【專題二】

臺灣證券交易所「監理科技」導入大數據與人工智慧之發展及應用

張益輔（臺灣證券交易所 上市一部 副組長）

壹、前言

證券市場是國家經濟發展的櫥窗，健全的證券市場除可提供企業、政府籌措長期資金，帶動國民經濟發展，更可提供社會大眾適當的投資管道，共享經濟發展成果。準此，為了實現證券交易法第 1 條「發展國民經濟，並保障投資」之目標，證券市場監理更形重要，因為唯有在社會大眾的投資獲得保障之後，才能吸引更多投資人參與證券投資，一旦證券交易市場趨於活絡，企業及政府方可在發行市場獲取資金，進而推動經濟發展，提高國民就業。

因此，世界各國證券主管機關及監理單位的目標，不外乎係為提供一個貫徹「資訊公開（full disclosure）」的證券發行市場，暨維持「自由（free）」、「公平（fair）」、「公開（full disclosure）」的證券交易市場。而在監理作業實務上，監理技術如何伴隨法令規章、交易制度及科技發展而與時俱進，就成為監理人員最大的挑戰。

在證券交易電腦化與自動化高度發展的今天，以電腦系統自動處理有價證券價量、投資人交易等結構化資料已不再是問題，真正的挑戰在於非結構化的輿情蒐集及資料庫管理。尤其是大數據時代的來臨，網路資訊管道多元、傳遞速度大幅增加、傳播範圍迅

14

中華民國一一一年十一月十六日出版

速擴展、對證券市場影響日益明顯，證券監理單位如何因應網路資訊的巨量增生及快速更新，以系統化且有效率的方式蒐集、萃取、分析、儲存、警示資訊，使用傳統的資訊系統架構進行分析運算與儲存已不堪負荷。如何導入並應用「大數據（Big Data）」及「人工智慧（Artificial Intelligence, AI）」技術進行輔助，就成為臺灣證券交易所（以下簡稱「證交所」）發行面與交易面監理最大的挑戰。

謹就證交所近年來導入大數據及人工智慧科技的進程，分階段說明如下。應注意者，有鑑於證券市場監理具有高度機密性，衡諸世界各國均不對外公開業務項目、作業標準及技術細節，故以下僅就證交所「監理科技」之核心概念進行說明與分享。

貳、第一階段：建置大數據平台

一、何謂「大數據」

2001 年，Gartner 公司研究員 Doug Laney 在一份研究報告中首次提出「大數據（Big Data）」的概念。「大數據」被譽為雲端運算後新一波科技浪潮，或被稱為「巨量資料」、「海量資料」或「大資料」，係指所涉及的資料規模龐大且複雜，超越一般資料庫軟體工具所能蒐集、儲存、處理、分析的巨型資料，以致難以利用傳統方式處理資料，必須透過新型硬體設備與演算法來管理與分析[1]。

從大數據的定義延伸，一般而言，大數據具有三大特性[2]，分別是：「大量性（Volume）」、「多樣性（Variety）」及「快速產生性（Velocity）」，俗稱 3V，分述如下：

（一）「大量性（Volume）」

是指需要分析的資料量很大。人類在網路上從事的一切行為都可以形成大數據，而其所產生之資料量每年以增加 50% 以上的指數性成長。資料隨時間快速累增，達到 TB、PB、ZB 等級，E-mail、Google 搜尋、圖片、音樂、影片、應用程式、電子商務、社群等網站都有上萬次的網頁瀏覽與點擊紀錄在不斷產生，甚至是後續發展的物聯網紀錄。資料量大幅增加導致傳統的方法已無法消化，必須產生新的資料處理系統，將大量資料進行擷取、處理、分析並轉化成為有用的資訊。

1 丁冠齊，「巨量資料與個人資料保護法之研究」，世新大學法律學研究所碩士論文，2015 年 7 月，頁 11；轉引自張益輔，「證券市場異常資訊偵測系統之規劃與應用」，臺灣證券交易所 105 年度專題研究計畫，2016 年 12 月，頁 7。

