中；而當無法在語料庫中找到適當答案時，「小圖」會依靠模式匹配和推理機制從儲存了四萬多條知識分類的AIML知識庫中取得推理後的答案（見圖2）。

（二）臺師大圖書館——參考諮詢機器人

陳宜琳（2019）建置的臺師大圖書館參考諮詢機器人，利用「模式匹配」規則和「關鍵字提取」結果來進行詢問意圖辨識，再採用檢索型的自然語言生成技術透過相似度比對問答語料庫後輸出相關答案作為回覆。該系統目前提供館藏查詢、開館查詢與問答語料庫回覆三大功能，而問答語料庫中僅收錄「服務申請與說明」、「硬體設備介紹」、「資源查詢指引」、「圖書館資訊」和「系統相關」這五大類適合機器人代為回覆的指引型參考問題，而使用的語料來自臺師大圖書館參考紀錄與常見問答集（見圖3）。其使用流程如下：當文字訊息送至自然語言問答系統時，優先經由人工撰寫的「模式匹配」規則判斷，若符合則逕行送

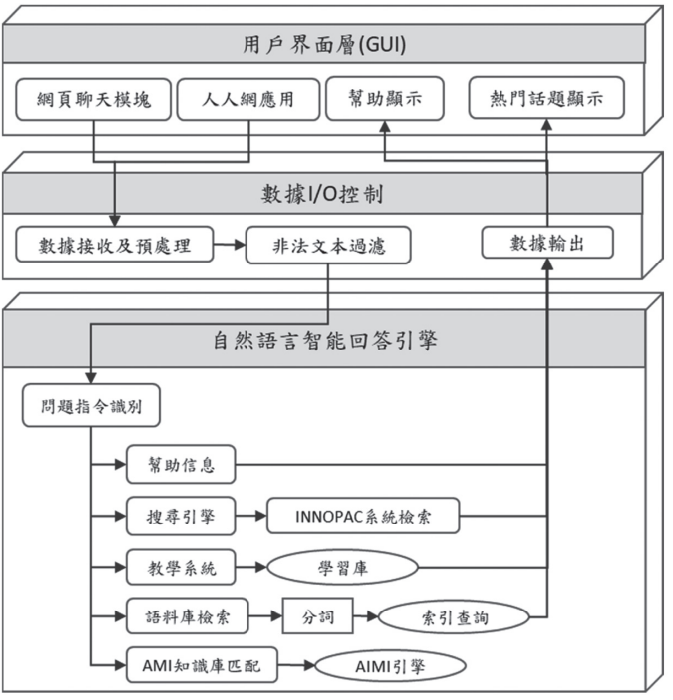

圖2　中國清華大學圖書館「小圖」的系統架構

資料來源：姚飛等人（2011）。

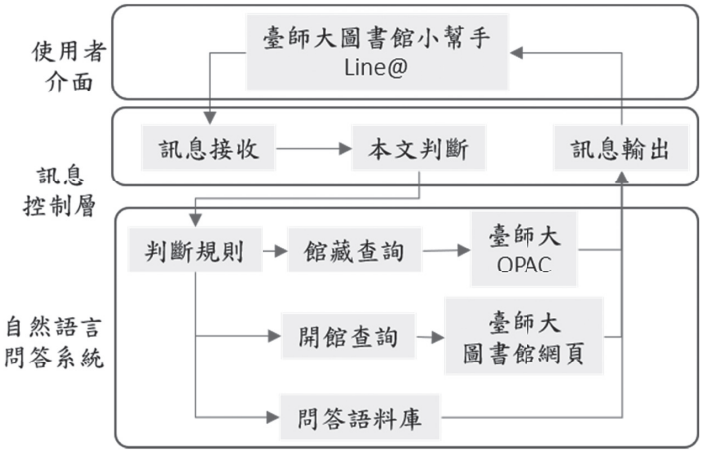

圖3　臺師大圖書館參考諮詢機器人的系統架構

資料來源：陳宜琳（2019）。

入相應的機器人功能模組中處理，接著才提取問句中的關鍵字，並與問答語料庫比對相似度、經排序後傳回最適答案。

系統架構上，「判斷規則模組」的評判依據是藉由使用者輸入的前置詞（如：找館藏）來選擇進入何種對話功能模組（即詢問意圖），這樣的做法雖然讓系統能確認讀者詢問範疇，但相對地也框限了使用者查詢問題方式。此外，系統建置加入一些查找規範的部分，類似由人工進行意圖判斷，這樣的做法可以建置一個簡易有效的分類器以使比對更加聚焦，可用於判斷相同意圖但問法不同的問題類型應該進入何種功能模組；或是人工訂定圖書館領域詞彙表、停用詞表和同義詞表，以便讓此機器人能夠處理圖書館領域的相關用詞，但缺點為需要較多專家知識介入。

（三）臺大圖書館——虛擬參考諮詢服務機器人

范蔚敏（2020）使用Google DialogFlow對話平臺建置臺大圖書館虛擬參考諮詢服務機器人。此一機器人的知識庫來源選用大學圖書館參考服務部落格的文章內容，並且透過「人工擷取」文章內容的核心關鍵字來建立同義字字典（entities）與問答對話組（intents），方便觸發並執行相應的回覆動作。使用上，經過語料訓練後的系統會自動判斷問題的intents，並將預先設計的對話內容輸出給使用者。

分析上述與圖書館相關之中文聊天機器人的建置架構，中國清華大學圖書館「小圖」在建置上雖然綜合了自然語言處理技術、問句相似度檢索匹配、教學系統以及知識推理機制，但在問句的處理上仍是仰賴人工規則的撰寫以及依靠大量訓練語料的支撐；而臺師大圖書館參考諮詢機器人，雖然也是利用訓練語料以及問句相似度檢索匹配進行建置，但在訓練語料的筆數上（僅1,540筆）比起中國清大圖書館的「小圖」少了許多，而且在問句的處理上較簡易，主要藉由人工撰寫的關鍵詞匹配規則來判斷，對於相同意圖但不同問法的問句無法彈性處理。而利用Google DialogFlow對話平臺進行建置的臺大圖書館虛擬參考諮詢服務機器人，雖然它能夠藉由訓練語料的學習來預測使用者的意圖，但在對話內容的範圍和設計上卻無法保有彈性，因為其必須仰賴「人工」設計聊天機器人的對話內容、問句與答句的對應設定（intents）、關鍵詞（entities）的觸發機制。作法上易於藉由人工針對讀者問題建置新的回覆訊息，適合規模較小、固定式的參考諮詢回覆系統建置。

綜整來看，在建置中文聊天機器人的過程中，會面臨以下挑戰（姚飛等人，2011；范蔚敏，2020；陳宜琳，2019）：

1. 中文語序相較於英、美語系更加自由，單字成詞情形繁多，且詞彙之

間無分隔標示，再者中文虛詞運用較多，無形中增加建置中文聊天機器人的困難度。

2. 中文自然語言處理技術發展時間相較英、美語系較晚，所開發與積累的工具與語料資源相差甚多。
3. 由於中文在使用上較自由且無固定的文法規則，蒐集、整理、分析讀者參考諮詢對話內容來建置具語法或語意標記的語料庫，需要耗費大量人力。
4. 當讀者「問題」內容範疇增廣時，就需要蒐集更大量的對話資料庫進行學習，甚至對應語料標記亦須大幅擴增，系統如何兼顧擴充性以及運作效能也是一大問題。
5. 僅以問句中關鍵詞來判斷檢索對應的問答資料庫，往往無法聚焦使用者詢問意圖，將使得判斷結果失準。所以若能藉由所蒐集對話語料，從問題與回應中訓練取得更多意圖關聯規則，來建構參考諮詢對話機器人，可以讓讀者得到的回覆更加精確。

參、參考諮詢意圖分群之研究設計

一、系統架構概述

本文擬由從各圖書館蒐錄「常見問答集」挖掘出讀者所提問句結構或模板（pattern），透過擷取讀者問題的意圖特徵，提供參考諮詢機器人更好的回覆選擇。同時也為了讀者在提問時不受限於特定語彙，實驗中也藉由詞彙擴展作法，讓讀者可以使用相似意圖的語彙進行詢問。在圖4系統實驗架構中，我們先從「常見問答集」語料庫進行「學習」（或「訓練」），也就是針對從網路上蒐錄各圖書館公開的問答集進行預處理，希望從中擷取出具代表的關鍵詞作為文本特徵向量，在此階段我們首先對文本資料進行斷詞、詞性標註、參考諮詢特徵群集向量擷取，為使擷取的特徵語彙能更正確，系統加入維基百科（Wikipedia）條目、圖書專有語彙以及從所蒐集語料中進行共現語彙分析結果，以強化詢問意圖分群的結果。

在選擇分群演算法上，因考量圖書館問答集語料具有多主題性、語義相似性、不規則的聚類結構以及雜訊數據的性質，所以我們會以「演算法要能夠依主題劃分為不同的群集，且具有一定程度的雜訊容忍度」的原則來挑選合的分群演算法進行「問題」重新分群。本文依前述原則

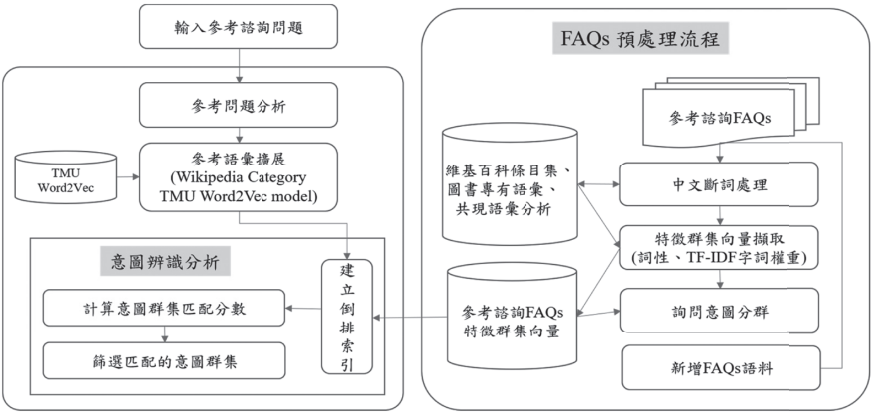

圖4　系統實驗架構

選擇Transformers、*K*-Means以及基於密度之含噪空間聚類法（Density-Based Spatial Clustering of Applications with Noise, DBSCAN）三種演算法對問答語料文本特徵進行初步意圖分群實驗，再就其分群結果評估選擇最適合的演算法，作為本研究中讀者詢問意圖分析之演算法。在「意圖辨識分析」模組系統先解析讀者輸入的參考諮詢問題後，再與經事先分群處理的訓練語料「意圖集群」之間的相似度（關聯程度）來評判得知讀者詢問意圖。

