結合詞向量技術與分群演算法於信用卡商戶名稱辨識
Combining Word Vector Technique and Clustering Algorithm for Credit Card Merchant Detection

Fang-Ju Lee
Soochow University
Dept. of Data Science
rukolee@yahoo.com.tw

Ying-Chun Lo
Soochow University
Dept. of Data Science
ginny880530@gmail.com

Jheng-Long Wu
Soochow University
Dept. of Data Science
jlwu@gm.scu.edu.tw

摘要

透過客戶消費資料萃取相關之使用者行為，是蒐集客戶資訊的方式之一。現行文字探勘的領域中，大多以文本分類之相關研究為主，顯少有文本分群之研究主題。從非結構化之交易消費說明中，找尋字詞之間的關係，運用不同詞向量技術，突破分類需事先區分條件之限制，建立自動化辨識分析方法，提升分群之準確率。在本研究中將以銀行信用卡交易消費說明內容，進行 Jieba 中文斷詞並採用 Word2Vec 特徵值萃取，搭配基於密度分群法(DBSCAN)和階層分群法，交叉組合進行實驗。預測結果以 MUC、B^3 和 CEAF 之 F1 平均值 67.58%較為顯著。

Abstract

Extracting relevant user behaviors through customer's transaction description is one of the ways to collect customer information. In the current text mining field, most of the researches are mainly study text classification, and only few study text clusters. Find the relationship between letters and words in the unstructured transaction consumption description. Use Word Embedding and text mining technology to break through the limitation of classification conditions that need to be distinguished in advance, establish automatic identification and analysis methods, and improve the accuracy of grouping. In this study, use Jieba to segment Chinese words, were based on the content of credit card transaction description. Feature extractions of Word2Vec, combined with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Hierarchical Agglomerative Clustering, cross-combination experiments. The prediction results of MUC, B^3 and CEAF's F1 average of 67.58% are more significant

關鍵字：詞頻-逆向文件頻率、Word2Vec、BERT、餘弦相似度、分群演算法、密度分群法、K-平均演算法、階層分群法

Keywords: TF-IDF, Word2Vec, BERT, Cosine Similarity, Clustering Algorithms, DBSCAN, K-means, Hierarchical Clustering。

1 緒論

隨著科技的發展，台灣已漸漸踏入數位經濟時代，網際網路與電腦的結合，已開始應用在各項領域中，而其中又以金融業在應用上佔有先天之優勢，交易的數位化及電子化等等的商務模式，不僅改變人們的支付行為模式，也縮短了全球商家與消費者彼此間的距離。經由客戶支付行為模式的改變，可獲得更多客戶生活消費資訊，也能提供更優質的客戶服務以及與商戶合作的機會，例如：精準推播客戶喜好的商戶行銷活動、發掘新星商戶等。

本研究旨在運用自然語言處理 (Natural Language Processing, NLP) 、文字探勘及分群演算法等技術，有效的發展文本分群演算法增進於金融應用之發展，落實人工智慧技術於真實場景。本研究期望經由不同的處理技術和訓練模型之方法，從非結構化之交易消費說明中，找尋字詞之間的規則，發展出更為自動與廣泛的分析方法，其透過信用卡交易時，所顯示的商戶名稱，進而了解字詞間關係、語義和對命名實體的理解，並透過交

易的大量消費記錄，以獲得相似之實體商戶群體歸納，提升辨識效果和建立自動化命名實體歸納演算法。

我們將以探討詞向量技術應用於信用卡商戶名稱分群為主題，交易記錄限制於國內交易，並將交易分成 10 種類別，而其中為防範個資外洩之風險，排除了六類含有客戶個資之消費類別，例如，保險、3C 電信通訊、公共事業代繳等等。以小樣本資料進進行研究，尋求如何發展詞向量自動分群，並建立不同之模型並進行反覆驗證和評估，找出最佳模型以確認結果是否符合預期。

2 文獻回顧

隨著文字探勘的技術不斷在精進，機器跨越與人之間的語言障礙。近兩年來 BERT、ERNIE 等非監督式訓練之技術，在語言文字的判斷、語義的相似度、命名實體識別和情感的分析等 NLP 任務更有重大的突破。