2 中華電信研究團隊，「BigData 於證券集中交易市場監理之應用」，臺灣證券交易所股份有限公司 104 年度委託研究計畫，2015 年 12 月，頁 5-7。

（二）「多樣性（Variety）」

是指需要分析的資料除了結構化資料外，還有非結構化資料。結構化資料是可以表示成欄位，且有固定格式儲存在資料庫，如買賣股數、交易金額等；半結構化資料包含結構化與非結構化的資料，例如xml、html、json檔案；而非結構化資料則無法透過預先定義好的資料模型表示，或無法直接儲存於資料庫，例如圖片、音訊、影片、網頁等。研究指出，有95%的資料屬於非結構性資料。然而傳統的資料分析方式處理結構化資料較得心應手，對於非結構化資料則左支右絀。

（三）「快速產生性（Velocity）」

意指資料並非靜態的，而是隨時都會產生。由於行動運算及社群網路風行，使得資料增生速度比傳統企業應用程式快很多，一旦資料增生速度愈快，資料處理、分析的速度也必須得跟上，以即時處理與回應。高時效資料超出傳統資料庫處理極限，須即時分析所擁有的最新資料，並萃取其中有附加價值的資訊，產生有利於決策的分析結果。

由以上這些特性可以得知大數據並非單一技術，而是眾多技術的集合體，而且近年來已經在吾人生活周遭被廣泛運用。最著名的例子，莫過於美國連鎖零售業商場Target針對女性消費族群的購買行為，研發出一套精準的「懷孕預測模型」，這個模型會列出25種孕婦最有可能購買的產品，從女性消費者購買資料的改變，計算出她們的懷孕預測分數。一旦推測這名女性消費者已經懷孕，Target就會立刻寄出相關商品的促銷廣告，締造「比父親更早知道女兒懷孕」的奇聞[3]。

至於國外證券主管機關及交易所運用大數據技術已行之有年，以紐約泛歐證券交易所集團（NYSE Euronext）為例，該集團發現高頻交易帶來一直成長的交易量，以及日益增加的行情資訊，使得傳統資料庫在發掘各類型異常交易時面臨效能瓶頸，故該集團不惜花費5年時間採用IBM Netezza（IBM PureData System）建立新市場監視平台，利用大數據分析技術來偵測各種新型態的非法交易行為，目前該集團每天能監控與處理約2TB的市場交易資

3　中華電信研究團隊，「Big Data於證券集中交易市場監理之應用」，臺灣證券交易所股份有限公司104年度委託研究計畫，2015年12月，頁35；轉引自張益輔，「證券市場異常資訊偵測系統之規劃與應用」，臺灣證券交易所105年度專題研究計畫，2016年12月，頁12。

料，自 2015 年起可監控與處理 PB 級資料[4]。

二、「大數據」在輿情分析之應用

以股市相關之網際網路資訊為例，除公開資訊觀測站之重大訊息係由上市櫃公司自行公告之外，其餘包括新聞報導及社群貼文在內，多屬未經證實之資訊。證券監理單位須將上開網路資訊予以收攏及優化，亦即運用大數據技術進行輿情蒐集和議題管理，進而取得重要輿情資訊，甚可作為監理業務執行之參考。例如：上市部門就特定新聞報導洽請公司加以澄清或說明，以避免誤導股東及投資大眾；監視部門就社群網站之市場謠言，查明是否涉有證券交易法第 155 條第 1 項第 6 款「散布流言或不實資料」等不法情事。

惟應注意者，面對瞬息萬變的網路資訊，加上社群平台的擴散效應，大數據新聞檢索平台及資料庫在功能設計上，必須要能解決下列問題[5]：

（一）非結構化資料的整合：

非結構化資料一直以極快的速度不斷增長，舉凡市場調查資料、網站點擊紀錄、網站流量或搜尋引擎關鍵字趨勢（Google Trend），乃至於聲音（Audio）、影像（Video）或社交媒體訊息（如 Twitter），都有可能成為市場監理的必要資訊。

