教育科技與學習　6：1（2018,1）：1-12

研究論文

關於 IBM Watson 結合主題模式應用在智慧問答系統建置的初探研究

林億雄

臺灣首府大學教育研究所助理教授
E- mail: yhslin@tsu.edu.tw

郭添財

臺灣首府大學教育研究所副教授
E- mail: drkuo@tsu.edu.tw

蔡奎如

國立高雄大學通識中心助理教授
E- mail: kjtsai@nuk.edu.tw

摘要

2011 年機智問答 Jeopardy 比賽，世人驚嘆 IBM Watson 在人工智慧上的大幅進步，根據 IBM 官網宣稱他們將擴展 Watson 的技能，讓它可以寫食譜、設計衣服、預測天氣，還能幫醫生診斷病情。在人工智慧興起的時代，資訊科技因電腦核心運算效率增強、機器學習演算法突破性的應用及大公司 (Google、Amazon、FB、Microsoft) 雲端開發平台的開放，使得過去人們認為無法處理的實務問題，首度出現解決曙光。這些科技發展，正充分展現出新型態資料分析的發展契機。新型態資料牽涉的領域可以包含：教育、醫療、商業、環境與公共行政等領域所收集的傳統資料。本研究依循統計學的設計精神從如何測定、收集、整理、歸納和分析，最後反映資料給出正確訊息，在系統建置研究上，不僅是探討智慧問答，更重要的是建立人、資料及流程的關係。最後，我們將會以一個人機模擬實例作為資訊科技在休閒教育互動的初步展示範例，同時透過人機模擬實例展示說明主題模式確實能成功適切地運用在建置智慧問答系統。

關鍵字：華生超級電腦、主題模式、人工智慧、大數據分析、智慧問答系統

壹、研究動機

1950 年代智慧問答系統因圖靈測試而誕生，該系統距今已經超過 60 年發展歷史。但，在學術與產業界獲得極大關注，則是在 2011 年 Apple Siri 和 IBM Watson 所帶來的成功示範效應。智慧問答系統與搜尋引擎最大差異在於智慧問答系統能夠更準確地以自然語言形式理解描述使用者的提問，並進行精確的匹配答案。相對於搜尋引擎提供的資訊在於模糊匹配的概念，智慧問答系統能更好地理解使用者提問的真實意圖，同時更有效地滿足使用者的資訊需求。智慧問答系統能獲得重視的主因有一方面歸功於機器學習與自然語言處理技術的大幅進步，另一方面得利於像維基百科等大規模知識庫的運用以及處理巨量網路資訊技術的出現。然而，現有的智慧問答系統所面臨的問題至今並沒有獲得完全解決。使用者問句的真實意圖分析、問句與答案之間的匹配關係判別仍然是智慧問答系統效能的兩個關鍵難題。然而，這些看似無法解決的實務問題，因為科技技術大幅進步及開源程式社群的分享精神，出現了解決的機會。有鑑於此，2016 年 Facebook 開源釋出該公司人工智慧實驗室所研發的深度學習模組 Torch，同年 11 月，科技龍頭 Google 開源釋出機器學習引擎 TensorFlow、微軟亞洲研究院也宣布在 Github 上開源釋出了分散式機器學習工具包 DMTK。不久，以 Watson 超級電腦在人工智慧領域著稱的 IBM 加入開源釋出了機器學習技術 SystemML。至此，多家科技大廠，紛紛開源釋出該公司人工智慧領域上的先進技術，希望可以藉由更多使用者提供原創內容與使用經驗來解決智慧問答所遇到的難題。關於處理巨量網路資訊技術，Appier 沛星互動科技創辦人兼執行長游直翰 (2016) 針對人工智慧 (Artificial Intelligence, AI) 和機器學習 (Machine Learning) 做出說明：「**人工智慧**是一種系統透過資料處理、演算法運算傳達資訊的方法，從而產出有意義的資訊，並讓機器擁有比人更聰明的智慧。**機器學習**是人工智慧的其中一個分支，讓機器可以自動學習、從巨量資料中找到規則，進而有能力做出預測。」故此，本研究期許能在智慧問答系統上加入人工智慧機器學習，讓智慧問答系統更能了解人類問句。綜合上述討論，本研究目的如下：