而實驗中考量的三種演算法中，Transformers演算法是由Hugging Face團隊基於Transformer架構（Vaswani et al., 2017）所開發，用於簡化複雜自然語言處理步驟的Simple Transformers函式庫，採用該函式庫本身支援的bert-base-chinese（BERT）模型，加上中研院提供的繁體中文BERT模型，並且基於從各圖書館蒐錄具有類別標記的原始問答語料來進行文本分類任務。*K*-Means是一種基於距離概念但分群效果佳的演算法，它需要事先指定群集的數量*K*，使得每個資料點都屬於最接近的群集的中心，但*K*-Means對初始群集中心的選擇敏感，如果初始群集中心選擇不佳，可能會導致演算法陷入局部最優解或者甚至發散，影響分群結果。而DBSCAN是一種基於密度的分群演算法，它通過識別高密度區域來構建群集，並將低密度區域視為雜訊。它不需要事先指定聚類的數量，並且可以發現任意形狀的群集。而第二、三種使用scikit-learn函式庫中的*K*-Means與DBSCAN方法，皆屬於分群演算法（clustering），是一種根據資料本身的特徵將其他同樣也具有相似特徵的資料逐步聚合為一

個集群的過程，具有「同一群集內資料相似度高，而群集之間相似度則較低」的特性，其他如*K*-Medoids、Hierarchical clustering……等演算法（Kanagala & Krishnaiah, 2016）亦有相同特性。

在本研究上，除了使用較常見的*K*-Means進行實驗外，還另外使用了DBSCAN演算法進行對照。一般來說，*K*-Means演算法適用於一般性分群任務（general-purpose），在設定欲劃分的集群數後，依據點與點之間的直線距離遠近進行分群，直到每個點被分群完畢。與*K*-Means不同的DBSCAN演算法，是一種基於密度的分群演算法，由Martin Ester等人於1996年提出（Ester, Kriegel, Sander, & Xu, 1996），主要是為因應挖掘任意形態特徵的資料而開發，其分群概念是依據資料點在特徵空間中分布的密集程度來進行聚合，亦即當某一區域的資料點特徵相近且密度超過一定門檻，則此區域內的所有資料點被歸為同一群。它同時考慮區域範圍的大小以及區域範圍內的點數量來進行分群，優勢在於可自動排除落於區域之外的點（離群值）。

對於讀者提出諮詢的問題所使用語彙不一定會出現在所蒐錄訓練語料的範圍內，本研究利用多種方法來為不存在的字詞進行查詢擴展，例如：維基百科條目、圖資領域慣用語、高出現頻率的共現*N*字詞（*N*-gram collocations），以及利用臺北醫學大學所訓練的Word2Vec模型找尋相關或相似詞。而擴展後的字詞連同中文斷詞結果，會被用於計算與意圖集群之間的相似度分數（score），評分在平均以上的意圖集群將被輸出。因此，我們最終能獲得與讀者問題相關的意圖集群，從而據此評估分群結果是否有助於理解讀者詢問意圖。

二、系統模組介紹

接下來我們針對系統模組進行說明：

（一）訓練語料蒐集與整理分析

首先利用Python程式語言撰寫網路爬蟲程式，抓取臺灣各圖書館網站中公開常見問答集（參考諮詢FAQs）作為訓練語料，藉此可以瞭解讀者進行參考諮詢使用的詞彙或是習慣問法。本文所收錄語料範圍包括大學圖書館、公共圖書館（包含直轄市圖書總館與省轄市立圖書館）和國家圖書館網站，共100間，總計4,975個問答。而每筆問答紀錄包含libraryname（圖書館名稱）、ID、question（問題）、answer（回覆）、category（所屬類別）、keyword（關鍵詞）以及relatedQ（相關問題）共計七個欄位，如表1所示。一般而言，keyword或可提供詢問意圖判斷，但在蒐錄各館問答集語料時keyword欄位鮮少有訂定（4,983個問答資料

表1

圖書館網站「常見問答集」語料範例

LibraryName	ID	Question	Answer	Category	Keyword	RelatedQ
國家圖書館	1	國家圖書館開放時間為何？	(1) 本館（臺北市中山南路20號）……	開放時間	false	false
國家圖書館	2	國家圖書館為何週一休館？	圖書館對於資訊系統軟、硬體設備……	開放時間	false	false
國家圖書館	3	國家圖書館為何四樓以上各專室……	本館四樓以上專科閱覽室有縮影資……	開放時間	false	false
國家圖書館	4	為何國圖讀者入館年齡需為年滿……	國家圖書館係以政府機關(構)、法……	辦證與入館資格	false	false
國家圖書館	5	如何辦理國家圖書館之閱覽證？	(1) 依據＜國家圖書館閱覽服務規……	辦證與入館資格	false	false
國家圖書館	6	為何國家圖書館不外借圖書而入……	為維持圖書館的秩序，保障館內閱……	辦證與入館資格	false	false
國家圖書館	7	館內是否可以使用行動電話？	讀者不可以在圖書公共空間使用行……	入館規範	false	false
國家圖書館	8	入館利用館藏資源時，可否將背……	本館依「圖書館法」為全國出版品……	入館規範	false	false
國家圖書館	9	國家圖書館的館藏資源主要包含……	(1) 中外文圖書、期刊、學報、報紙……	典藏資源	false	false
國家圖書館	10	如何查詢利用國家圖書館之「電……	因應資訊科技的發展，本館近年來……	典藏資源	false	false
國家圖書館	11	國家圖書館所建置提供的電子資……	為方便讀者利用電腦及網路查……	典藏資源	false	false

中僅有318個問答有設定keyword欄位），僅占6.38%，比例極小，故未考慮使用該欄位屬性。question、answer、category三欄位中資料較為完備，本文目標是進行問答相關的自然語言處理任務，從理解問句結構、意涵及詢問意圖著手，所以著眼於這三個欄位資料處理。

分析蒐集本文的句法結構後，可以發現「問題」與「回覆」之間語彙間的關聯，並且整理獲得「問題類型」與「疑問詞短語」之間的對應關係（見表2）。基本上從各圖書館所整理的常見問答集，依據詢問的目標可區分為具體時間、持續時間、規定、地點、數量、原因、方法、是否和定義這九種問題類型。根據所蒐集語料的問題模式，發現多數的疑問詞（如：什麼、哪些……）需要搭配其前後的詞語（如：時候、資料庫……），方能完整表達使用者想獲得的答案類型。

另外再藉由自然語言處理技術中常見的語彙分析方法，例如：*N*-gram （Majumder, Mitra, & Chaudhuri 2002）、字詞頻率以及共現詞分析（co-occurrence analysis）（Jiao, Liu, & Jia, 2007）等，可以進一步挖掘出讀者參考諮詢經常共同出現的字詞組合，從而得知特定主題下具有關聯性或相似性的習慣用語。舉例來說，詞彙層次（word-level）的雙字母組（bigram），在綜合字詞頻率以及共現詞分析方法後，可以得知哪些詞彙往往共同出現且出現頻率極高，如圖5所示。

除了對問答語料進行語彙分析外，更要能辨識讀者所提問題的詢問意圖，則需回歸問句本身意涵的語意重心和句法結構，將具有相似問法或相同詢問目的（意圖）的問句歸類至同一群。而審視原始蒐集的問答語料庫後，發現蒐集語料中「category」（所屬類別）欄位並無法作為意圖判別依據，因為每間圖書館整理讀者「常見問答集」並未無類別分類規範，雖各館或有相互參考，但仍會因各館館員因各館特性或主觀分類概念不同，使得各館對於「問題」分類類別差異極大。

表3中顯示「常見問答集」語料庫中的問句經自然語言處理、TF-IDF擷取重要語彙作為該問句之關鍵詞向量，再經過加權運算以及向量餘弦相似度演算法計算兩兩問句相似度後，顯示「如何辦理圖書續借」此概念問題相似度分數達70%以上的常見問答中相似問題，其原始分類歸屬相當混亂（包含「流通（借還書服務）」、「借書問題」、「借閱服務」……），然而這些問題從句法結構與用詞可以看出其具有相同的詢問意圖，經系統計算重新歸屬於「借還書」的意圖群集下。

（二）訓練語料預處理

欲理解一個句子需要從表達語意的最小單位「詞彙」（token）著

表2

九種問題類型的常見疑問詞問法

問題類型	疑問詞短語	關鍵疑問詞	例子
項目（具體時間）	開放時間為何	(N)+為何_D	圖書館開放時間為何？
	什麼時間／什麼時候／何時／哪個時候／何時	何時_Nd	在網路上預約書籍完成後，何時可到館取書？
持續時間	多長時間／多少時間	多久_Nd	借閱多久呢？
項目（規則）	規定為何／限制為何	(N)+為何_D	使用館際合作服務的資格限制為何？
地點	什麼地方／什麼地點／哪裡／哪兒／何處	哪裡_Ncd／何處_Nc	請問我要到哪裡找過期的期刊資料呢？
數量	多少／幾	幾本_Na	我可以借幾本書？
原因	為何	為何_D+(N+V)	為何收到逾期通知？
	為什麼	為什麼_D	請問為什麼我介購的圖書遲遲未到館呢？
方法	如何辦理／如何申請／如何利用／如何捐贈／如何得知	如何_D+(V)	如何辦理閱覽證？
	該怎麼辦呢／該怎麼辦／怎麼辦	怎麼辦_VH	我想借的書，圖書館沒有收藏，怎麼辦？
	該如何／要如何	要_D/該_D+如何_D+(V)	該如何辦理借書手續呢？
是否	是否／是否有／是否可以／是否提供	是否_D+(V)	館內是否可以使用行動電話？
	可否／可否用	可否_D	館藏圖書資料可否外借？
	有……嗎／可以……嗎／可以嗎／可以	有_V_2/可以_VH+……+嗎_T	校友可以使用圖書館資料庫嗎？
	可不可以	可不可以_D	在家或宿舍可不可以使用圖書館的資料庫呢？
定義	包含哪些／哪些／有那些／有哪些	包含_VJ／有_V_2+哪_Neqa+(N)	校外人士進入圖書館，有哪些規定？
	有什麼／是什麼／什麼管道	什麼_Nep	如果圖書逾期罰款一直不處理，那麼會有什麼後果嗎？