（二）新興資料源之運用：

包含社群媒體、影音平台在內的新興資料源，乃至於衍生出來的位置資料、開放資料等，應如何取得，進而分析及整合出各式應用，都有可能對監理產生助益。

（三）議題的交互渲染與蔓延：

當同一議題被多則輿情轉載傳播時，即能產生連鎖效應的「共鳴效果（consonance effect）」；另外，同一議題之概念被抽取並建構出其他議題後，即產生「外溢效果（spill-over effect）」。就輿情分析的角度觀察，既然一

4 中華電信研究團隊，「Big Data 於證券集中交易市場監理之應用」，臺灣證券交易所股份有限公司 104 年度委託研究計畫，2015 年 12 月，頁 45；轉引自張益輔，「證券市場異常資訊偵測系統之規劃與應用」，臺灣證券交易所 105 年度專題研究計畫，2016 年 12 月，頁 26。

5 張益輔，「證券市場異常資訊偵測系統之規劃與應用」，臺灣證券交易所 105 年度專題研究計畫，2016 年 12 月，頁 17-18。

個議題可以代表多則輿情，那麼如何透過議題分析機制，從而適時掌握各議題趨勢的篇數總量、情緒評價觀感、社群互動程度，在輿情蒐集上也扮演重要角色。

三、建置「大數據」新聞資料庫及搜尋引擎平台[6]

證交所於 2016 年規劃上開大數據平台之目標，係為建置以網路資訊驅動之視覺化與自動化資訊系統。藉由網路資訊之自動化蒐集、萃取、轉換、儲存等一系列過程，結合標的證券同期間價量走勢與投資人交易，產製視覺化分析介面，進而觸發警示或進一步分析。系統功能設計共區分為三大模組，分述如下：

（一）網路擷取及同步監控平台

針對公開資訊觀測站、新聞網站在內的網路公開資料，運用網路擷取技術，依使用者需求精確切割目標網頁版型，針對網頁名稱、標題、描述……等欄位進行資料萃取，並儲存於資料庫中。

（二）證券市場輿情分析系統

運用廠商文字探勘工具之專利技術與公開資訊蒐集工具作為基礎，進行資料擷取、資料轉換及資料載入之作業。流程如下：

✓ 將所有非結構化資料，轉換為結構化資料

✓ 對資料進行斷字抽詞

✓ 由系統自動建立詞彙關聯、比對及分類

✓ 過濾資料不一致性與重複的問題，提供使用者高可信度與精確性的有效資料

為便利使用者全方位檢視各類主題資訊，完整掌握情資動向，系統運用全文檢索、數位儀表板、分類樹 / 資訊策展、輿情警示及趨勢統計分析報表等功能，在前台首頁以視覺化圖表來呈現。

以「數位儀表板」為例，系統依監理人員所轄管區公司名單，建置客製化儀表板，讓使用者在登入系統的第一時間，即可在首頁檢視管區公司當日

6　張益輔，「證券市場異常資訊偵測系統之規劃與應用」，臺灣證券交易所 105 年度專題研究計畫，2016 年 12 月，頁 52-116。

網路訊息的觸及率（即聲量）、多空屬性（即情緒），再輔以個股成交價、漲跌幅度等參考資訊，有助於使用者在第一時間發掘風險較高或需進一步分析處理的公司、個股或新聞資訊，俾利即時掌握業務焦點。

再以「輿情警示」為例，使用者可設定多個關鍵字詞組合作為告警條件。當系統擷取網路訊息、分析發現符合使用者設定的條件時，即自動啟動告警功能，並以電子郵件等方式通知使用者，俾利監理人員隨時掌握資訊趨勢走向，即時回應處理。

（三）視覺化分析報表

提供使用者多樣化的統計分析報表功能，可根據主題、資料類型、特定議題、熱門檢索詞彙，設定時間區間等條件，系統即自動將數字量化成統計報表，並以視覺化方式呈現。統計報表類型包括「曝光率分析」、「熱門主題分析」、「分類主題關聯地圖／智慧關聯地圖 (KMAP)」、「意見領袖」等，共計十餘種。