（1）提出新的智慧問答系統核心建置方法，讓電腦更能理解人類對話內容。

（2）將此智慧問答系統運用在休閒教育互動領域，並實際進行人機展示。

貳、文獻回顧與探討

Amazon 首席科學家 Nikko Strom (2017) 在 AI Frontier 大會，闡述 Amazon Alexa 的大規模深度學習基本架構、語音識別、語音合成等內容，Nikko Strom 提到 Alexa 替「雞尾酒派對難題」找到了有效的解決方法。Nikko Strom 認為讓電

腦更智慧的關鍵，不是演算法，而是如何擁有更多的資料。根據 Amazon 的研究發現人的耳朵並非隨時都在蒐集語音資訊，「聽」的時間大約佔 10%，所以一個人成長到 16 歲的年紀，他所聽到的語音訓練時間大概有 14,016 小時。關於這個數據，Nikko Strom 認為一個人要花 16 年的時間來學習 1.4 萬小時的語音，而 Amazon Alexa 系統，大約 3 個小時就可以學習完成。傳統語音識別系統框架主要包括四大塊：訊號處理、聲學模型、解碼器和後處理。語音識別系統運作方式包含 (1) 將文本規範化、(2) 把文字轉換成語音，由此得到語音串 (3) 將語音生成波形，也就是真正的聲音，將聲音播出。本文研究主題為建置智慧問答系統，除包含語音識別系統外，更加入核心專業知識的對話設計部分。核心專業知識的對話部分主要搭配 IBM 公司所提供的雲端應用運算服務。IBM Watson 超級電腦在智慧問答對話設計上可分為三個部分，可分為意圖、實體、對話 (intent、 entities、dialog)。傳統上，這內容方面仰賴領域專家進行設計，本文將採用主題模式來進行分類設計。關於本文所探討的主題模式，主要使用的文字探勘技術為簡易主題模式 (Topic Models)，此方法為一種機器學習可以進行語料文本分析分類，關於簡易主題模式的介紹可以參閱 Blei, Ng, & Jordan (2003) 一文。文本分析技術可以參閱 Mihalcea, & Tarau (2004), Page, Brin, Motwani, & Winograd (1998) 及 Erkan, & Radev (2004) 等文獻。在運用文字探勘於顧客留言資料的研究，林億雄與賴美嬌 (2013) 曾提出使用簡易主題模型應用於 Moodle 平台系統，該研究目的為針對在 Moodle 平台系統所產生的巨量數據運用主題模型進行探勘研究，該研究以 Java Based 程式語言建構的軟體進行文本探勘研究，所研究的文本資料以英文語料庫為主。2014 年為了擴充簡易主題模式對於中文文本資料的分析能力包含中文斷詞、分詞及識別動作，林億雄 (2014[a]、2014[b]、2015[a]、2015[b]) 以 R 軟體先進行中文斷詞與中文分詞動作，並完成簡易主題模式分析中文文本資料的探討。另外，郭添財與林億雄 (2017) 針對大數據時代教育現場所產生的教育大數據，說明文字探勘技術對於教育大數據的分析研究將能提供教育現場工作者實質幫助。

參、研究方法及步驟

本文將使用軟體 R 撰寫耙梳線上數位休閒教育資料的程式碼，對於使用者原創內容的文本清理、分詞動作及詞庫管理也將會一併處理。接續，本文將使用簡易主題模式（或稱為潛藏狄利克里分配，請參閱 Blei et al., 2003）來進行文本探勘的統計分析；最後，我們將結果套入 IBM Watson Conversation，以便讓機器能理解人類的問句。

研究方法及步驟：程式設計將文件資料轉換建立用於 LSI 之字詞 - 文件矩陣。

研究方法：使用 R 軟體相關套件，及撰寫耙梳數位資料程式碼。

研究步驟：首先，將待檢索的文件集中放置程式根目錄，透過本文所設計之程式進行讀檔輸出（程式含關鍵詞判斷、stop word 判斷、一詞多義、多詞一義）自動輸出字詞 - 文件矩陣。如下圖 1A：