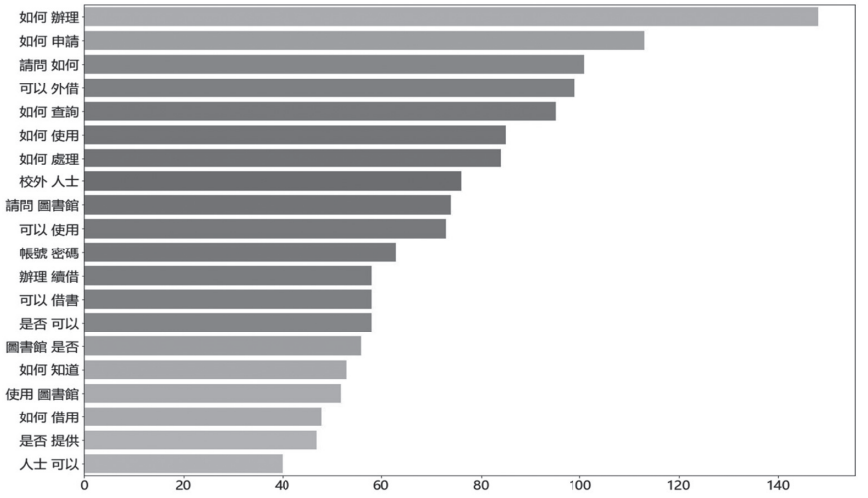

圖5　詞彙（word）層次的bigram次數統計長條圖

手，才能從根本上提升自然語言理解的準確度。對於中文處理而言，能正確標示出文本中詞彙（含詞組）對於理解自然語言是最重要的第一步。在實驗中我們利用中央研究院詞庫小組所開發CkipTagger斷詞器，它是一套以深度學習模型為基礎，針對繁體中文文本進行斷詞和詞性標註的工具，也支援實體辨識（named entity recognition, NER）功能，能夠識別文本中的命名實體，這些特性均有助於文本字詞的正確解析。為了改善斷詞結果，研究中將維基百科條目、圖資領域慣用語以及高出現頻率的共現*N*字詞加入CkipTagger斷詞器的自訂字典之中以強化斷詞結果的正確性。所以在語料的預處理分析上，先使用CkipTagger將問答語料集的「問題」進行中文斷詞與初步詞性標註，再透過TF-IDF計算字詞權重以理解句子本身的語意重心，最後分別以Transformers、*K*-Means以及DBSCAN三種方法對於「問題」進行重新分類或分群，作為蒐錄「問題」重新意圖分群標準的依據，藉此來挑選最為適合的演算法以及分析集群的使用者意圖。

由於中文文本短語出現頻率高，而辭典中亦無法蒐錄所有詞彙，所以斷詞器會將文本中未蒐錄於辭典的長詞切割為幾個連續單詞或短詞，造成後續處理的困難。有不少研究者將維基百科導入應用於聊天機器人實作上（Augello, Pilato, Vassallo, & Gaglio, 2009; Hussain & Athula, 2018），因為維基百科是集合全球群眾智慧共同進行協作、編輯的線上

表3

整併相似類別與修正問句分類之範例（借還書）

問句：如何辦理圖書續借？		類別：借還書
相似度	問句	類別
1.0000	如何 辦理 圖書 續借？	流通（借還書服務）
0.8870	如何 辦理 續借？	借書問題
0.8870	如何 辦理 續借？	False（尚未分類）
0.8870	如何 辦理 續借？	圖書流通問題
0.8870	如何 辦理 續借？	借閱服務
0.8721	如何 辦理 圖書 續借？ 何時 可以 辦理 續借？	續借
0.8367	如何 續借 圖書？	False（尚未分類)
0.8367	如何 續借 圖書？	圖書借閱
0.8082	圖書 可以 續借 嗎？ 要 如何 辦理 續借 呢？	借書、還書及預約
0.7969	請問 如何 辦理 續借？	圖書館借閱查問問題
0.7963	如何 辦理 續借？	False（尚未分類）
0.7963	如何 辦理 續借？	續借
0.7883	要 如何 辦理 續借？	借閱問題
0.7575	如何 辦理 續借 呢？	續借問題
0.7226	我 如何 在 線上 辦理 圖書 續借？	入館閱覽及借閱服務
0.7189	如何 辦理 圖書 續借？何時 可以 辦理 續借？什麼 樣 的 情況 下 無法 辦理 續借？	續借圖書
0.7025	請問 該 如何 辦理 續借？	圖書館

參考工具書，且具備多國語言、即時更新且持續成長的特性，及其蒐羅資源規模、廣度、深度與敘述風格多樣等各種優勢，許多研究中利用其擴增系統詞彙來強化斷詞器分詞結果正確度。

再者我們在語言的使用上往往會有某些詞彙在語句中會高頻的連用，所以透過*N*-gram技巧挖掘出經常問答語料集中緊密相連使用的*N*字詞，如圖6例句中的「找不到」、「本館」、「到館」這類「詞彙」極易被斷詞器斷開，但就語意上來說，它們比起被斷開，合併在一起成為「詞語」才能更好地表達完整的意思，故而將這些連用高頻共現語彙組合納入自訂字典之中。而從圖書館常見問答集中也發現有引號此種標點符號（如：「」、『』）所包夾的詞語，其為數不多且多半為名詞短

語，具有強調詞語的特徵，實驗中將其蒐錄為「圖資領域慣用語」（見圖7）。藉由前述對於斷詞器辭典的調整，可強化從訓練語料中抽取中心語彙的準確度。

將所蒐集的常用問答集的「問題」與「回覆」經中研院開發CkipTagger工具進行中文斷詞（word segmentation）後，再利用TF-IDF演算法對詞彙進行重要程度的加權計算篩選，來擷取語句中所傳達重要的訊息焦點，亦即語意重心。透過TF-IDF詞彙加權調整，我們將文本映射成關鍵詞向量空間，方便後續藉由分類或分群演算法把具有特定樣式或具相近關鍵詞向量文本進行群集分析，亦即藉由向量空間相似度計算將具有相同詢問意圖的問句重新形成集群。

（三）意圖分類／分群演算法分析評估

為了將具有相同詢問意圖的「問題」重新進行分群，研究中嘗試三種具有代表性的方法進行實驗，分別為Transformers中的BERT分類模型與分群演算法的*K*-Means和DBSCAN進行實作測試。首先我們直接將從各圖書館蒐錄具有類別標記的原始問答語料（文字）視為一種

圖6　尚未考量*N*字詞的共現關係之斷詞範例

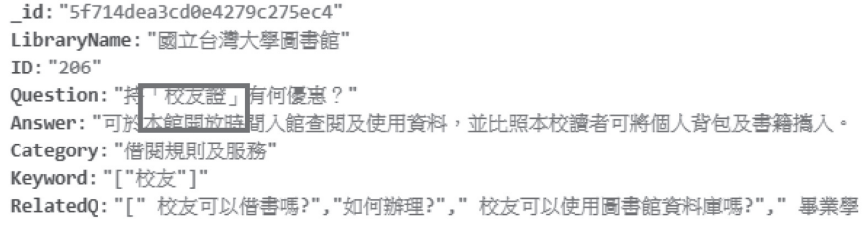

圖7　將引號內的詞納入圖資領域慣用語之範例

多分類（multi-class classification）問題的學習素材（訓練語料），再應用2018年席捲各大NLP競賽排行榜且有著優異表現的BERT模型對全部的問答語料進行重新計算，重新將原始問答語料重新分配新的類別結果，並評估使用者問題意圖重分類後之正確性。而在分群演算法*K*-Means和DBSCAN實驗中，我們將常見問答集語料先利用中研院釋出的CkipTagger工具進行分詞處理，並利用TF-IDF詞彙加權，參考公式(1)、(2)與(3)，調整後作為「問題」的特徵向量，經各自參數調整後再分別透過*K*-Means和DBSCAN進行分群的實作（Ester et al., 1996; Rong, Yan, & Guo, 2004），觀察新劃分群集類別進行演算法優劣的評估。

$$TF = tf(t, d) \tag{1}$$

$$\text{idf}(t, D) = \log\left(\frac{1 + N}{1 + \text{df}(t, D)}\right) + 1 \tag{2}$$

$$TF\_IDF = tf\_idf(t, d, D) = tf(t, d) \times idf(t, d) \tag{3}$$

透過初步實驗我們以「在架上找不到書，該怎麼辦？」此一問句分群結果為例（分群結果範例見表4、表5和表6），藉由BERT模型重新分類預測的結果中，可得到與問句相關的類別為「館藏使用相關問題」，此類別下共有17個文本（問句），觀察判斷後發現僅有4個文本相關度較強，表示與該類別真正相關的文本僅占比24%；而*K*-Means和DBSCAN演算法所得分群結果，其各自文本相關度分別占比為67%和47%。從初步實驗結果觀察BERT模型之問句重新分類結果最差，而*K*-Means和DBSCAN的分群結果較好。假如進一步分析所屬集群中較不相關的語料，則可發現經BERT模型分類預測的不相關語料與問句間詢問意圖上毫無規則或關聯可言；相較之下，透過*K*-Means和DBSCAN進行分群結果較不相關語料範例中，其問句之間仍是比較相關的，因為*K*-Means和DBSCAN演算法是回歸問句本身的描述進行分群，所以可發現這些問答語料在句法結構或關鍵詞上是相像的，譬如這些不相關語料大多保留著「不在架上」、「找不到」或「期刊」的字眼，只是詢問的面向不同而已。

表4

使用Transformers演算法的重新分類結果——以問句「在架上找不到書，該怎麼辦？」為例

演算法	Transformers
所屬群集／類別	館藏使用相關問題
所屬群集／類別之語料總數	17個
與問句相關的語料占所屬集群下的比例：4 / 17 = 0.2353	
所屬集群內與問句相關的語料範例	在架上找不到書，該怎麼辦？ 找不到書，圖書館可以幫忙找？ 資料不在架上，遍尋不獲，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？
與問句不相關的語料占所屬集群下的比例：13 / 17 = 0.7647	
所屬集群內與問句不相關的語料範例	想要贈書給圖書館，要如何處理？ 我想要學習如何查詢資料，要如何參加？ 圖書館的電腦可以使用隨身碟？ 如何知道圖書館有何新的館藏？ 忘記【我的帳戶】登入帳號密碼，怎麼辦? 各類資料放置何處？ 如何知道圖書館有何新的館藏？ 各類資料放置何處？ 我想找考古題？ 如何知道圖書館有何新的館藏？ 如何推介資料／圖書館處理推介資料的流程？ 參考書／博碩士論文／教師課程指定參考書可以外借嗎？ 考古題都只有題目，是否可以提供詳解呢？

註：本表中出現有重複問題例如：「資料不在架上，遍尋不獲，該怎麼辦」，其原因為不同館讀者詢問問題相同，但其問題來源圖書館不同，甚至分類歸屬亦不同，故並未作同整刪除。

在上述實驗問句「在架上找不到書，該怎麼辦？」經重新歸屬群集，似乎*K*-Means的分群效果比DBSCAN好，且兩者均優於使用Transformers中的BERT模型。但細觀兩者分群結果，以「圖書館開放時