2.1 命名實體識別

深度學習方法：隨著詞向量技術的發展，Mikolov 等 (2013) 提出 Continuous Bag-of-Words Model (CBOW) 與 Continuous Skip-gram Model (Skip-gram) 兩個模型結構，用於學習單詞的分佈式表示式，減少計算複雜度。而神經網絡的深度學習發展，始其更有效的處理 NLP 任務，如循環神經網絡 (Recurrent Neural Network) 擴展的 RNN-CRF、卷積神經網絡 (Convolutional Neural Network) 的 CNN-CRF。簡國峻與張嘉惠 (2019) 提出延伸記憶增強條件隨機場域於中文的命名實體擷取，利用門控卷積網路及雙向 GRU (Gated Recurrent Unit) 網路來增強記憶條件隨機場域，提升模型抓取長距離的文章資訊。藉由特徵探勘擷取命名實體的前後詞彙以及前綴後綴詞彙特徵 (Common Before、Common After、Entity Prefix、Entity Suffix , BAPS)，使模型可自動訓練的參數，自動調整詞向量及 BAPS 詞彙特徵，在社群資料中具穩定性和效能，高度依賴特徵設計，對於不同資料集是否有同樣效能值得再研究。而 BERT (Bidirectional Encoder Representations from Transformers) 是近年來較熱門的深度學習方法，許多研究指出 BERT 應用於自然語言上都有獲得不錯的成績。Gong 等 (2019) 將 CRF 層加到雙向 GRU 模型的隱藏層以限制每次的輸出，提高模型的識別性能 (BGRU-CRF model)，並將 Bert Embedding 和 Radical Embedding 串聯在一起做為輸入 Embedding 放入 BGRU-CRF 模型中。王子牛、姜猛、高建瓴與陳婭先(2019) 提出了基於 BERT 的神經網路方法進行命名實體辨識，結合 BERT 和 BiLSTM-CRF 模型，實驗結果均表示中文實體識別以無需添加任何特徵的方式，明顯提升了準確率、召回率和 F1 值。

2.2 字詞相似度

向量空間模型：李琳與李輝 (2018) 研究指出，已提出的非結構化文本相似度計算方法，主要包含基於詞袋 (Bag of Words, BOW) 模型、主題 (Topic) 模型、知識本體和詞向量等方法，然而這些模型方法仍有一些關鍵問題待解決。於是他們嘗試結合依存句法分析和詞嵌入方法，提出一種基於概念向量空間 (Concept Vector Space) 語義相似度的計算方法。曹錫 (2021) 使用 BERT、RoBERTa、ALBERT 三種預訓練語言模型，進行法律判決書案件情境相似檢索實驗，搭配餘弦相似度 (Cosine Similarity)、歐式距離 (Euclidean Distance) 和向量內積 (Inner Product) 三種演算法，以案由分群亂度 (Average Entropy of the Offence-charged Clustering, AEOC) 為指標評估，判斷檢索的優劣程度，其 AEOC 值愈小愈好，代表各分群內的類別蒐集愈收斂。

2.3 Bag-of-Word Model

TF-IDF (Term Frequency-Inverse Document Frequency) 是一種用於資訊檢索與文字探勘的傳統機器學習統計方法，用來評估一字詞對於一個檔案中的重要程度。王美淋(2020)提出結合擷取和萃取兩段式模型方式處理 NLP 任務，其實驗結果較 Transformer 良好。劉賢鈞 (2019) 以 Kaggle 的 Fake News 資料集進行預測假新聞之研究，使用 TF-IDF 找出文本的字詞特徵，並使用線性區別分析 (Linear Discriminant Analysis, LDA) 進行降維，搭配 Random Forests、XGBoost、Naïve Bayes 和羅吉斯迴歸四種分類進行比較，經實驗結果得

知，以羅吉斯迴歸分類方法最佳，準確率高達 96.32%。TF-IDF 雖簡單、容易快速理解，但僅使用詞頻評估文章某一字詞的重要性，缺乏整體性；有時關鍵的字詞出現可能不多，無法表達字詞位置與上下文字的重要性。