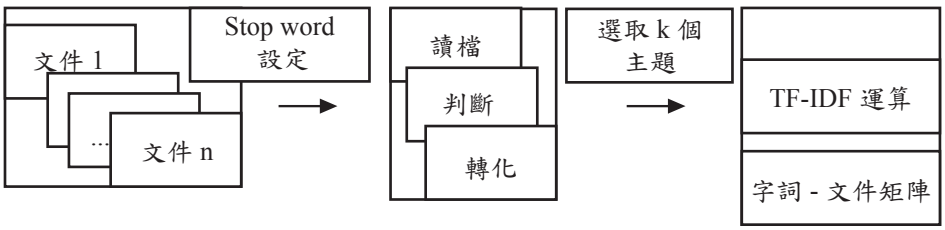

圖 1A 使用者原創內容的文本分析

簡易主題模式（潛藏狄利克里分配；LDA）：

主題模式是一種機率生成模型，近年來被廣泛應用於文字探勘與信息檢索領域。主題模式從被提出以來就受到許多領域研究人員的廣泛關注，該模型除在文字探勘與文本分析上有優異表現外，在計算機視覺領域（Fei-Fei & Perona, 2005; Luo et al., 2015）與社會網絡研究（Jiang et al., 2015）也有成功的應用。關於主題模式的發展，最早從 Deerwester 等人 (1990）提出潛藏語義索引（LSI）而來，該方法一般被視為主題模型最早的起源；LSI 是主題模型發展的基礎，但 LSI 並不是一個機率模型，因此，並不是一個真正的主題模式。Hofmann（2001）基於 LSI，提出機率潛藏語義分析（PLSA），至此 PLSA 才正式稱為一個真正的主題模型。在 PLSA 發表之後，Blei 等人 (2003) 延伸 PLSA 的想法，正式提出潛藏狄利克雷分配（LDA），而此 LDA 正式成為一個更完整的機率生成模型。Blei 等人 (2003) 所提出潛藏狄利克里分配，此為主題模式機率建模最基礎的統計機率模型，也稱為簡易主題模式。目前，有越來越多基於不同目的延伸 LDA 發展出的變形機率生成模型，用於進行文本的無監督主題文字探勘分析。運用主題模式的目的在於發現隱藏在大規模文檔資料中的主題結構。隨著網際網路的快速發展，網路數位文本資料與數量越來越多，單單僅靠人力已經無法全部閱讀和研究所有的文本資料。所以重新思考一種可以利用電腦使用機器學習自動分類的演算法就顯得格外重要且迫切，該演算法可以用來發現和標記大規模網路數位文檔的主題信息。主題模式是一種統計方法，它通過分析原文本中的詞，用以發現蘊藏於其中的主題以及主題隨時間的演變，而且不需要事前對文檔進行標記。換句話說，人力所無法完成的網路龐大數位文檔標記，主題模式演算法能夠進行組織和歸納。簡易主題模式 LDA 的介紹如下：在 LDA 模式中，文字 (terms) 的機率分佈 p（z | N_d）和 p（w | z）均設定為多項式分佈。Blei 等人 (2003) 在 LDA 模式中對於文件中主題的先驗分佈 (the topic distributions in all documents)，與主題中文字的先驗分佈 (the word distributions of topics) 均假設為 Dirichlet 分配。透過統計機

率理論計算，可以得到在共軛分布 (conjugate prior) 的貝氏機率分布性質下，每個主題下文字 (the word distributions of topics) 的後驗分配將是 Dirichlet 分配。最後，透過繁複的蒙地卡羅迭代電腦運算，簡易主題模式可以解出文件中各個主題所佔的百分比，及組成各個主題下的重要關鍵字。下圖 1B 為 LDA 的圖形理論模式 (Graphical Model)，圖 1B 引自 Liu 等人 (2016) 一文中的潛藏狄利克里分配的圖形理論模式。

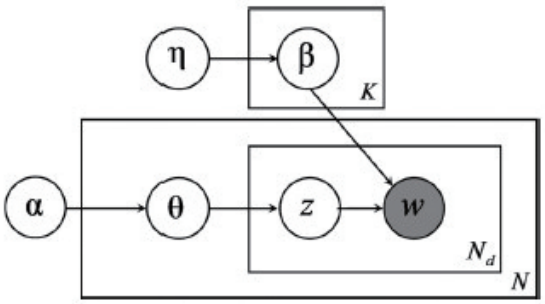

圖 1B 潛藏狄利克里分配的圖形模式

關於將簡易主題模式應用在生物資訊的領域，Liu 等人 (2016) 於其發表的文章引用 Blei (2012) 所發表的例子說明，該例子使用簡易主題模式探討醫學、疾病與計算訊息，在透過簡易主題模式運算後列出其中 3 個主題及該主題中最常出現的前 5 大關鍵字，如下表 1:

表 1 Blei (2012) 列出三個主題中最常出現的前五大關鍵字

主題	Protein	Cancer	Computation
關鍵字	Protein	Tumor	Computer
	Cell	Cancer	Model
	Gene	Disease	Algorithm
	DNA	Death	Data
	Polypeptide	Medical	Mathematical