表5

使用*K*-Means演算法設定分群數目為270的分群結果——以問句「在架上找不到書，該怎麼辦？」為例

演算法	*K*-Means
所屬群集／類別	Cluster 4
所屬群集／類別之語料總數	30個
與問句相關的語料占所屬集群下的比例：20 / 30 = 0.6667	
所屬集群內與相關的語料範例	館藏查詢圖書為「在架上」但卻找不到書，其原因？該如何處理？
	常常要找書，系統明明顯示在架上，可是就是找不到？如果又急著要用，怎麼辦？
	「歡迎外借」的書，為什麼在架上找不到呢？
	請問在網路上查書顯示在架上，但到書架上卻找不到書怎麼辦？
	為何在架上找不到「在架」的書？是否遺失了呢？
	為何在架上找不到「在架」的書？是否遺失了呢？
	在架上找不到書，該怎麼辦？
	為何圖書館藏狀態顯示「在架上」，卻在書架上找不到該圖書？
	資料顯示在架上，但實際卻找不到？
	書況顯示「在架上」的書找不到，怎麼辦？
	為什麼使用圖書館查詢系統顯示書本在架上，實際到架上尋找的時候卻找不到書？
	資料不在架上，遍尋不獲，該怎麼辦？
	系統顯示資料未借出，但不在架上，要怎麼辦？
	查詢館藏目錄，電腦顯示「在架上」，可是為何找不到書？該怎麼辦呢？
	資料不在架上，遍尋不獲，該怎麼辦？
	為何在架上找不到我要的書籍或刊物？
	在網路上查書顯示在架上，但到書架上卻找不到書怎麼辦？
	館藏紀錄上有，但書不在架上怎麼辦？
	我要借的書不在架上，怎麼辦？
	我在電腦查到書在架上，為何找不到？
與問句不相關的語料占所屬集群下的比例：10 / 30 = 0.3333	

表5

使用*K*-Means演算法設定分群數目為270的分群結果——以問句「在架上找不到書，該怎麼辦？」為例（續）

演算法	*K*-Means
所屬集群內與不相關的語料範例	我怎麼在架上會找不到語言學習類的期刊呢?
	為什麼現期的期刊卻在架上找不到？
	請問期刊不在架上該怎麼辦？
	為什麼從電腦目錄上查到的期刊卷期，卻在架上找不到？
	請問期刊不在架上該怎麼辦？
	我想找的期刊，在架上卻找不到？
	請問現期期刊不在架上該怎麼辦？
	為什麼在架上找不到以前的期刊？
	請問期刊不在架上怎麼辦？
	期刊並不在架上？

間」相關問句分群結果來看（見表7），由DBSCAN而得的語料分群更為精細，而且會再依不同共現語彙「圖書館」、「影印室」、「多媒體視聽區」、「24H閱讀區」等再區隔細分不同群集，能更細緻的判斷讀者的詢問意圖。

綜合上述三種演算法的初步實驗結果，總結出藉由Transformers中BERT模型進行重新分類預測的結果最差，原因可能是所收錄的問答語料各分類語料筆數規模不足，造成Transformers在學習和預測類別上容易預測成具有較多語料筆數的類別，導致結果不如預期。而*K*-Means和DBSCAN這兩種演算法會依據問句本身特徵逐步將其他同樣也具有相似特徵的問句聚合為一個集群，其分群效果均優於Transformers。從問答集語料中發現有些問句的句法結構極為相近，只是詢問標的有所差異；舉例而言「大學生借書有哪些限制？」、「研究生的借書規定？」和「老師身分的借閱規定為何？」這三個讀者問題均與借書規定相關，只是身分對象不同。而有些詢問的關鍵詞是相似的，只是詢問的面向不同而已，再者如「學生證遺失，該如何處理？」、「找不到我要的書，怎麼辦？」和「如何申請研究小間？」以上三種問題皆是詢問「方法」此種問題類型，只是面向證件遺失、書籍找尋、空間借用不同的圖書館常見問題類別而已。

表6

使用DBSCAN演算法設定半徑0.8、最小資料數目為2的分群結果——以問句「在架上找不到書，該怎麼辦？」為例

演算法	DBSCAN
所屬群集／類別	Cluster 284
所屬群集／類別之語料總數	17個
與問句相關的語料占所屬集群下的比例：8 / 17 = 0.4706	
所屬集群內與問句相關的語料範例	書找不到怎麼辦？ 我找不到我想借的書，怎麼辦？ 書找不到怎麼辦？ 在架上找不到書，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？ 資料不在架上，遍尋不獲，該怎麼辦？ 我要借的書不在架上，怎麼辦？ 找不到書，該怎麼辦？
與問句不相關的語料占所屬集群下的比例：9 / 17 = 0.5294	
所屬集群內與問句不相關的語料範例	為什麼現期的期刊卻在架上找不到？ 請問期刊不在架上該怎麼辦？ 請問期刊不在架上該怎麼辦？ 期刊並不在架上？ 找不到館藏期刊怎麼辦？ 請問期刊不在架上怎麼辦？ 我想找的期刊，在架上卻找不到？ 為什麼在架上找不到以前的期刊？ 請問現期期刊不在架上該怎麼辦？

面對上述情況，*K*-Means和DBSCAN演算法會回歸問句本身特徵進行訓練，藉此進行分群較能區分讀者詢問問題之意圖。其中又以DBSCAN無須事先決定欲劃分的群集數量、透過文本特徵學習來細分不同集群詢問意圖，同時也能排除與其他群集差異極大的語料（離群值），能更為細緻的判斷詢問內容的意圖（見表8）。

因此，本文選擇基於密度的DBSCAN演算法來為問答語料重新分群，理由如下：1. Transformers中BERT模型進行多類別分類任務的做法是對原始類別進行特徵學習後再重新預測其類別，因此在分類標準不一

表7

使用*K*-Means和DBSCAN演算法的分群結果——以圖書館開放時間相關的問句為例

問句	*K*-Means演算法下的所屬群集	DBSCAN演算法下的所屬群集
從何處可以得知圖書館開放時間及休館日？	Cluster 59	Cluster 118
從何處可以得知圖書館開放時間及休館日？	Cluster 59	Cluster 118
請問可由何處得知圖書館開放時間及休館日？	Cluster 59	Cluster 118
圖書館開放時間？國定假日有開館嗎？	Cluster 59	Cluster 158
圖書館開放時間?國定假日有開館嗎?	Cluster 59	Cluster 158
圖書館開放時間？國定假日有開館嗎？	Cluster 59	Cluster 158
請問圖書館的開放時間？假日或寒暑假有開館嗎？	Cluster 59	Cluster 158
影印室的開放時間？	Cluster 59	Cluster 327
影印室開放時間？	Cluster 59	Cluster 327
請問視聽區的開放時間？	Cluster 59	Cluster 440
請問多媒體視聽區的開放時間？	Cluster 59	Cluster 440
24H閱讀區開放時間？	Cluster 59	Cluster 515
國家圖書館開放時間為何？	Cluster 59	Cluster 65
請問圖書館開放時間？	Cluster 59	Cluster 65
請問圖書館開放時間？	Cluster 59	Cluster 65
請問圖書館的開放時間是？	Cluster 59	Cluster 65
圖書館開放時間？	Cluster 59	Cluster 65
請問圖書館開放時間？	Cluster 59	Cluster 65
圖書館開放時間為何？	Cluster 59	Cluster 65
請問圖書館開放時間？	Cluster 59	Cluster 65
圖書館的開放時間？	Cluster 59	Cluster 65
圖書館開放時間？	Cluster 59	Cluster 65
非圖書館開放時間如何還書？	Cluster 59	Cluster 65
請問圖書館開放時間？	Cluster 59	Cluster 65
圖書館開放時間	Cluster 59	Cluster 65
開放時間	Cluster 59	Cluster 65

表 7

使用*K*-Means和DBSCAN演算法的分群結果——以圖書館開放時間相關的問句為例（續）

問句	*K*-Means演算法下的所屬群集	DBSCAN演算法下的所屬群集
請問圖書館開放時間？	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
圖書館開放時間?	Cluster 59	Cluster 65
請問圖書館開放時間為何？	Cluster 59	Cluster 65
期中、期末考週有延長開館嗎？	Cluster 67	Cluster 157
期中、期末考期間，週末有延長開館嗎？	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎？	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎？	Cluster 67	Cluster 157
期中、期末考週有延長開館嗎？	Cluster 67	Cluster 157
桃園市立圖書館各館開館時間？	Cluster 67	Cluster 85
圖書館開館的時間為何？	Cluster 67	Cluster 85
請問高空大圖書館開館時間？	Cluster 67	Cluster 85
圖書館開館時間及借還書櫃臺服務時間為何？	Cluster 67	Cluster 85
請問高空大圖書館開館時間？	Cluster 67	Cluster 85
圖書館的開館時間？	Cluster 67	Cluster 85
中市圖各分館開館時間？	Cluster 67	Cluster 85

的情況下學習特徵傾向會預測為擁有較多訓練語料的大類別，而大幅影響預測結果。反觀DBSCAN和*K*-Means，由於分群是根據問答語料本身的句法結構或語意資訊來進行聚類，因此結果較為準確；2. DBSCAN不需要事先確定集群數目，而是依據所設定的參數自動分群，因此勝過需要預先決定集群數目的*K*-Means；3. DBSCAN不受極端值（noise）影響，因為極端值不會被硬性的歸類為任意一群，也就比較不會有不相似的問答語料歸類在集群內，勝過強硬將每一個問答語料歸類於某一集群的*K*-Means；4. DBSCAN藉由設定半徑範圍（ε或Eps），決定集群的數量和大小，從而彈性地調整集群內詢問問法的嚴謹或籠統，即界定詢問意圖的細緻度。

DBSCAN演算法中主要的參數設定有兩個：Eps和min_samples。Eps

表8

Transformers、*K*-Means和DBSCAN的實驗結果整理

演算法	Transformers	*K*-Means	DBSCAN
類別	分類	分群	分群
實驗做法	使用BERT模型對原始問答語料重新分配新的類別結果。	使用CkipTagger對原始問答語料先進行分詞處理，再利用TF-IDF詞彙加權調整後作為「問題」的特徵向量，經參數調整後得出分群結果。	使用CkipTagger對原始問答語料先進行分詞處理，再利用TF-IDF詞彙加權調整後作為「問題」的特徵向量，經參數調整後得出分群結果。
實驗分群數量	490	270	531
範例集群之語料總數（*T*）	17	30	17
與問句相關的範例集群內語料數目（*n*）	4	20	8
範例集群內相關比例（*n*/*T*）	0.2353	0.6667	0.4706
優劣	• 較易受訓練語料類別數目影響，而傾向被預測成原本就擁有較多語料筆數的類別 • 分類預測結果較差	• 必須事先決定欲劃分的集群數量 • 較易受極端值影響，而強硬被歸類 • 回歸問句本身特徵進行訓練，能區分讀者詢問問題之意圖	• 依據參數自動分群，彈性調整集群的數量和大小 • 不受極端值影響 • 回歸問句本身特徵進行訓練，能區分讀者詢問問題之意圖 • 從分群結果觀察，群集劃分更加細緻，且集群內共現語彙現象明顯，能更細緻的判斷讀者的詢問意圖（意圖的顆粒度）