綜上，「大數據」新聞資料庫及搜尋引擎平台之監理目標，係整合現有監控的網路資料來源、以大範圍且 24 小時不停機的方式，持續且有效的追蹤目標知識，提供領域相關產業之新聞與異常言論，以利使用者快速掌握輿論資訊，即時追蹤網路訊息，提供管理者迅速有效的決策及回應依據參考。

四、建置關聯分析系統

證券市場監理常有關聯分析之必要。就發行面監理而言，不論是集團企業間投資及持股之關係、關係人交易之金流與物流，乃至於獨立董事之適格性等，監理人員均有查核分析前開法人與自然人彼此間關聯性之必要；另就交易面監視論，舉凡操縱案件中集團成員之歸納、內線交易案件中投資人與公司內部人是否具有關聯性，均為不法交易案件查核之重點所在。

傳統作法上，關聯分析之佐證資料散見於內部申報資料（如上市公司申報之內部人股權交易資料）及外部公開資訊之多種知識庫，乃至於新聞網站、企業徵才網站等均有其參考價值。為便利使用者整合查詢，證交所爰規劃建置關聯分析系統，將前開知識庫及網路公開資訊等非結構化資料，以大數據技術蒐集、萃取並就關聯性加以整合，最後以視覺化圖形呈現，俾供使用者得以最直觀的方式進行查詢與分析。

參、第二階段：導入人工智慧科技

前開新聞資料庫及搜尋引擎平台以大數據技術整合多種網路新聞、重大訊息等資料源，雖大幅減少使用者在不同網站間切換及查找之困擾，惟因中文之複雜性，無可避免衍生出「雜訊偏多」及「個股分類錯誤」等問題，原因如下：

- 上開系統於文字探勘及文章分類之邏輯係採「規則式分類法 (rule base)」，致部分上市公司特殊的簡稱及代號，引發系統錯誤爬取雜訊或無法正確分類。常見上市公司簡稱之誤判情形如「傳奇」、「卓越」、「幸福」、「國產」、「全國」、「新興」、「全台」[7]、「全新」、「冠軍」、「台南」[8]、「大量」等；上市股票簡稱相似造成誤判者，如「大成」與「大成鋼」、「統一」與「統一超」、「台南」與「台南 -KY」、「長榮」與「長榮航」等；上市股票代號造成誤判者，如官田鋼之公司代號「2017」[9]。
- 指標性權值股如「台積電」、「鴻海」新聞資訊篇數過多，經查多屬盤勢分析類訊息，與該等上市公司本身財務業務並無直接相關。
- 部分新聞重複刊載於各大新聞網站，或被轉分享至社群平台，致系統載入之新聞則數虛增。

為克服上開問題，導入以「自然語言處理（Natural Language Processing， NLP）」為核心之人工智慧科技實有必要。證交所自 2017 年與委外廠商共同規劃導入「類神經網絡」、「深度學習分類模型」、「支持向量機（Support Vector Machine，SVM）分類模組」等人工智慧技術，有效提升新聞訊息自動分類之精準度；另導入新聞資料整併功能，以「相似新聞演算法」將相同或相似的新聞進行整併歸類。

在作法上，委外廠商先從證交所提供 8 萬則人工分類的個股新聞資料中提取特徵、機器學習並建立初始模型。嗣後當網路新聞被擷取進系統後，即自動進行斷字抽詞、分析詞組向量，最後賦予個股貼標。上開語意分析系統經由證交所投入領域知識，結合廠商最新人工智慧技術，經由長達兩年的共同研發，2019 年上線時即有 8 成以上的精確度，隨著近幾年訓練資料不斷累增，重複機器學習的結果，精確度仍在持續提升。