2.4 分群模型評估

黃宇翔、王品鈞與方志強 (2017) 考量現今資料屬性為多樣式，為改善 K-means 演算法之處理效能，提出了將資料依數值、類別和順序三種屬性分別做 K-means，以取得較好之初始中心點後再進行組合找到質心。黃郁豪與張芳仁 (2017) 探討在網路資訊眾多的環境之下，如何讓閱讀者更容易獲取有興趣之相關文章，提升讀者點擊意願。研究以 Word2Vec 和 Doc2Vec 模型進行詞向量處理，並取每則新聞前 3%、5%、7%之 TF-IDF 權重較大為特徵關鍵字，與 Word2Vec 相乘轉換後產生新聞字詞向量，採用階層式聚合分群法將文章分群，以 Purity 和 Entropy 評估結果好壞。

3 研究方法

在本節中，我們將描述數據收集和清理、數據註釋、用於解決 NER 任務的模型和學習方法。

3.1 數據收集與清理

本研究以銀行 2020 年度信用卡交易消費為資料來源，銀行依據 VISA 與 MasterCard 國際組織所定義之行業代碼 (Merchant Category Code, MCC) 將資料區分為 15 大類，排除研究限制之含有客戶個人資料和國外消費，預計收集 10 萬筆國內十大消費類別，且不重複的刷卡消費交易記錄之樣本為本研究資料來源。為預防收集之消費類別筆數過少而無法抽樣取得全部消費類別之情形，單一消費類別之母體筆數占總筆數小於 2% 者全數收集，其他則依據母體筆數之比例抽樣收集。刷卡交易除了商戶名稱之外，多數含有分店資訊、使用的支付工具、分期期數或金額等訊息，以下將以各範例分別描述說明。

- 一般商戶
  同一商戶但在不同行銷通路或透過網購平台上架之賣家名稱內容，但商戶傳送交易的中文說明並無統一格式，如：「富邦 momo-EC」、「愛貝金流－momore25」
- 商戶且有分店或分期
  這類型資料常因交易消費說明過長，導致資料傳送時會將資料截斷，造成分店資訊不完整，如：「三澧－MoMo Paradise 復興牧」
- 可使用支付工具之商戶
  屬於非現金交易之掃碼行動支付或銀行與商戶自行開發之行動支付 APP 軟體，如：「全聯門市－PX Pay」、「街口電支－２派克脆皮雞排」。
- 提供自動加值功能之商戶或分店
  現行提供悠遊卡、一卡通 (iPASS) 和愛金卡 (i-cash) 三家公司之自動加值功能，如：「悠遊卡自動加值－比漾廣場摩斯漢堡」。

3.2 分群模型評估資料處理

為了讓資料在格式上能達成一致標準，處理內容含有分期資料的雜訊，移除不必要的資訊，以提升資料品質。在進行特徵植萃取前，本組使用 Jieba 與 CKIP Transformers 先行斷詞。

Transformer 多用於處理連續資料之任務，與 RNN 不相同的是，Transformer 不需要依照順序處理資料，因此減少了訓練時間，在近年的諸多 NER 任務中，Transformer 已取代了舊的遞歸神經網絡模型，迅速成為 NLP 問題的首選模型。

3.3 特徵值萃取

本研究以斷詞工具辨識內容中含有特定意義之名稱所產生的資料集，採用 TF (Term Frequency)、TF-IDF 以及 Word2Vec 三種方法進行文字轉特徵，萃取文本中關注的成分，並將這些詞句轉換為詞向量，以及 BERT 中文預訓練模型技術，計算其相似度供分群模型建立使用。

- TF、TF-IDF

依斷詞後的字詞，TF 採用計算字詞在消費說明內容中出現的次數、TF-IDF 以評估字詞在消費說明的重要程度產生關鍵詞，分別建立 TF 和 TF-IDF 不重複的字詞向量，分別產生

特徵矩陣，供後續計算每筆記錄之間的相似度。

- Word2Vec

具有考慮上下文之特性，將字詞投射在向量空間，其訓練模型有 CBOW 和 Skip-gram 兩種，本研究採用 CBOW 模型架構，給定一個商戶名稱的前後鄰近的交易消費說明字詞，預測商戶名稱出現的機率。

- BERT

本研究採用 Cui 等人(2021) 提出的 Chinese-MacBERT-Base 預訓練模型，輸入資料集之每一筆商戶交易明細，訓練出每一筆 768 維詞向量的記錄，計算兩筆記錄之間的相似度，產生相似矩陣。