調整的是每個資料點「要搜尋的周圍範圍需要多大」，而min_samples設定的是一個範圍內「需要有多少個資料點以上」才算密度夠高、能夠形成一個集群。藉由參數調整的優點，讓資料點自動形成一個個集群，而無須事先決定集群數目。在實驗中我們調整不同Eps和min_samples兩個參數組合，觀察產出群集數量與分群效果，從實驗數據中發現將參數條件設定Eps = 0.8和min_samples = 2，這樣效果較佳，使得具有相似關鍵詞的語料盡可能地被劃分在同一集群內。

我們將經過預處理的問答語料庫經DBSCAN演算法重新計算分群後，得到較一致且準確的意圖集群。然而，由於參數設定得較為嚴苛，使得集群內的訓練語料傾向擁有相同的問法（句法結構），而那些同樣具有相同意圖但不同問法的語料則會被另外自成一群，導致得到較多的意圖集群，共計528個。接著，為了將那些「具有相同意圖但不同問法」的集群合併在同一個意圖集群內，首先會對每個意圖集群的字詞進行TF-IDF加權運算，得到每群前6名的關鍵詞，而這些關鍵詞代表各意圖集群的語意重心。隨後，藉著這些語意重心的共現關係，篩選出具有相似關鍵詞分布的意圖集群，再以人工來判斷合併意圖集群（見圖8）。

肆、意圖集群結果分析

圖書館有許多資料本身因由人對資料進行分類或欄位資料著錄，所以隨著人為認知上差異或資料登載的疏漏，使得後續資料使用時造成困擾，研究中圖書館常見問答集館員在對於問題回復或問題歸屬分類並無一致性或強制的規範，所以各館究其本身對問題認知自行對問題進行類別定義與歸屬，這樣一來造成讀者透過網頁歸屬類別瀏覽找尋問題解答時有著不小的落差，本研究希望透過問題本身意圖分析，重新定義問題

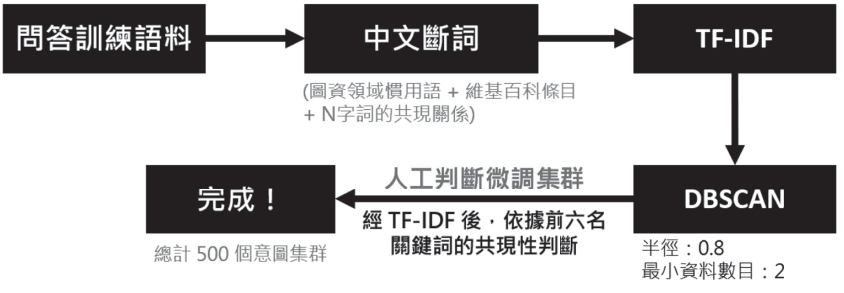

圖8　應用DBSCAN分群的實驗流程

類群屬性，將讀者相同詢問意圖的問題重新聚類，便於讀者在瀏覽問答集時更快能理解所想要獲取的相關答案。

一、實驗範例闡述

本文透過密度為本的聚類分析演算法DBSCAN對於從各圖書館常見問答集蒐錄原始問答文本語料進行訓練，嘗試以使用者意圖將讀者問題進行重新分群，以下就分群輸出結果進行討論：

（一）重新調整因人為認知所產生分類群集差誤

從各館常見問答集網站蒐錄的問答集，因為各館館員對於問題認知與各館任務上的差異，在整理讀者詢問問題與館方回應「常見問答集」問題分類標準並無一致，所以相同或類似問題不同圖書館在問題歸類上有不小的差異，造成讀者透過網站公開「常見問答集」分類類別架構找尋答案時的困擾。本文希望透過所蒐錄讀者「常見問答集」中「問題」間的詢問語彙相似度進行重新分群，藉由辨識讀者意圖重新調整問題類別歸屬，把詢問意圖相同的問題重新歸類於同一群集，如此更有助於讀者查詢到所需詢問的意圖群集，也能更快速尋找所需的「答案」（見表9）。

（二）彙整詢問語法結構不同但具相同意圖群集

從表10可知「資料未借出，可是架上找不到？」的這類問題，其原始類別分歸於「借書問題」、「借書」、「館藏利用」、「圖書館藏」甚至有些是沒有類別的，研究中透過DBSCAN的分群結果（Cluster 175）卻能將這些具有類似問法結構的問題歸為同一種意圖集群中。由此可觀察到，所蒐錄的原始問答語料其分類類別的差異，極大程度取決於各館人員對於問題分類主觀概念。而這樣的分類差異，使得在詢問問題的認知上館員自行定義的問題類別與讀者有所不同，從而導致意圖辨識不如預期。簡言之，分群結果的正確性，不只影響詢問意圖辨識的正確性，也會對後續諮詢回應造成影響。

（三）藉由問句中詢問語彙集合運算解析細分讀者的詢問意圖

再者某些群集具有相同核心意圖，但在觀察問句中其他附屬語彙，可以再細分不同詢問意圖，從圖9顯示「Cluster 174：臨時學生證可以借書嗎？」、「Cluster 256：可以到其他學校借書嗎？」和「Cluster 448：可否用其他同學的借書證借書？」這三種集群間均具「借書」的核心意圖語彙，但從三個群集語彙集合間的關係，我們可以透過「問句」中其他詞彙來細分讀者詢問意圖。舉例來說，當讀者的問題中出現「借書」

表9

因人為認知分類差誤範例

來源	語料	原始類別	DBSCAN分群
Cluster 172：以預約書問題為例			
淡江大學覺生紀念圖書館	預約資料到館，會通知嗎？如何借？	借書問題	Cluster 172
聖約翰科技大學圖書館	預約之館藏到館，會通知嗎？如何借？	FALSE	Cluster 172
嘉藥學校財團法人嘉南藥理大學圖書資訊館	預約資料到館，會通知嗎？如何借？	預約	Cluster 172
國立交通大學圖書館	預約資料到館，會通知嗎？如何借？	借還書	Cluster 172
屏東大學圖書館	預約資料到館，會通知嗎？如何借？	資料借還	Cluster 172
臺南市公共圖書館	請問預約到館，如何通知我？	借書	Cluster 172
Cluster 122：以閉館還書問題為例			
國立聯合大學圖書館	閉館後要如何還書？	FALSE	Cluster 122
淡江大學覺生紀念圖書館	閉館可以還書嗎？	還書問題	Cluster 122
嘉藥學校財團法人嘉南藥理大學圖書資訊館	閉館可以還書嗎？	還書	Cluster 122
國立成功大學圖書館	閉館時，應如何還書？	資料借還	Cluster 122
國立金門大學圖書館	閉館時，應如何還書？	圖書借閱	Cluster 122
國立東華大學圖書資訊處	閉館時如何還書？可寄回還書嗎？	借還書	Cluster 122
臺南家專學校財團法人臺南應用科技大學圖書館	請問閉館時，我該如何還書？	資料借還	Cluster 122
中山醫學大學圖書資訊處	我到圖書館還書時，已經閉館了，該怎麼辦？	圖書資料借還	Cluster 122
中國文化大學圖書館	閉館後可以還書嗎？	館藏使用規則	Cluster 122

表9

因人為認知分類差誤範例（續）

來源	語料	原始類別	DBSCAN分群
Cluster 183：以逾期罰款問題為例			
淡江大學覺生紀念圖書館	逾期罰款的計算方式？	還書問題	Cluster 183
逢甲大學圖書館	逾期罰款如何計算？	與借書相關	Cluster 183
嘉藥學校財團法人嘉南藥理大學圖書資訊館	逾期罰款的計算方式？	借閱問題	Cluster 183
屏東大學圖書館	逾期罰款的計算方式？	資料借還	Cluster 183
國立彰化師範大學圖書與資訊處	逾期罰款有緩衝期嗎？如何計算？	借閱問題	Cluster 183
國立彰化師範大學圖書與資訊處	逾期罰款如何計算？	借閱問題	Cluster 183
國立東華大學圖書資訊處	放假時逾期罰款怎麼計算？	寒（暑）假優惠借期	Cluster 183
樹德科技大學圖書館	逾期罰款如何計算？	流通服務問題——罰款及遺失賠償	Cluster 183
長榮大學圖書館	逾期罰款計算方式？	借閱服務	Cluster 183
中山醫學大學圖書資訊處	逾期罰款如何計算？	圖書資料借還	Cluster 183
中原大學張靜愚紀念圖書館	我借的書過期了，罰款如何計算？	借閱相關問題	Cluster 183
國立臺灣科技大學圖書館	逾期罰款如何計算？	圖書資料的借還問題	Cluster 183

一詞時，我們無法確切得知讀者想詢問哪一方面的問題；而當讀者的問題中同時出現「借書」和「學生證」的詞彙時，我們能將範圍縮小至「Cluster 174」和「Cluster 448」這兩個意圖集群；假如又有「同學」此一詞彙加入的話，則我們將能得出讀者所問的問題屬於「Cluster 448：可否用其他同學的學生證借書？」此一意圖集群，從而能夠將相應的答覆傳回給讀者。

（四）依據疑問句型或詢問目標進行分群統整

從所蒐集語料訓練所得之意圖群集Cluster 38為例，透過所擷取關

表10

問法不同但具相同意圖群集(以借書問題為例)

來源	語料	原始類別	DBSCAN分群
淡江大學覺生紀念圖書館	資料未借出，可是架上找不到？	借書問題	Cluster 175
嘉藥學校財團法人嘉南藥理大學圖書資訊館	資料未借出，可是架上找不到？	借書	Cluster 175
屏東大學圖書館	資料未借出，可是架上找不到？	借書問題	Cluster 175
長榮大學圖書館	資料顯示在書架上，可是架上找不到書？	館藏利用	Cluster 175
高雄醫學大學圖書資訊處	架上找不到我想借閱的書該怎麼辦？	圖書館藏	Cluster 175
慈濟學校財團法人慈濟大學圖書館	找不到架上的書？	無分類	Cluster 175
聖約翰科技大學圖書館	館藏未借出，但架上找不到？	無分類	Cluster 175