7　2017 年 8 月 15 日「全台大停電」事件，造成隔日上市公司「全台」（全名為「全台晶像股份有限公司」，公司代號 3038）在系統內的新聞數量大幅增加。

8　上市公司「台南企業股份有限公司」（代號 1473）簡稱為「台南」，當新聞出現台南市、台南地區等關鍵字時，該篇新聞就會被錯誤分類至台南（1473），導致台南（1473）的新聞雜訊過多。

9　上市公司「官田鋼」之公司代號為「2017」，造成 2017 年度該公司之新聞聲量居高不下，因為系統自動將含有「2017 年」、「2017 年度」、「2017 年 X 月 X 日」之新聞全部歸類至「官田鋼」所屬新聞。

除此之外，證交所自 2021 年起研究以人工智慧技術建構「情緒（利多 / 利空）指標自動分類模型」，運用類神經網絡、文字探勘等技術進行個股新聞「利多 / 利空」分數的自動分類演算，今（2022）年度將上開功能模組整合至現有「新聞資料庫及搜尋引擎系統」，則為導入人工智慧科技的另一成功案例。

肆、推動自動化及流程改善

由於自動化工作流程之成本效益日益顯著，自動化已成為企業韌性與風險管理的全球熱門議題。

證交所近年來積極導入「機器人流程自動化（Robotic Process Automation）」，簡稱 RPA，係模擬人類在電腦介面上按規則自動執行作業流程，善於處理成熟穩定、遵循明確規則及高度手動、重複性或高頻率、多系統切換之流程，代替或輔助人類完成電腦操作，進而有效重置人力配置，降低成本、減低人為誤差、提升效率及資訊穩定性。

以證交所上市一部為例，該部自 2020 年起積極導入流程自動化，截至 2021 年底已完成 11 項流程自動化作業之導入，廣泛應用於管理性報表編製、自動化寄發提醒通知及稽催、緊急通報等各項監理職能，經評估每年可節省同仁作業時間約 604 小時，個別項目節省時間之比例則約 13% ～ 100% 不等。

除了節省作業時間外，RPA 自眾多分散來源自動擷取資料，更可降低數據錯誤風險、縮短龐雜資訊梳理時間、釋放人力時間以投入更需聚焦與專業分析判斷之工作、有效增進監理效率、縮短資訊彙報時間。

伍、監理科技之未來規劃

伴隨證券市場全球化、網路化，金融商品與證券交易制度的推陳出新，以及電腦科技的不斷演進，世界主要證券監理機關包括美國證券交易委員會（SEC）、日本交易所自主規制法人在內，均積極發展「監理科技」。

證交所自 2016 年導入大數據及人工智慧系統迄今已邁入第 7 年，未來除將持續規劃引進新興科技，精進「新聞資料庫及搜尋引擎平台」、「關聯分析系統」之準確率及資料探勘分析之深度、廣度，更將結合並深化監理業務面之實務應用。

舉例來說，證交所刻正規劃「大數據平台」，區分交易、發行、券商監理三大面向，建置可即時取得監理資料之視覺化分析平台。將採循序漸進之方式，第一階段先引進視

覺化分析工具，同步訓練監理人員熟悉資料分析原理與工具操作；第二階段則利用資料科學，研究彙整資料、設計出新的資料集，並將前開資料抽取、轉換及載入大數據平台；第三階段將精進分析工具之應用，讓資訊呈現更多層次，輔助決策面向更深而廣。

證券市場瞬息萬變，隨著市場狀況轉變、法規制度演進以及資訊科技進步，都將對證券市場帶來潛在的衝擊。準此，證券監理單位如何應用新興科技，並以系統化且有效率的方式蒐集、萃取、分析、警示、處理資訊，已成為未來必然的趨勢。

然而「監理科技」之發展並非一蹴可幾，小從輿情監控及交易分析，大到全市場的創新應用，都不是短時間投入大量人力、物力即可成就。如何分階段持續規劃、評估與導入，應屬最具體可行之實作方法。期許臺灣證券交易所導入新興科技及建置數位化環境的努力，能有效強化對證券市場的監理效能與預警能力。