3.4 資料訓練與模型建置

本研究以四種特徵值方法分別計算消費說明資料彼此之間的相似度，依相似矩陣轉換成距離特徵資訊，作為密度聚類演算法(DBSCAN) 、 DBSCAN + K-means 以及 DBSCAN + 階層分群法三種演算法之分群基準，將資料分成數個群集，目標找到群內差異小、群外差異大之群集，並配合特徵值萃取之技術，進行訓練與建立模型。

3.5 評估指標

本研究以共指消解作為評估方式，共指消解，是將文字中指向同一 Entity 的詞語劃分到同一個等價集的過程，其中被劃分的詞語稱為表述或指稱語（Mention），形成的等價集稱為共指鏈（Coreference Chain）。在共指消解中，指稱語包含：普通名詞、專有名詞和代詞，因此可以將顯性代詞消解看作是共指消解針對代詞的子問題。研究使用任務中最常使用之評估指標包括 MUC、B^3、CEAF 作為評估方式。

- MUC

MUC score 計算將預測的共指鏈映射到標註的共指鏈所需插入或者刪除的最少的鏈接數量，但 MUC 的缺點為無法衡量系統預測單例實體的性能。

- B^3

B^3 算法可以克服 MUC 的缺點，該算法主要是對每個 mention 分別計算 precision 和 recall，然後以所有 mention 的平均值作為最終的指標。

- CEAF

CEAF 是一種基於實體相似度的評估算法，相比於前兩個評估指標的算法更加直觀的表現評估共指簇劃分的好壞，就是對應地比較每個共指簇劃分。

3.6 相關參數設定

資料點之半徑距離會影響分群個數，而分群個數會直接影響結果。在 K-means 和 AGC 需事先設定群組個數，其分群數則由 DBSCAN 分群演算法而來。DBSCAN 依據不同之半徑距離Ɛ、在Ɛ之內最少有 1 個資料個數 (MinPts = 1)，計算可歸納為 *n* 個分群數。本研究以驗證集資料進行調參，並經由觀察不同特徵值方法之分群數遞減變化。下表為本研究所設定各特徵值萃取方法之參數。

特徵值方法	參數設定
TF	單詞在消費說明內容中出現的次數。
TF-IDF	重新計算 IDF 權重，若文檔不含關鍵詞時，無需對 IDF 做平滑 (不考慮分母為 0 的情形)。
Word2Vec	採用 CBOW 的方式，根據目標字的左右 5 個字進行預測，將訓練出對映到 300 維度空間的詞向量，迭代次數設定為 50 次。
BERT	Model 和 Tokenizer 採用 Chinese-Macbert-Base 預訓練模型，訓練出每一筆消費記錄之 768 維度空間的詞向量。

表 1. 特徵值萃取參數設定

分群演算法	參數設定
DBSCAN	半徑距離Ɛ之內最少有 1 個資料個數。Ɛ之範圍為 TF = 1.1~2.9，每次調整 0.2、TF-IDF = 1.1~2.9，每次調整 0.2、Word2Vec = 1.1~3.8，每次調整 0.3、BERT = 1.1~2.9，每次調整 0.2。距離計算方式使用預設值 Euclidean。依據上述參數設定，獲得 n 個 clusters

分群演算法	參數設定
DBKM	依據 DBSCAN 所以獲得 n 個分群數為參數，距離計算方式採用預設值 Euclidean，其隨機種子為 0
DBAGC	採用 CBOW 的方式，根據目標字的左右 5 個字進行預測，將訓練出對映到 300 維度空間的詞向量，迭代次數設定為 50 次。
BERT	依據 DBSCAN 所以獲得 n 個 clusters 為參數，資料點之間的距離採用 Euclidean 計算方式，群與群之間的距離則使用 Ward 方法。

表 2. 分群演算法參數設定

4 實驗結果

4.1 特徵值方法與演算法分析

依據斷詞工具加三種特徵值方法，以及 BERT 萃取特徵值結果，經由不同半徑距離所獲得之分群數，再依據表 2 分群演算法的參數設定，分別帶入 DBSCAN、DBKM、DBAGC 三種演算法，各自計算評估指標再取四種演算法 F1 平均值之最大值，作為 DBSCAN 半徑距離最佳參數設定，實驗結果分析如下圖 1 至圖 4。