建置智慧問答系統：互動式對話機器人

關於建置智慧問答系統，在軟硬體的結合方面，我們企圖讓互動式對話機器人可以擁有人類一般的常識能力。在硬體部分包含有微電腦晶片 Raspberry Pi 3 Model B (如圖 2)、感控模組及小型收音麥克風等。其中，使用的微電腦晶片 Raspberry Pi，透過裝入含有作業系統的 SD 卡，Raspberry Pi 本身就將會是一台小型電腦。軟體設計部分為使用 IBM Watson Conversation 雲端應用服務，該人工智慧機器學習應用服務將使用 (1) 將聲音轉為文字 :Watson Speech to Text、(2)

將請求後的答案進行匹配 :Watson Conversation to process the text and calculate a response 及 (3) 將答案從文字轉為語音 : Watson Text to Speech。

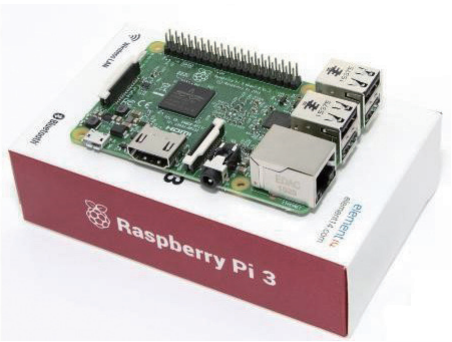

圖 2 微電腦晶片 Raspberry Pi 3 Model B

在智慧問答系統建置中，我們使用微電腦處理晶片 Raspberry Pi 3 Model B 與 IBM Watson Conversation 雲端應用服務連接，這部分是本研究組建對話式機器人最重要的部分。在進行智慧問答系統設計上，軟體方面由 IBM Watson 雲端應用服務提供語音辨識，語音合成及對話設計；硬體方面需要整合微電腦處理器、伺服馬達及感控元件，透過軟體與硬體結合，萌型互動式對話機器人可依據指令進行合宜訓練與學習。關於 IBM Watson Conversation 對話設計如圖 3 與圖 4: (以下圖片來自 IBM Bluemix 官網文件說明)