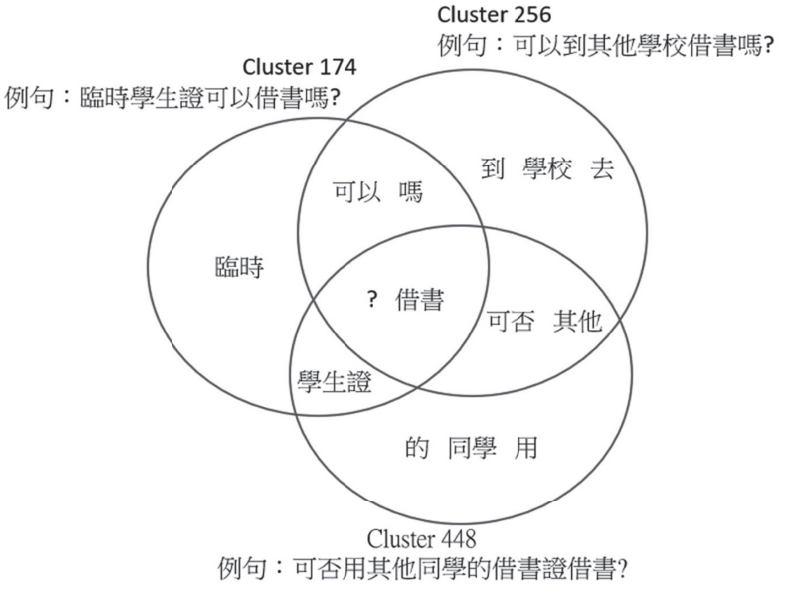

圖9　DBSCAN分群結果群集間關鍵語彙關係（以借書問題為例）

鍵詞計算依其權重產生Cluster 38文字雲（見圖10），從中可以清楚看到該意圖集群的主要關鍵詞有「怎麼辦」、「遺失」和「處理」等，由此可推知該意圖集群主要在詢問「關於遺失東西該如何處理的問題」。回溯該群集原始蒐集的問題集（見表11），可以驗證「遺失」、「如何處理」和「怎麼辦」這些詞的權重對於該群集的影響相較於其他字詞更為巨大。再者對於問句中其他語彙進行關聯分析，則可發現其詢問目標可細分為館藏、證件、圖書附件、圖書……等項目。這使得建置對話系統時在辨識到讀者詢問與遺失物品相關問題時，能進一步地擷取讀者的詢問目標，可以更準確產生對應回覆結果。

（五）讀者問題句法結構相似但詢問「標的」不同

根據表2本研究從收錄的圖書館常見問答集中歸納的九種「問題」類型，藉由「問題」語彙與句法結構分析，我們發現許多「問題」在句法結構與詢問意圖大致相同，但因詢問「標的」不同而被歸類到不同的類別，表12顯示「期刊是否可外借」這類群問題，具相同詢問意圖且句法結構也相近但詢問「標的」不同，可能被歸屬於不同的類別（包括FALSE〔無分類〕、借閱相關問題、報紙及期刊雜誌、期刊及報紙、期刊常見問題，以及與借書相關）。經DBSCAN重新分群將這些「問題」歸類為Cluster 108（該集群與「期刊、報紙或雜誌借閱」相關）。如此一來，除了解決館員分類標準不一的問題，更能在重新分群後將原始類別納入考量，賦予其更明確的內容界定。未來當讀者詢問「光碟是否可外

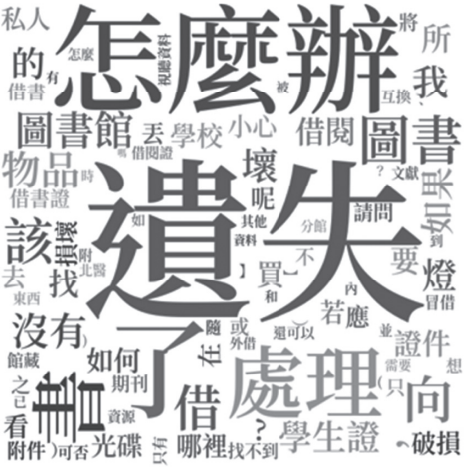

圖10　依關鍵詞權重產生的文字雲（以Cluster 38為例）

表11

意圖集群Cluster 38的訓練語料

Cluster38訓練語料	詢問目標
向圖書館外借的館藏遺失了，該如何處理？	館藏
學生證遺失怎麼辦？	學生證
遺失圖書館證件，我應如何處理?	證件
借書遺失該如何處理？若只遺失隨書所附之光碟呢？	書、光碟
圖書附件（如光碟）遺失如何處理？	圖書附件
資料損壞或遺失要怎麼辦？	資料
圖書遺失了如何處理？	圖書
如果借的圖書遺失了怎麼辦？	圖書
圖書遺失如何處理？	圖書
借閱圖書遺失怎麼辦？	圖書
向圖書館借閱的圖書不小心遺失了怎麼辦？	圖書
請問圖書如果找不到怎麼辦？	圖書
書丟（壞）了，怎麼辦？	書
書遺失了怎麼辦?	書
借的書找不到（遺失了）或破損了怎麼辦？	書
我在圖書館借的書丟了怎麼辦？	書
向圖書館借的書若遺失了，該怎麼辦？	書
如果向圖書館所借之書已遺失，該如何處理？	書

表12

重整「期刊、報紙或雜誌借閱」相關的語料（Cluster 108）

來源	語料	原始類別	DBSCAN分群
慈濟學校財團法人慈濟大學圖書館	期刊是否可外借？	FALSE	Cluster 108
中原大學張靜愚紀念圖書館	期刊是否可外借？	借閱相關問題	Cluster 108
新北市立圖書館	期刊雜誌可外借嗎？	報紙及期刊雜誌	Cluster 108
開南大學圖書資訊處	圖書館期刊及報紙是否可外借？	期刊及報紙	Cluster 108
國立陽明大學	期刊（雜誌）是否可外借？	期刊常見問題	Cluster 108
亞洲大學圖書館	各種期刊是否可外借？	與借書相關	Cluster 108

借？」（於訓練文本未曾出現的「問題」），系統亦能判斷此「問題」的類別，如此對於「答案」回覆會有極大助益。

在表13顯示群集Cluster 105中問句語料大都與「影印設備與服務」詢問意圖相關。不過，在詢問「問法」有些許不同，詢問意圖相似但詢問「標的」含括影印卡購買、計費方式、列印服務、列印用紙的提供……等。研究中依據詢問目標（關鍵詞）進行關聯性擴展，可以將詢問意圖相同但標的不同的「問句」限縮歸屬同一意圖群集。

（六）「問題」具相同關鍵詞但前後關聯語彙不同，對應意圖集群不同

從分析自然語言的角度來看，不管是問句或是敘述句，在傳達資訊時必定會有訊息焦點，亦即語意重心，因此句子內每一個詞彙其重要性應有大小之別，賦予不同的權重。藉由理解句子的語意重心，我們能夠抓出讀者所欲詢問的方向，再透過同樣具有較高權重的上下文或關鍵詞，將有助於逐步限縮範圍，聚焦主題（即詢問意圖）。此一概念若拓展到意圖集群（intent cluster）的層次，將能夠識別集群與集群之間在詢問意圖上的關聯與差異，有助於在讀者不清楚該如何描述其問題的情況下，藉由篩選出具有共同語意重心（即高權重的關鍵詞）的意圖集群來幫助讀者縮減問題的範圍，或是得知具有共同語意重心的意圖集群其大致有哪些詢問方向。

舉例來說，假如讀者的問題過於模糊，僅僅知道其問題的大致方向是關於「遺失」和「證件」兩個關鍵詞的組合，我們藉所擷取高權重關鍵詞從意圖集群進行篩選，如表14取得包含關鍵詞「遺失」的意圖集群。接著，再挑選同樣具有「證件」的意圖集群，從而得到Cluster 379（關於證件遺失如何處理）和Cluster 392（關於撿到他人遺失的證件）兩種詢問方向的意圖集群（見表15例句）。

二、實驗結果綜述

綜上所述，研究中透過前導分群演算法實驗中，Transformer模型通常需要大量標註的數據來進行有效的訓練，圖書館「常見問答集」沒有足夠大量類別標註資料可能是導致其表現不如預期的原因之一；*K*-Means需要事先指定群集的數量*k*，且對初始群集中心的選擇敏感。如果初始群集中心選擇不佳，可能會導致演算法陷入局部最優解或者甚至發散，影響分群結果。DBSCAN算法它不需要事先指定聚類的數量，並且可以發現任意形狀的群集，而問答集中文本資料的聚類結構屬於不規

表13

重整「影印設備與服務」相關的語料（Cluster 105）

來源	語料	原始類別	DBSCAN分群
東海大學圖書館	有關影印服務	其他	Cluster105
國立臺北護理健康大學圖書館	查詢資料庫，圖書館有提供列印嗎？	電腦檢索與網路	Cluster 105
臺北市立大學圖書館	哪裡提供影印、列印服務？	FALSE	Cluster105
國立交通大學圖書館	圖書館有提供列印服務嗎？	影印、列印、掃描	Cluster105
臺南家專學校財團法人臺南應用科技大學圖書館	圖書館有無提供影印服務，影印卡何處購買？	其他	Cluster105
臺中市立圖書館總館	圖書館的影印服務？	其他	Cluster105
亞洲大學圖書館	圖書館有無提供影印服務，影印點數如何儲值？	關於空間與設備	Cluster105
淡江大學覺生紀念圖書館	圖書館可否提供列印用紙？	環境與設備	Cluster105
屏東大學圖書館	圖書館哪裡可影印或列印？	場地與設施	Cluster105
逢甲大學圖書館	圖書館哪裡可影印或列印？如何計費？	關於空間與設備	Cluster105
嘉藥學校財團法人嘉南藥理大學圖書資訊館	影印服務	影印服務	Cluster105
新北市立圖書館	請問圖書館內有無提供影印服務？	閱覽環境、空間設備	Cluster105
長榮大學圖書館	請問圖書館是否提供列印／影印的服務？	環境與設備	Cluster105
國立成功大學圖書館	請問圖書館是否提供影印的服務？	其他	Cluster105
嶺東科技大學圖書館	館內哪裡可影印或列印？如何計費？是否可以彩色列印？	FALSE	Cluster105

