大數據決勝點：圖控式文本分析平台

Match Point of Big Data: Graphical Control of Text Mining Platform

黃家祚[1*]、梁家隆[2]

[1] 皮托科技股份有限公司 軟體系統處 經理

[2] 皮托科技股份有限公司 軟體系統處 工程師

摘要：大數據與文本分析是現今各產業、政府機構、學校機關最頭痛的議題。因文本為非結構化資料，無直觀的特徵可循，而現今非結構化資料占資料比率八成以上、隱含的可用知識比結構性化資料還要更多，若能在非結構化資料取得先機，勢必能引領業界、掌握未來趨勢。本文簡要敘述大數據與文本分析的發展概況及分析工具選擇的重要信，並嘗試使用圖控式分析平台 PolyAnalyst 來進行案件應用分析，將文本資料透過轉換量化等統計資訊，並進一步找尋資料的規則與模型。

Abstract：Text mining in Big Data Analytics is challenging in industrial applications. There is no intuitive feature to follow because of unstructured text data. According to research, 80 percent of the data is unstructured. Valuable business information is often concealed in these unstructured data. Thus, unlocking hidden knowledge in unstructured data is very important for making business decisions. This article introduces the trend on text mining for big data. PolyAnalyst is used as demonstration to transfer qualitative data to quantitative data. Rules and models of text data are created based on statistical analysis.

關鍵詞：圖控大數據、文本探勘、PolyAnalyst

Keywords：Graphic control of big data, Text mining, PolyAnalyst

前言

科技日新月異，資料不停地產生，無論是機台參數變化或者是銷售紀錄的不斷寫入，大數據可視為一個非常接近母體的樣本，蒐集到如此龐大的資料，若未透過合適統計方法及大數據機器學習演算法處理，就只是公司的資產。而資料未經資料前處理，結合該領域的專業知識、有效判讀欄位參數對於目標的貢獻度，仍是 GIGO (Garbage In Garbage Out) 無法提供企業效益。

在工業 4.0 時代，大數據分析有著舉足輕重的地位，大數據主要以萃取其隱含知識，並驗證其效益，以確立對企業的貢獻度。常見的資料型態為已定義欄位格式的結構化資料，但其實大數據分析隱含更重要的問題：「非結構化」資料，其主要原因在於資料來源可包含網路、文字檔、客戶行為資訊以及影像檔等。其中網路資訊包含大量的自然語言、網路用語、口語或地方性方言，即時出現又短暫消逝的新興詞彙，一般除了特定的幾種關鍵字外，往往只能透過詞性，分割至單詞的程度而無法繼續執行與分析，了解文本之間的關聯性。資料科學家皆期盼在茫茫的資料，透過數據分析挖掘隱含的知識，短期可作為預測；長期可作為企業思維、營運策略，提升企業競爭力。

什麼是大數據

“Big data is the asset and data mining is the

"handler" of that is used to provide beneficial results." 這段話可以完整地詮釋大數據，大數據只是公司的資產，重要的是如何透過資料探勘將大數據資料變現，挖掘隱含且有用的知識。

資料的重要性

資料量大不代表分析結果就能準確；但資料量不足可能導致並非呈現真實資料的原貌，造成分析上的失真，信賴度不足。如**圖 1** 表示當資料量不足容易造成樣本分配無法表達真實的母體分配，而造成失真。實線表示資料的實際母體分配，黑點為所蒐集的樣本資料點，而虛線即為透過樣本資料點所建立的分配，可知資料蒐集上若資料量不夠，或是未收集到關鍵資料，容易造成資料偏誤、影響後續分析。

大數據常見痛點

1. 耗時的資料前處理

在瞬息萬變的時代，每分每秒都有資料不停地產生，而資料科學時代是講求時效性與即時性，當企業或資科學家收集到資料時，應立即分析並提供即時的決策。資料前處理通常約占總大數據分析時間 80%，若資料前處理階段不夠即時，或資料前處理完成後的資料分析不能確實的反映現況，仍是 GIGO (Garbage In Garbage Out)，容易造成徒勞無功的情況產生，而在無效益的情況下投入極大的成本。