圖 3 IBM Watson Conversation (一) : Dialog 執行步驟

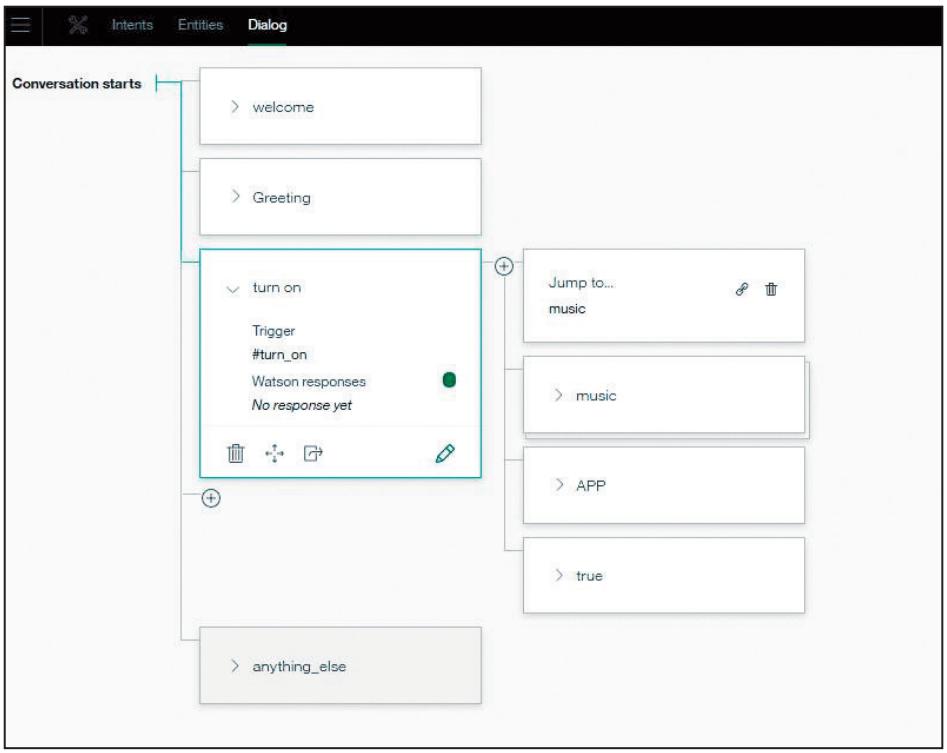

圖 4 IBM Watson Conversation (二) : Dialog 對話樹狀圖展開測試

肆、結果與討論

本研究為資訊科技導入休閒教育互動應用的初探，我們於研究中先爬梳網路上休閒教育的數位文件檔與運用簡易主題模式，分析這些數位文件檔。接續，將簡易主題模式的分析結果，依據所得主題及關鍵字進行分類，並將結果導入 IBM Watson 雲端的互動對話設計 (Dialog)。我們透過模擬實例展示，初步發現主題模式確實能適切地運用在智慧問答系統建置。關於本研究，因為考量科技設備資源花費與大規模運算的時間問題，我們應用市面上流通的現成組件，設計整合出一款萌型互動式對話機器人 (如圖 5 所示)。這款萌型對話機器人主要用以實際模擬人機交談，作為資訊科技在休閒教育互動應用的初步展示。在實際測試展示例子上，虛擬互動式對話機器人可以理解主題內容為「休閒活動」、「天氣預報」、「播放音樂」及「新聞播報」等意義，並且做出相對應的執行動作，例如：使用者可以要求機器人播放音樂，機器人會進一步詢問使用者喜歡哪種類型的音樂，並搜尋該類型音樂進行播放；或者，使用者可以詢問機器人明天天氣如何，機器人會透過網路爬梳氣象資料，告知使用者所在地區的明日天氣預測，並給予使用者在衣服穿搭上的建議。

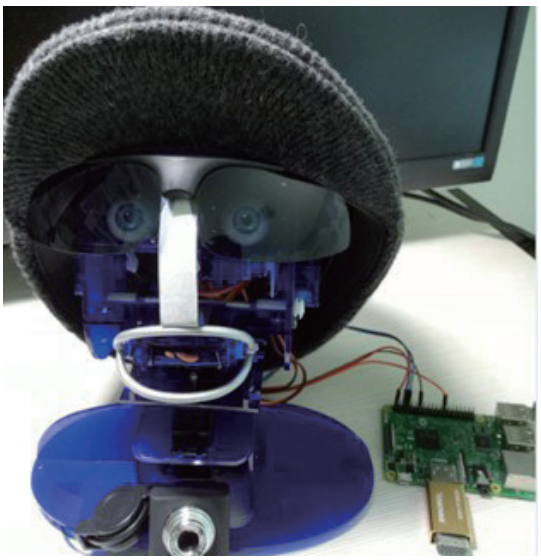

圖 5 互動式對話萌型機器人

目前，我們的研究成果展示出互動式機器人的可能性，未來，我們將運用所實作的萌型機器人與使用者互動交談及爬梳更多文字文件資料，讓智慧問答系統能不斷累積收集人類常識資料，用於改善機器人對於人類問題的理解能力，以利於後續作為資訊科技導入在休閒教育情境學習的有效學習利器。這樣的互動功能，我們認為可以廣泛運用在課堂上協助教師進行課後輔導及提高學生學習興趣與動機。在本文中，我們初步運用簡易主題模式來分析休閒教育的主題，並將結果彙整至 IBM Watson 雲端服務進行智慧問答系統的軟硬體建置。我們在軟體設計上搭配國際大廠 IBM 的雲端應用服務，在硬體上使用一般使用者能拿到的資源，如：微電腦處理器晶片 Raspberry Pi3 model B 及相關控制模組等，透過軟體與硬體結合順利完成資訊科技在休閒教育範例展示。除了在休閒教育互動展示外，我們認為互動式機器人在英文教學上，也可提供英文授課教師課後輔助與幫忙。未來，我們將結合更多新科技產品，例如：HoloLens、Magic Leap 等 VR、AR 相關設備與平台，並且討論在學習過程中導入基於區塊鏈 2.0 (Blockchain 2.0) 智慧合約執行教育素養的可行性。我們期待能建構更強大的智慧問答系統，並配合中文主題模式分析的 R 模組相關套件與中文詞庫管理工具，作為相關產業情境學習最佳的學習利器。

參考文獻

林億雄、賴美嬌 (2013)。主題模型應用於 Moodle 平台系統海量數據之探勘研究。2013 年醫療雲端 - 數位學習 - 第四屆醫療數位學習研討會暨臺灣醫療數位學習學會。嘉義大林慈濟醫院。