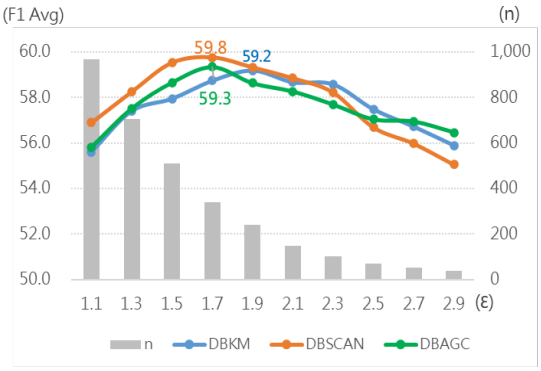

圖 1. TF 特徵值方法之評估結果

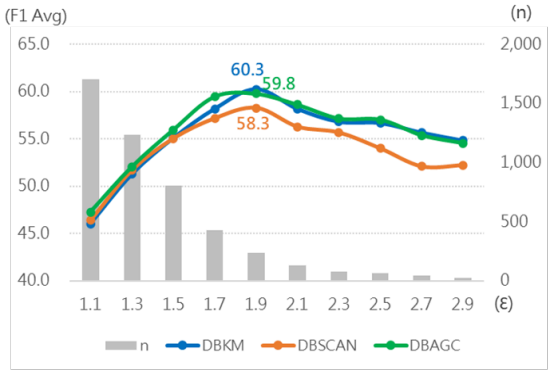

圖 2. TF-IDF 特徵值方法之評估結果

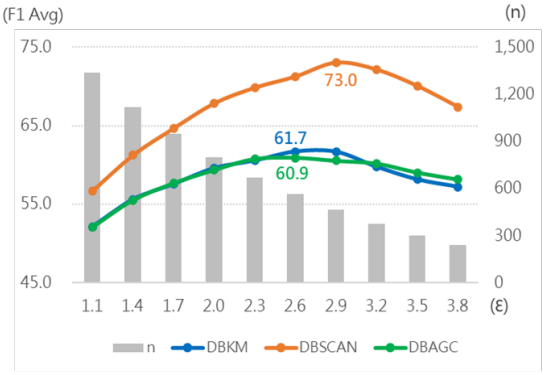

圖 3. Word2Vec 特徵值方法之評估結果

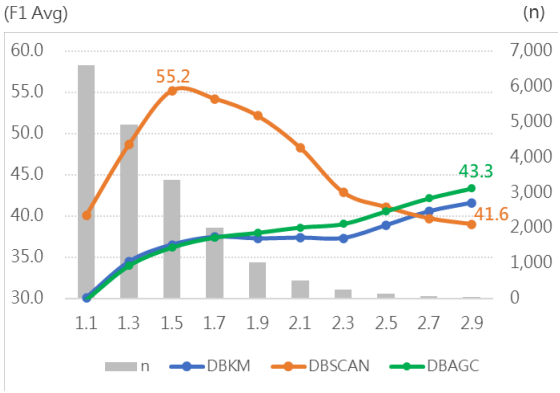

圖 4. BERT 特徵值方法之評估結果

TF 在半徑距離 1.9 時，DBKM 之 F1 為 59.2 最高；DBSCAN 和 DBAGC 在半徑距離 1.7 時均為最高，其 F1 分別為 59.8 和 59.3。三種演算法之 F1 平均值以 59.3 為最高，故 TF 特徵值方法最佳半徑距離設定為 1.7。

TF-ID 在半徑距離 1.9 時，DBKM、DBSCAN 和 DBAGC 之 F1 均最高，分別為 60.3、58.3 和 59.8。三種演算法之 F1 平均值為 59.4。可以發現 DBSCAN 之 F1 平均值計算結果均較

低，且當分群數愈少時，F1 平均值比 DBKM 和 DBAGC 明顯較低，TF-IDF 較不適合 DBSCAN。

Word2Vec 在半徑距離 2.6 時，DBKM 之 F1 為 61.7 最高；DBSCAN 在半徑距離 2.9 時最高，其 F1 為 73.0； DBAGC 在半徑距離 2.6 時最高，其 F1 為 60.9。雖 DBKM 和 DBAGC 二種演算法之 F1 平均值計算結果相似，無明顯之差異，但三種演算法之 F1 平均值以 65.1 為最高，故 Word2Vec 最佳半徑距離設定為 2.9。