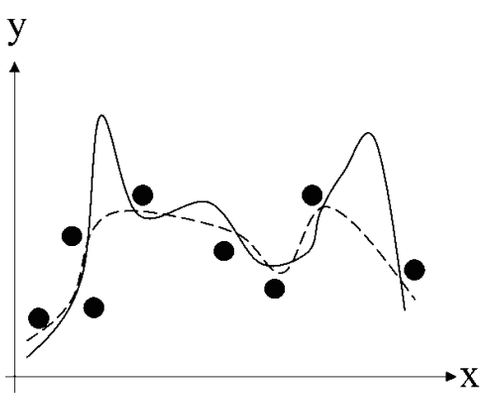

圖 1 資料量不足之影響

2. 預測與分類模型規則難尋

資料前處理常見的痛點 GIGO (Garbage In Garbage Out) 解決後並非一勞永逸。其實現實中就算擁有完整、乾淨的資料，若沒有適合資料分析演算法、手法、甚至是參數設定的錯誤，最後的結果仍是徒勞無功。

若資料科學家對於演算法具有參數設定的知識，由數據中萃取適合規則或預測模型又是另一大課題，尤其在模型的選擇上，通常推論規則驗證週期非常長。

3. 文本分析的挑戰

隨著資料探勘技術日漸純熟，文本分析更為重要，文本分析是一種對非結構資料搜索、指定關鍵字詞、語句的過程。搜索大量文本後，可以建立出這些文本之間的各種詞關聯規則、時序性。舉例來說：想確認公司的服務品質，要求客戶給予回饋意見作為例子，當然一般使用數值分析即可獲取大部分所需的資料與資訊，但若同時需要複雜、無結構化的問題來獲取回饋。此時，若是人工的方式進行回饋的分析則會相當費時與費力，是對於回饋中推導出通用準則是個艱鉅的挑戰。

因文本為非結構化資料，非結構化資料亦為無直觀性的可循，而現今非結構資料占資料比高達八成以上如**圖 2**，甚至更高，且非結構資料隱藏的有用知識比結構性資料還要更多，且非結構性資料的文本所需的資料前處理比結構性資料更為複雜及高難度，往往需要投入更多的人才成本與

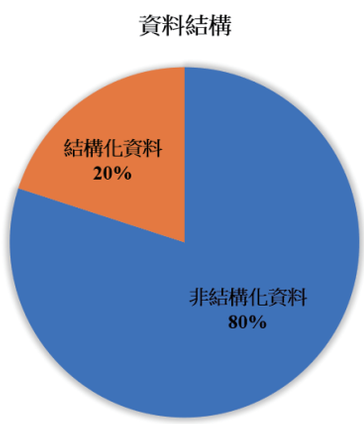

圖 2 現今的資料結構

時間，將其做有效分析才能達到目的。

文本分析概述

文本分析有許多不同的方法，但大多數並未能成功地應用到實務上。今日所有可用的資訊一大部份是以未結構化的形式存在，如書本、雜誌文章、研究論文、產品手冊、備忘摘要、電子郵件，網際網路的網頁等，全部都包含了以自然語言型式的文字資訊。

文本分析技術能夠自動地抽取一段文字的意義，以簡明表列其中最重要概念的方式來儲存分析結果。這個程序將提供一個新而有效的機制來瀏覽文章，自動產生文件的摘要，文字的分群及分類，同時擷取出自然語文的資訊，達成這些功能就能完全應用在日常處理語文資訊的作業上。

將文本自動分類到各不同優先順序主題群組下的功能，在資料庫中或網路上搜尋時增加關聯性和精準性的語義資訊擷取能力，可節省數以百萬計的人工小時。將表達最新市場狀況，可協助公司與競爭者在行銷管理策略的效率上。這些功能與自然語文的資訊擷取能力全部結合在一起，將創新下一世代強力而智慧的知識組織解決方案。

文本分析技術

傳統技術是以特定方式在個別語句的層次上分析，目的是在組成一個語句的重要單字間，產生結構化的關係形式來表達語義。

人工智慧的自然語言人機溝通領域，專門著重於文字的自動化處理。已經產生的語意規則集合僅能良好地運作於開發它們的主題上，因此所執行的分析非常依賴其領域的背景知識。這也暗示人類專家必需介入一主題的語義規則開發階段。這種方式可以針對只專門應用在單一領域的專家系統，但是如果要成功地分析任意領域的文字時，我們需要利用更普遍的演算法。