表14

含「遺失」關鍵詞的意圖集群

意圖集群編號	Cluster 38	Cluster 86	Cluster 140	Cluster 219	Cluster 329	Cluster 379	Cluster 392	Cluster 410	Cluster 425	Cluster 436	Cluster 511
Top1	**遺失** **0.3359**	賠償 0.4886	閱覽證 0.9965	補辦 0.7203	臨時閱覽證 0.7979	損壞 0.2656	撿到 0.5075	損毀 0.5975	在架 0.4482	出版社 0.3724	存物櫃 0.9308
Top2	怎麼辦 0.1817	**遺失** **0.3248**	怎麼辦 0.5299	**遺失** **0.4788**	**遺失** **0.5304**	借書 0.2426	他人 0.4707	**遺失** **0.3534**	在架上 0.3332	書局 0.3724	鑰匙 0.8632
Top3	處理 0.0692	毀損 0.2094	沒 0.5185	校友借書證 0.4192	歸還 0.3610	證件 0.2093	證件 0.3708	處理 0.2770	**遺失** **0.2650**	買不到 0.3724	怎麼辦 0.3975
Top4	書 0.0679	借閱 0.1320	**遺失** **0.4124**	時 0.2850	怎麼辦 0.2270	學生證 0.1990	財物 0.2856	向 0.1942	找不到 0.2418	已經 0.2889	忘了 0.3888
Top5	圖書 0.0638	圖書 0.1259	帶 0.4041	如何 0.1249	忘記 0.2061	**遺失** **0.1904**	物品 0.2356	時 0.1402	為何 0.1717	**遺失** **0.2476**	**遺失** **0.3092**
Top6	向 0.0583	如何 0.0847	—	借書證 0.1094	逾期 0.1475	學生 0.1831	怎麼辦 0.2168	借 0.1072	是否 0.1535	怎麼辦 0.1591	帶 0.3030
Top7	借 0.0536	—	—	—	處理 0.1385	處理 0.1492	**遺失** **0.1689**	圖書 0.0959	—	書 0.1298	—
Top8	圖書館 0.0497	—	—	—	—	如何 0.1489	圖書館 0.1053	書 0.0926	—	—	—

註：關鍵詞「遺失」在各意圖集群的出現機率

表15

含「遺失」關鍵詞的意圖集群之問句範例

意圖集群編號	問句範例
Cluster 38	向圖書館外借的館藏遺失了，該如何處理？
Cluster 86	圖書遺失或毀損怎麼辦？
Cluster 140	閱覽證遺失怎麼辦？
Cluster 219	校友借書證遺失時，要如何補辦？
Cluster 329	臨時閱覽證忘記歸還或遺失該如何處理？
Cluster 379	學生證件遺失了，該如何處理？又如何借書？學生證損壞了如何借書？
Cluster 392	在圖書館撿到他人遺失的物品或證件，怎麼辦？
Cluster 410	向圖書館借的書損毀或遺失時，該如何處理？
Cluster 425	為何在架上找不到「在架」的書？是否遺失了呢？
Cluster 436	我遺失的書在出版社或書局已經買不到了，怎麼辦?
Cluster 511	存物櫃鑰匙遺失了怎麼辦？

則且具有不同的密度，相較下DBSCAN是一個適合的選擇。它能夠發現任意形狀的聚類，並且對雜訊數據具有較好的穩健性（robustness）。

選擇DBSCAN作為本文意圖分群演算法，將所蒐集各圖書館常見問答集進行訓練學習重新將「問題」進行意圖分群，的確能將具有類似問法結構的問題（固定或相似問題模板Template）含括視為同一種意圖集群，使得原本由各館館員主觀分類的圖書館常見問答能夠透過自然語言處理技術對於問答中使用的關鍵語彙關聯進行分析，其結果依據詢問意圖進行分群歸屬所獲得的分群標準也較趨於一致，人為主觀認知差異大幅降低。而那些具有共同語意重心（即高權重的關鍵詞）的意圖集群，也能因關鍵詞權重的高低（語意重心）加以區分成不同的意圖集群，有助於應用在幫助讀者縮減問題範圍，或是能得知具有共同語意重心的意圖集群其大致上的詢問方向。利用Word2Vec 和維基百科分類架構，進行查詢詞彙（同義詞）擴展，也有初步的成效，讓未來實作參考對話機器人中讀者詢問語彙不再受限訓練語料庫語彙，再者透過DBSCAN分群演算法重新建立的群集，其群集產生對應的關鍵詞彙向量亦可視為該意圖群集特徵，對於後續參考諮詢機器人產生對應對話結果有一定依循方向。

伍、結論

從所蒐錄各圖書館網站整理公開的「常見問答集」發現，不同圖書館對於讀者提出的「問題」因為思考角度與觀點不同，所以對於問題類別定義用語與歸屬有不小的差異，造成「問題」歸類紛亂，使得相同或類似的問題與回應因為各館視角不同造成問題類別歸屬認定的落差，導致讀者透過圖書館提供的常見問答集尋找答案時經常遭遇困難。

在利用自然語言技術處理文本資料，關鍵詞彙擷取的正確性對於後續一些應用（例如文本分類、分群）的影響相當明顯，為補足CKIP Tagger斷詞器詞彙的不足，實驗中我們利用問答集語料中取得圖資專有語彙並輔以N-Gram計算共現詞彙，大幅修正CKIP Tagger斷詞器因辭典詞彙不足導致斷詞時中文短詞的問題。針對查詢詞彙（同義詞）擴展實驗中導入Word2Vec和維基百科分類詞語架構來輔助，再利用相似度分群演算法進行自動意圖分群，能夠有效地分析讀者提出的問題意圖，並從訓練語料中搜尋對應的答案進行回應。此外，本研究還探討了相似句法結構下不同詢問目標的意圖集群，發現從中擷取不同的詢問目標能夠改善答覆並限縮答題範圍，使之更加具針對性。

而以目前能從網路上爬蟲所獲得的圖書館問答集體量較小，僅擷取4,975個問答，現今有些圖書館漸漸有提供線上參考諮詢服務（如messenger服務），可能由各館負責同仁直接回覆，不一定有存檔歸納整理。讀者通常透過這類線上即時參考諮詢服務詢問各館負責「小編」希望獲取即時解答回覆，未來或可從這類互動平臺或類似本研究建置的Linebot蒐集更大量的對話資料來擴增訓練語料，但此類語料可能僅有「問」（讀者提問）與「答」（館員回覆），因為線上即時互動對話，其他標註欄位建置要由人工進行建立較為困難，但若能擷取這類互動諮詢平臺的資訊，對於建置對話機器人所需問答語料集的擴增應有極大幫助，所以透過「問」與「答」文本資料中關鍵詞彙自動識別讀者詢問意圖也是本文希望達成的目標。

總體而言，本研究目標是希望藉由目前從各網站蒐錄的讀者常見問答集進行訓練學習，透過自然語言處理技術來判斷讀者提出問題的意圖，以期後續自動參考諮詢機器人能精準地提出回應，協助圖書館對於讀者問題進行解答，並且能夠透過不斷的學習來提高準確度和效率。未來，我們將繼續強化特徵詞彙向量擷取與文本相似度計算來提升讀者問題意圖分析成效，進而讓圖書館參考諮詢機器人能更準確的回覆讀者問題。

誌謝

本研究得到科技部「基於自然語言處理技術整合維基百科（wikipedia）之圖書館藏查詢參考機器人建置與使用評估計畫」支持，計畫編號：MOST 109-2410-H-030-077。

參考文獻

王明玲、杜立中、曾彩娥（2011）。國家圖書館數位參考服務之使用研究。國家圖書館館刊，*100*(1)，65-98。【Wang, M.-L., Tu, F., & Tseng, T.-E. (2011). The use study of the digital reference service from the National Central Library. *National Central Library Bulletin*, *100*(1), 65-98. (in Chinese)】

王愛珠（2007）。參考諮詢服務的發展：從傳統、線上到線上合作。臺灣圖書館管理季刊，*3*(4)，28-40。【Wang, A.-C. (2007). The development of the reference services—From tradition, online to online cooperation. *Interdisciplinary Journal of Taiwan Library Administration*, *3*(4), 28-40. (in Chinese)】

李金霙、高淑容、劉翠瑤、朱怡蓁（2022）。應用聊天機器人於COVID-19病人照護之成效。北市醫學雜誌，*19*(1)，10-15。doi:10.6200/TCMJ.202010/PP.0012【Li, C.-Y., Kao, S.-J., Liu, T.-Y., & Chu, I.-C. (2022). Effectiveness of chat robot in COVID-19 patient care. *Taipei City Medical Journal, 19*(1), 10-15. doi:10.6200/TCMJ.202010/PP.0012 (in Chinese)】

吳美美、龐宇珺（2011）。虛擬參考服務平臺功能探究。圖書資訊學研究，*6*(1)，1-30。【Wu, M.-M., & Pang, Y.-C. (2011). A prototype for virtual reference system. *Journal of Library and Information Science Research*, *6*(1), 1-30. (in Chinese)】

林妤蓁、高東慶、李祐丞、林紋如、洪名呈（2020）。AI聊天機器人在澎湖旅遊諮詢之開發與應用。島嶼觀光研究，*13*(2)，50-74。【Lin, Y.-C., Kao, T.-C., Li, Y.-C., Lin, W.-J., & Hung, M.-C. (2020). Development of AI chatbot and its application in Penghu tourism consulting. *Journal of Island Tourism Research*, *13*(2), 50-70. (in Chinese)】

林靖文、謝寶煖、莊馥瑄（2013）。以科技準備度與科技接受模型探討

公共圖書館使用者運用數位服務科技之意願。圖書資訊學研究，*7*(2)，81-122。【Lin, C.-W., Hsieh, P.-N., & Chuang, F.-H. (2013). A study of E-Service technology in public library based on technology readiness and technology acceptance. *Journal of Library and Information Science Research*, *7*(2), 81-122. (in Chinese)】

姚飛、紀磊、張成昱、陳武（2011）。實時虛擬參考諮詢服務新嘗試——清華大學圖書館智能聊天機器人。現代圖書情報技術，*27*(4)，77-81。doi:10.11925/infotech.1003-3513.2011.04.13【Yao, F., Ji, L., Zhang, C.-Y., & Chen, W. (2011). New attempt on real-time virtual reference service—The smart chat robot of Tsinghua University Library. *New Technology of Library and Information Service*, *27*(4), 77-81. doi:10.11925/infotech.1003-3513.2011.04.13 (in Chinese)】

柯皓仁（2004）。公共圖書館之數位參考諮詢服務。臺北市立圖書館館訊，*22*(2)，8-19。【Ke, H.-R. (2004). Digital reference services in public libraries. *Bulletin of the Taipei Public Library*, *22*(2), 8-19. (in Chinese)】

范蔚敏（2020）。聊天軟體機器人技術應用於大學圖書館虛擬參考諮詢服務建置過程與評估研究。圖書館學與資訊科學，*46*(1)，4-31。doi:10.6245/JLIS.202004_46(1).0001【Fan, W.-M. (2020). The implementation and evaluation of chat robot application on virtual reference service of university library. *Journal of Library & Information Science*, *46*(1), 4-31. doi:10.6245/JLIS.202004_46(1).0001 (in Chinese)】

陳宜琳（2019）。國立臺灣師範大學圖書館參考諮詢機器人建置與評估（未出版之碩士論文）。國立臺灣師範大學圖書資訊學研究所，臺北市。【Chen, Y.-L. (2019). *Development and evaluation of a chatbot for reference service in National Taiwan Normal University Library* (Unpublished master's thesis). National Taiwan University, Taipei. (in Chinese)】

蘇小鳳（2004）。淺談即時數位參考諮詢服務中的人：參考館員之角色、工作模式與態度。中國圖書館學會會報，*73*，109-123。【Su, S.-F. (2004). The human element in real-time live digital reference services: Reference librarians' roles, work models, and attitudes. *Bulletin of the Library Association of China*, *73*, 109-123. (in Chinese)】

Al-Zubaide, H., & Issa, A. A. (2011). OntBot: Ontology based chatbot. In

Proceedings of International Symposium on Innovations in Information and Communications Technology (ISIICT) (pp. 7-12). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ISIICT.2011.6149594

Altinok, D. (2018, May). *An ontology-based dialogue management system for banking and finance dialogue systems*. Paper presented at the 1st Financial Narrative Processing Workshop. Miyazaki, Japan. doi:10.48550/arXiv.1804.04838

Augello, A., Pilato, G., Vassallo, G., & Gaglio, S. (2009). A semantic layer on semi-structured data sources for intuitive chatbots. In L. Barolli, F. Xhafa, & H.-H. Hsu (Eds.), *Proceedings of the 2009 International Conference on Complex, Intelligent and Software Intensive Systems* (pp. 760-765). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/CISIS.2009.165

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020). A smart chatbot architecture based NLP and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*. doi:10.1145/3386723.3387897

Deryugina, O. V. (2010). Chatterbots. *Scientific and Technical Information Processing*, *37*(2), 143-147. doi:10.3103/S0147688210020097

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226-231). Menlo Park, CA: AAAI Press.