BERT 模型在半徑距離 2.9 時，DBKM 之 F1 為 41.6 最高；DBSCAN 在半徑距離 1.5 時最高，其 F1 為 55.2； DBAGC 在半徑距離 2.9 時最高，其 F1 為 43.3。DBSCAN 演算法對群數之多寡差異較為明顯，DBKM 和 DBAGC 之計算結果較為相似，且可以發現群組數愈少其評估指標 F1 平均值愈高。三種演算法之 F1 平均值以 43.0 為最高，故 BERT 最佳半徑距離設定為 1.7。

BERT 模型訓練以學習完整句子為主，非以簡短之交易特店說明，雖然訓練集實驗結果以 DBSCAN 評估指標 F1 平均值 55.2%最高，DBKM 和 DBAGC 評估指標 F1 平均值均不超過 45%，整體而言此特徵值方法較不理想。

4.2 測試集結果

測試集正確商戶分群數為 264 分群數，其中單一商戶之分群數有 157 個，占正確總分群數 59.5%。由表 3 評估 F1 結果得知。

- 三種演算法的評估指標 F1 平均值與驗證集之實驗結果接近，以 Word2Vec 特徵值方法搭配 DBSCAN 演算法之 F1 平均值 67.58 最高。
- BERT 萃取特徵值方式受不同資料集筆數之多寡影響，依訓練集實驗結果之半徑距離進行測試，其分群數僅有 5 群，F1 平均值 36.93 最低。
- MUC 在不同特徵值方法搭配不同演算法之差異較不顯著，主要因 Precision 和 Recall 計算時，單一商戶之個數減 1 後相抵消失，評估指標無法計算含單一商戶之準確率。
- Word2Vec + DBSCAN 之分群數最多，其評估指標 B3 因排除單一商戶之分群，故 Precision 較 DBKM、DBAGC 偏低；評估指標 CEAF 經由調整後，包含單一商戶分群數，其 F1 較 TF-IDF + DBKM 和 TF + DBAGC 分別高出 30%和 20%；其中又以 TF-IDF + DBKM 之 Recall 為 10.56%最

演算法	DBKM				DBAGC				DBSCAN				DBSCAN			
特徵值	TF-IDF				TF				Word2Vec				BERT			
半徑距離	1.9				1.7				2.9				1.7			
總分群數	60				112				145				5			
單一商戶	0				4				93				2			
多個商戶	60				108				52				3			
評估指標	MUC	B^3	CEAF	平均	MUC	B^3	CEAF	平均	MUC	B^3	CEAF	平均	MUC	B^3	CEAF	平均
Precision	85.66	71.89	46.49	68.01	85.52	80.34	45.53	70.47	89.77	48.37	66.32	68.15	86.24	8.65	61.69	52.19
Recall	96.38	61.05	10.56	56.00	93.50	55.75	19.32	56.19	96.32	89.51	36.42	74.08	99.94	99.77	1.17	66.96
F1	90.70	66.03	17.22	57.98	89.33	65.83	27.13	60.76	92.93	62.80	47.02	67.58	92.58	15.93	2.29	36.93

表 3. 各模型於測試集之辨識效果評估結果

低，單一商戶分群數之多寡對 CEAF 影響最為明顯，是影響實驗結果主要原因。

- Word2Vec + DBSCAN 之 F1 平均值較 TF + DBAGC 高出 6.82%，亦較 TF-IDF + DBKM 高出 9.6%，其特徵值萃取之方法對評估結果具有影響力。

綜合整體研究測試結果得知，中文採用 Jieba 斷詞並以 Word2Vec 之特徵值萃取方式，搭配 DBSCAN 分群演算法之評估指標 F1 平均值 67.58%表現最佳；BERT 特徵值方法受限於模型訓練之方法不同，其測試結果表現最差。