而現今可應用處理無結構化文字的方法：類神經網路 (Neural Networks, NN)，是以人工處理的媒體元件連結，模仿人類大腦神經單位實際處理

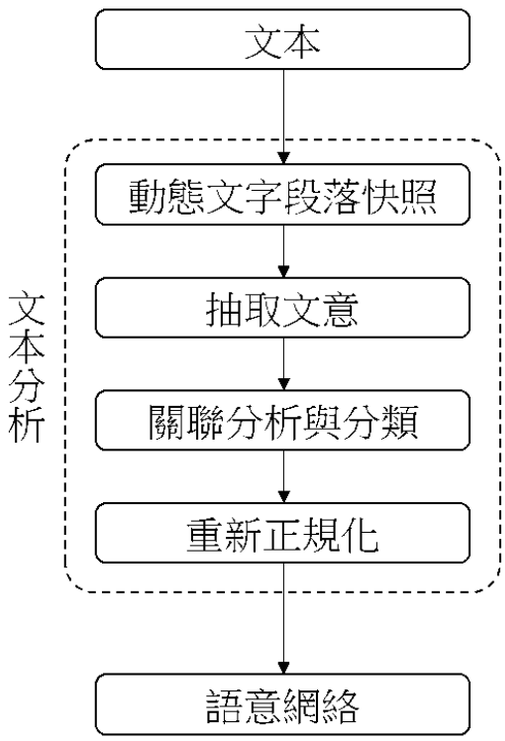

圖 3 文本分析流程圖

資訊的同性質方式而研發出來的。

文本分析技術流程如**圖 3**，透過動態文字段落快照將非結構化文本資料進行斷詞，抽取斷詞中的文意，挖掘關鍵字詞、關鍵字之間關聯性、詞量化、詞分類、時序性 … 等分析，建立語意網絡。語義網絡代表文件分析語言學方面精確和簡明的描述，完全自動地萃取出文件的語義網絡，而不需由人類專家事前開發與特定主題相關的字典。使用者不需提供任何與主題有關的背景知識。此系統會自動地取得這個知識。

產業應用面 - 汽機車業文本分析為例

大數據分析應用層面多元，可於各產業結構化與非結構化資料，目前已成功導入半導體、鋼鐵業、製造業、銀行業、服務業、政府機關、學校、國家研究單位 … 等。以汽機車設計優劣文本分析為例，根據網路論壇廣大使用者對於九大汽機車商，分別探討八大裝置：煞車、加速、引擎、手把、燈、發動、避震、離合器之優劣。文本分析最強

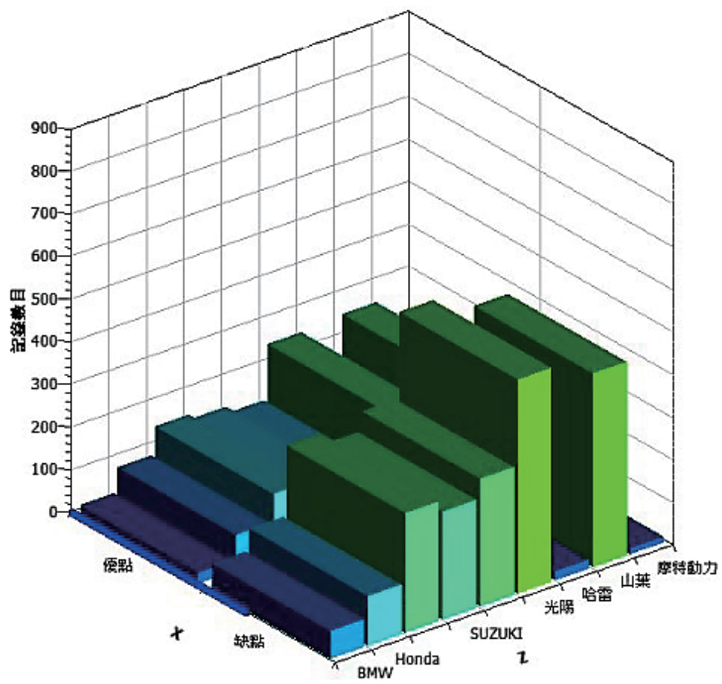

圖 4 使用者需求與優劣分析

大的地方在於能夠簡易且有效地分析非結構性的文本資料。

本研究原使用自由分析軟體，遇到上述常見的資料分析痛點外，資料前處理階段經常發生當程式邏輯更改或是錯誤時，舊有程式已改變資料面貌，此時需重新讀取原資料；或是在執行大型專案時，團隊工程師分別處理不同部分資料，在銜接上難以合併，且知識難以傳承等，導致專案耗時、難以完成。