Hussain, S., & Athula, G. (2018). Extending a conventional chatbot knowledge base to external knowledge source and introducing user based sessions for diabetes education. In L. Barolli, M. Takizawa, T. Enokido, M. R. Ogiela, L. Ogiela, & N. Javaid (Eds.), *Proceedings of 2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (pp. 698-703). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/WAINA.2018.00170

Jiao, H., Liu, Q., & Jia, H.-B. (2007). Chinese keyword extraction based on N-gram and word co-occurrence. In Y. Wang, Q. Zhang, H. Liu, & X. Niu (Eds.), *Proceedings of 2007 International Conference on Computational*

Intelligence and Security Workshops (CISW 2007) (pp. 152-155). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/CISW.2007.4425468

Kanagala, H. K., & Krishnaiah, V. V. J. R. (2016). A comparative study of K-Means, DBSCAN and OPTICS. In *Proceedings of 2016 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICCCI.2016.7479923

Kane, D. A. (2016). The role of chatbots in teaching and learning. In S. E. Rice & M. N. Gregor (Eds.), *E-learning and the academic library: Essays on innovative initiatives* (pp. 131-147). Jefferson, NC: McFarland.

Majumder, P., Mitra, M., & Chaudhuri, B. B. (2002, November). *N-gram: A language independent approach to IR and NLP*. Retrieved from https://aclanthology.org/www.mt-archive.info/ICUKL-2002-Majumder.pdf

Nagarhalli, T. P., Vaze, V., & Rana, N. K. (2020). A review of current trends in the development of chatbot systems. In *Proceedings of 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 706-710). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/ICACCS48705.2020.9074420

Rong, Q. S., Yan, J. B., & Guo, G. Q. (2004). Research and implementation of clustering algorithm based on DBSCAN. *Computer Applications*, *24*(4), 45-46.

Rosruen, N., & Samanchuen, T. (2018). Chatbot utilization for medical consultant system. In *Proceedings of 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)* (pp. 1-5). Piscataway, NJ: Institute of Electrical and Electronics Engineers. doi:10.1109/TIMES-iCON.2018.8621678

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433-460. doi:10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In U. von Luxburg, I. Guyon, S. Bengio, H. Wallach, & R. Fergus (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (pp. 6000-6010). Red Hook, NY: Curran Associates.

Young, J. R. (2019, June 14). Bots in the library? Colleges try AI to help

researchers (but with caution). *Edsurge*. Retrieved from https://www.edsurge.com/news/2019-06-14-bots-in-the-library-colleges-try-ai-to-help-researchers-but-with-caution

Library Reference Service Chatbots for Library FAQ Intent Analysis

Shun-Der Chen
Associate Professor
Department of Library and Information Science
Fu Jen Catholic University

Feng-Chia Huang
Graduate Student
Department of Library and Information Science
Fu Jen Catholic University

Introduction

Reference consultation services are notably an essential service libraries provide to readers, offering invaluable assistance in navigating information resources. However, the traditional model of reference service desks demands considerable human resources, making it challenging for libraries to sustain these services. With the advent of the internet and the digital age, there has been a significant shift in how readers interact with libraries, with a growing preference for accessing online platforms to seek information and guidance.

However, insufficient human resources mean that many libraries can only provide asynchronous digital reference services and cannot support readers' needs 24 hours a day. Therefore, introducing chatbots to serve readers is a promising option for libraries. Chatbots can enable libraries to provide reference services to readers without being limited by time and space when reference librarians are in short supply.

Against this backdrop, this study examined Chinese natural language processing, drawing insights from the vast repository of frequently asked questions that libraries encounter. Through meticulous analysis and algorithmic refinement, we aimed to develop an intent-based chatbot prototype that can

comprehend reader inquiries while engaging in meaningful, contextually relevant conversations akin to human interactions.

Chatbots have been widely used in various industries recently, with relevant cases ranging from medical care to the tourist industry. However, the current natural language understanding (NLU) of the core engines of chatbots is primarily based on English, and applying them to Chinese conversations presents particular challenges. Therefore, this study applied Chinese natural language processing. Based on the corpus of library Frequently Asked Questions (FAQs), we constructed an intent-based chatbot that can understand readers' query intentions and simulate conversations with humans in Chinese.

Methodology

This study aimed to improve the accuracy of Intent-Based Recognition in determining the intent of a user's query by chat robots. To do so, we proposed a method to mine the structure or pattern of a user's query from a library's "Frequently Asked Questions" (FAQ) corpus. During the training corpus preprocessing, we first utilized the CkipTagger lexer developed by Academia Sinica to preprocess Chinese lexical disambiguation and lexical annotation. Furthermore, to allow readers to ask questions without being restricted to a specific vocabulary, this study extended the vocabulary by including words from Wikipedia entries, idiomatic phrases in graphical domains, high-frequency N-gram collocations, and related words using the Word2Vec model. This enabled readers to ask questions using words with similar meanings. Next, we combined the TF-IDF algorithm for vocabulary weighting to understand the semantic focus of the sentences while extracting the readers' question feature vectors.

Based on the library's "Frequently Asked Questions" (FAQ) corpus, we found no consistent specification for categorizing patron questions across library settings. Patron questions were also characterized by diversity, semantic similarity, irregular clustering structure, and the nature of miscellaneous data. Therefore, the study experiments adopted the principle that "the algorithm should be able to be categorized into different clusters according to the topics and have a certain degree of noise tolerance" to select a suitable

clustering algorithm. To provide the training corpus, a standardized intentional clustering criterion using three algorithms, namely, the Bidirectional Encoder Representations from Transformers (BERT) classification model, *K*-Means, and density-based spatial clustering of applications with noise(DBSCAN) in Transformers, were selected for re-clustering the "questions" in this experiment. By analyzing the experimental results, we evaluated the effect of different methods on intention clustering and summarized the advantages and disadvantages of each algorithm.

Results

The experimental results indicate that the Transformer model typically required a large amount of labeled data for effective training. A lack of sufficient category-labeled data from the Library FAQ collection may be one reason for its less-than-expected performance; furthermore, an uneven number of categories in the collected Q&A corpus caused the model to predict more diverse categories. This affected the accuracy of the prediction results. In contrast, the *K*-Means and DBSCAN algorithms were based on the syntactic structure or semantic information of the question itself. They were, therefore, more effective in distinguishing the intention of the reader's question, which was more effective than Transformers regarding the clustering results.

Furthermore, we found that the DBSCAN algorithm did not need to specify the number of clusters in advance. Moreover, it was unaffected by extreme values. It could automatically determine the number and size of clusters according to the parameter settings and could discover clusters of arbitrary shapes. The advantages above make the DBSCAN algorithm superior to the *K*-means algorithm. At the same time, due to the irregular structure and varying densities of the textual data clustering in the quiz set, which made DBSCAN, an algorithm robust against noisy data, a suitable choice to discern the user's intention more clearly.

After evaluating the three algorithms, this study selected the density-based DBSCAN clustering algorithm to regroup library FAQs. By analyzing the syntactic structure and key terms used in patron questions, we effectively solved the inconsistency in question categorization due to librarians' subjective

cognitive differences. Moreover, we were able to more flexibly adjust the degree of rigor of the inquiries in the cluster, i.e., we defined the refinement of the inquiry intention to better satisfy the needs of practical applications.

Conclusions

Past studies building Chinese chatbots indicate they faced similar challenges, including the complexity of the Chinese language, difficulty in collecting and organizing the corpus (particularly the tagged corpus), accurate analysis of conversation intentions, etc. Currently, the volume of library FAQs obtained from web crawlers is relatively small (only 4975 FAQs were retrieved). As such, some libraries are gradually providing online reference and consultation services (e.g., messenger services). These may be answered directly by library colleagues and may not necessarily be summarized in the archives. Readers typically ask "editors" in charge of each library through this type of online real-time reference and consultation service to obtain real-time answers. In the future, researchers may collect additional conversation data from this type of interactive platform or Linebot similar to the one constructed in this study to expand the training corpus. However, this type of corpus may only exhibit "questions" (asked by readers) and "answers" (librarian's replies). This is due to the nature of online real-time interactive conversations. Moreover, building other markup fields manually is challenging. Capturing the information from this type of interactive consulting platform would be beneficial to expand the question-and-answer corpus required for building the dialogue robot. This would enable the automatic recognition of readers' query intention through keywords in the "Q" and "A" text, a goal this study aims to achieve.

The goal of this study was to use the "Frequently Asked Questions" (FAQs) collected from various websites for training and learning to determine the intention of the questions readers ask through natural language processing techniques. This led to the subsequent construction of reference and consultation robots to accurately propose responses to assist libraries in answering the questions asked by the readers. This can improve the accuracy and efficiency through continuous learning. During the experiment, we

proposed solutions to address challenges, such as improving Chinese word breaks to enhance the accuracy of keyword extraction, expanding the Chinese vocabulary, N-gram collocations, and training the automatic clustering of corpus maps, etc. These measures will help improve the accuracy and efficiency of the chatbots, which in turn can enhance the quality of a Library's reference counseling services. Our future research will continue to enhance the featured vocabulary vector extraction and textual similarity calculation to improve the effectiveness of reader question intent analysis. This will enable library reference counseling robots to respond to reader questions more accurately.