林億雄 (2014 a)。主題模式運用在系所專業課程核心能力設定檢驗 – 以健康與美容事業管理學系單一課程為例。2014 觀光、休閒暨餐旅管理學術研討會。臺灣首府大學。

林億雄 (2014 b)。虛擬社群使用者原創內容的統計分析。103 年農業統計暨學術研討會。國立成功大學。

林億雄 (2015 a)。運用大數據分析進行飯店產業客戶服務革新之初探。2015 觀光、休閒暨餐旅管理學術研討會。臺灣首府大學。

林億雄、萬金生 (2015b)。巨量資料的資料探勘與文字探勘。臺南應用科技大學金融投資社群

郭添財、林億雄 (2017)。教育大數據時代的創新發展。臺灣教育，*708*，17-24。

游直翰 (2016)。人工智慧和機器學習有何不同？哈佛 AI 專家告訴你，檢自：http://www.ithome.com.tw/news/101031。

Blei. D. (2012). Probabilistic topic models, *Communications of the ACM, 55*(4), 77-84.

Blei D., Ng A., & Jordan M. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research, 3*, 993-1022.

Deerwester S., Dumais S. T., Furnas G.W., Landauer TK, & Harshman R. (1990). Indexing by latent semantic analysis, *Journal of the American society for information science, 41*(6), 391-407.

Fei-Fei L., & Perona P. (2005). A. bayesian hierarchical model for learning natural scene categories. *IEEE computer society conference on computer vision and pattern recognition* (CVPR’ 05), *2*, 524-531.

Hofmann T. (2001). *Unsupervised learning by probabilistic latent semantic analysis. Mach Learn, 42*(1-2), 177-196.

Jiang S, Qian X, Shen J, Fu Y, & Mei T. (2015). Author topic model-based collaborative filtering for personalized POI recommendations. *IEEE Trans Multimedia, 17*(6), 907-918.

Liu L., Tang L., Dong W., Yao S., & Zhou W. (2016). An overview of topic modeling and its current applications in bioinformatics, *Springer Plus, 5*(1), 1608-1630.

Luo, W., Stenger, B., Zhao, X., & Kim, T. K. (2015). *Automatic Topic Discovery for Multi-Object Tracking*. Paper presented at the AAAI.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web.

Erkan G. & Radev, D. R. (2004). LexRank: Graph-Based Lexical Centrality As Salience in Text Summarization, *Journal of Artificial Intelligence Research*, vol. *22*, 457-479.

Nikko S. (2017). Deep Learning in Alexa, AI Frontiers Santa Clara, CA, USA.

Mihalcea, R., & Tarau, P. (2004). *Textrank: Bringing order into text*. Paper presented at the Proceedings of the 2004 conference on empirical methods in natural language processing.

教育科技與學習　6：1（2018,1）：1-12

研究論文

A Preliminary Study on Applying IBM Watson and Topic Models to the Wisdom Q&A System

Yi-Hsiung Lin

Assistant Professor
Graduate Institute of Education, Taiwan Shoufu University
E- mail: yhslin@tsu.edu.tw

Tien-Tsai Kuo

Associate Professor
Graduate Institute of Education, Taiwan Shoufu University
E- mail: drkuo@tsu.edu.tw

Kuei-Ju Tsai

Assistant Professor
Generation Education, Kaohsiung University
E- mail: kjtsai@nuk.edu.tw

Abstract

In 2011, people were surprised to see a milestone in the development of artificial intelligence represented by the IBM Jeopardy Challenge. The IBM's official website announced that they are continuing to expand Watson's capabilities to incorporate ability to write recipes, design clothes, predict weather, and help doctors diagnose disease. The rise of artificial intelligence in the era of higher compute efficiency, significant breakthroughs in application of machine-learning algorithms, and open cloud application platforms (such as Google, Amazon, FB, Microsoft) assisting people in coping with difficult situations. The technological development provides a glimpse at the potential uses of new opportunities to explore data in educational, clinical, business, environmental, and public administration practices. This study follows the statistics framework of collecting, organizing, summarizing, and analyzing, and making inference from data. The construction of Wisdom Q&A not only to explore whether

followers provide better answers but also build the relationship among human, data and process. A video showing a display of information technology in the leisure industry education are presented in this study, and it proves that Topic Models can be used in the wisdom Q&A system successfully.

keywords: IBM's Watson, Topic Models, Artificial Intelligence, Big Data, Wisdom Q&A System