避免上述問題，研究過程改使用圖控式分析介面 PolyAnalyst，可以有效降低資料分析的難度，透過該分析平台從資料匯入、資料前處理、分析、產生可視化報表等皆可在平台完成。運用 AI 自然語言之分析演算法，與 NN(Neural Network) 及 TF-ITF 詞頻分析法來進行資料分析，並關聯方法採用詞頻矩陣來進行詞與詞之間的關聯分析，再採用語詞分類方法來進行語詞歸類。如**圖 4**，有效挖掘客戶需求與市場趨勢，分析各汽機車廠與功能優劣之差異比較。

PolyAnalyst 與分析軟體差異比較

首先圖控式分析平台可克服上述的大數據常見痛點。

1. 耗時的資料前處理

使用圖控式 PolyAnalyst 可快速建立自動化資料前處理流程模型，能夠讓初學者快速上手，學習上與操作上都能縮短大量時間，有效降低分析上的時間成本以及人力訓練成本，進而快速完成資料前處理。

2. 預測與分類模型規則難尋

PolyAnalyst 已大量參數化及模型化，可建立不同模型及萃取規則縮短驗證週期，資料科學家可快速的同時執行多種演算法進行比較，可由數據中找尋相關規則及建立預測模型 (如製程黃金路徑或金融預測模型)，進而找尋適合資料的演算法，可有效降低推論規則與驗證的時間成本。

3. 文本分析的挑戰

Polyanalyst 可提供多達 21 種文本分析演

算法，使用 AI 的自然語言之分析演算法，與 NN(Neural Network) 及 TF-ITF 詞頻分析法來進行資料分析，並關聯方法採用詞頻矩陣來進行詞與詞之間關聯分析；在採用語詞分類方法來進行語詞歸類，最後可利用 OLAP 方法或一般圖表來進行詞類歸納進行事件總和計算，達到分析預測的能力。

市面上自由分析軟體 R 與 Python 相似度極高，主要可讓資料科學家透過不同程式達成所需的表現模式，對應的對於程式函數、參數設定的熟悉度以及程式撰寫的細心度要求甚高。導致 R 與 Python 學習門檻高且需要相當大的人力成本。另外介面上操作：Python 通常需要另外下載 IDE 才能方便操作，R Code 的介面設計上簡易，但缺乏可視化，而 Polyanalyst 透過物件拉取式的方式撰寫，具有流程的概念，且操作簡易。建模時間，PolyAnalyst 使用圖拉式建模快速，且具有 AI 自動搜尋功能，縮短建模時程，而 R 與 Python 耗時長，且所需傳承與知識管理成本高。比較如**表 1**。

PolyAnalyst 可多元結合，另外支援之功能如下：

1. 可支援於 GIS 空間分析、影像辨識，可結合 R、Python，高彈性分析平台。
2. PolyAnalyst 具有排程功能，可即時、快速獲得不易取得的隱含資訊。
3. 提供互動式 Web report，輕鬆分享於手機、電腦、平板系統。
4. 可支援 Windows、Linux 作業系統。

表 1 圖控式平台與分析軟體差異比較

功能	PolyAnalyst	Python &R
介面比較	圖控式	純程式
讀檔	點選式，可預覽	需指定路徑，不可預覽
瀏覽資料	表格標準化	無表格化
資料前處理	簡單且部分自動化	程式且需下載套件
建模	AI 自動搜尋	程式且需下載套件
驗證	圖拉式	程式
結果可視化	內鍵直接計算	需另外執行
學習門檻	低	高
人力成本	低	高
功能	PolyAnalyst	Python &R

結論

工業 4.0 大時代來臨，從自動化進入智能自動化的工業大革命，已經不單是利用數據與統計管理產線，應利用邏輯與人工智慧管理的時代。將既有的資料有效變現，挖掘從未發現的有用知識，創造新價值以提升競爭力。

且台灣產業正面臨嚴峻人才短缺問題與老師傅高齡化及需知識傳承的問題，可透過 PolyAnalsyt 圖控式建模，以流程節點來解釋資料分析的流程與目的，可保留模型操作歷程與知識思維，減少用戶的重複性工作。

可應用於各產業結構化與非結構化資料，目前已成功導入半導體、鋼鐵業、製造業、銀行業、化工業、材料業、能源業、服務業、交通業、組裝業、政府機關、學校、國家研究單位 … 等。

參考文獻

[1] 皮托科技，“資料科學與大數據文本實務應用-以 PolyAnalyst 為工具”，2019 年 6 月