5 結論

整體研究結果發現，Jieba 斷詞所訓練之 Word2Vec 模型，以 DBSCAN 演算法經由半徑距離設定，較容易獲得單一商戶之分群數，對於商戶名稱分群之效果最好，有助於自動分群之應用。依據 CEAF 評估指標可以發現，測試集正確商戶分群數為 264 個，DBSCAN 演算法預測之商戶分群數為 145 個，Recall 分數為 36.42；DBAGC 演算法預測之商戶分群數為 112 個，Recall 分數為 19.32；DBKM 演算法預測之商戶分群數為 60 個，Recall 分數為 10.56；正確分群數與預測分群數之差異多少，對於 CEAF 評估指標之 Recall 影響最為明顯。在 B^3 評估指標中，測試集正確商戶分群數中，多個商戶之分群數為 107 個；DBSCAN 演算法預測多個商戶之分群數為 52 個，Precision 分數為 48.37；DBKM 演算法預測多個商戶之分群數 60 個，Precision 分數為 71.89；DBAGC 演算法預測多個商戶之分群數為 108 個，Precision 分數為 80.34。多個商戶之分群數多寡，對於 B3 評估指標具有影響。若單一商戶 (非連鎖) 較多時，其評估指標可能失去之可性度。

References

Bagga, A., & Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In The first international conference on language resources and evaluation workshop on linguistics coreference (Vol. 1, pp. 563-566).

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3504-3514. doi:10.1109/TASLP.2021.3124365

Gong, C., Tang, J., Zhou, S., Hao, Z., & Wang, J. (2019). Chinese named entity recognition with bert. DEStech Transactions on Computer Science and Engineering. doi:10.12783/dtcse/cisnrc2019/33299

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers), 687–692. doi:10.18653/v1/N18-2108

Liu, Y. (2019). Fine-tune BERT for extractive summarization. arXiv:1903.10318.

Luo, X. (2005). On coreference resolution performance metrics. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 25-32)

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020). ERNIE 2.0: A continual pre-training framework for language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8968-8975). doi:10.1609/aaai.v34i05.6428

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.

Xiao, X., Ye, S.-Z., Yu, L.-C., & Lai, K. R. (2017). 應用詞向量於語言樣式探勘之研究 (Mining Language Patterns Using Word Embeddings) [In Chinese]. In Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017) (pp. 230-243)

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

王子牛, 姜猛, 高建瓴, & 陈娅先 (2019)。基于 BERT 的中文命名实体识别方法。计算机科学, 46(S2), 138-142。

王美淋 (2020)。結合擷取式與萃取式兩段式模型以增進摘要效能之研究。

车万翔, 刘挺, 秦兵, & 李生 (2004)。基于改进编辑距离的中文相似句子检索。高技术通讯, 14(7), 15–19。

吳政育, 陳冠宇 (2019)。EBSUM：基於 BERT 的強健性抽取式摘要法. 中文計算語言學期刊, 24(2), 19–35。

张占英, & 王中立. (2003)。中文文本中公司名简称的识别。许昌学院学报, 22(2), 99-101

李琳, & 李辉 (2018)。一种基于概念向量空间的文本相似度计算方法。数据分析与知识发现, 5。doi:10.11925/infotech.2096-3467.2018.0007

施瑞朗 (2018)。基于社交平台数据的文本分类算法研究。电子科技, 31(10), 69-70。doi:10. 16180 /j. cnki. issn1007 － 7820. 2018. 10. 016

胡若云, 孙钢, 丁麒, 沈然, & 谷泓杰 (2021)。基于雙向傳播框架的客服對話文本挖掘算法。沈阳工业大学学报。

郭家清, 蔡東風, 王智超, & 劉浩公 (2007)。一種基于條件隨機場的人名識別方法. Journal of Communication and Computer, 4(2), 22–25。

黃宇翔, 王品鈞, & 方志強 (2017)。混合型資料集的 K-means 分群演算法。電子商務學報, 19(1), 1–28。doi:10.6188/JEB.2017.19(1).01

黃郁豪, & 張芳仁 (2017)。新聞分群方法之比較研究及應用 (Doctoral dissertation)。

簡國峻, 張嘉惠 (2019)。應用記憶增強條件隨機場域與之深度學習及自動化詞彙特徵於中文命名實體辨識之研究。中文計算語言學期刊, 24(1), 1–14.