當　代　會　計
第十二卷第一期
民國一〇〇年五月
頁 85-114

Journal of Contemporary Accounting
Vol. 12 No. 1, May 2011
PP. 85-114

運用文字探勘與 XBRL 技術提升企業資訊擷取與整合效益之研究

周濟群[*]　連子杰[**]

摘要：除傳統財務報表資訊外，企業財報附註或企業年報中的非財務資訊等補充性資訊，亦包含許多重要的企業公開資訊，且兩者間應具有資訊整合的綜效。然而，受限於兩者資料格式之不一致，投資人必須透過繁瑣且錯誤率較高的人工處理程序，方能擷取、整合或流通此兩類資訊。因此，市場參與者開始重視如何利用資訊技術提升企業資訊擷取與整合效益。本研究應用設計科學研究方法，探討不同的企業報告資訊應如何整合與塑模，首先運用本體知識概念分類方法與文字探勘技術，處理股東會年報中與企業策略相關之非財務性資訊，以協助使用者有效率地分類閱讀這些非結構化文字資訊。其次則運用可延伸企業報導語言 XBRL 作為財務報表資料來源，以解決傳統財報格式不易處理之問題。最後，在資訊整合方面，則透過連結機制之設計以整合非財務與財務資訊。本研究並運用雛型系統設計法驗證所提出的企業資訊整合架構確實可行，能協助投資人能於短時間內擷取、整合或流通企業資訊，提升決策品質。

關鍵詞：企業資訊整合、概念階層、文字探勘、可延伸企業報告語言、資訊擷取

[*] 國立台北商業技術學院會計資訊系副教授
[**] 勤業眾信聯合會計師事務所查帳領組

99 年 05 月收稿
100 年 01 月接受

Journal of Contemporary Accounting
Vol. 12 No. 1, May 2011
PP. 85-114

Enhancing Effectiveness of Business Information Retrieval and Integration via Text Mining and XBRL Technology

Chi-Chun Chou* **Zih-Jie Lian****

Abstract: In addition to the information provided in traditional financial statements, the supplementary data contained in footnotes and non-financial disclosures in the annual report proves to be decision relevant. They have the potential to provide a synergy for information integration between data sources. However, due to the format inconsistency between two data sources, investors confront inefficient and ineffective process for retrieving, integrating and exchanging information. Hence, market participants are starting to put an emphasis on utilizing information technology to advance information retrieval and integration. This research proposes an "annual report knowledge construct" framework to enhance the effectiveness of data retrieval and integration through the design science methodology. It uses text mining technology for reclassifying the semi-structured or unstructured business strategy-related non-financial data. In addition, financial data in XBRL format is used to mitigate the format inconsistency problem. Finally, this paper develops a linking mechanism to integrate financial and non-financial data. This paper also adopts the prototyping method to prove the concept of the business information integration framework.

Keywords: business information integration, concept hierarchy, text mining, XBRL, information retrieval

* Associate Professor, Department of Accounting Information, National Taipei College of Business
** Auditor, Deloitte & Touche

Submitted May 2010
Accepted January 2011

壹、研究背景與目的

企業報告資訊應該忠實表達出企業經營活動之成果，不僅用以遵循監理機關的強制揭露需求，也是增進企業與投資人間信任程度的根本；而企業報告對於使用者有無資訊價值，則應可由資訊品質特性來加以分析。例如，財務會計學者 Maines, Bartov, Fairfield, Hirst, Iannaconi, Mallett, Schrand, Skinner, and Vincent (2002) 曾歸納數個企業現行報告揭露實務的品質問題，如攸關性、可靠性、可比較性與一致性問題等。Maines and McDaniel (2000)、Hirst, Hopkins, and Wahlen (2002)、Hodge, Kennedy, and Maines (2004) 等則針對現行財報表達格式的可比較性（presentation format comparability）、資訊可取得性（accessibility）等提出質疑。

基於現行企業報告的限制，Pincus (1989) 提出，企業應將投資人視為即將購買公司的人，來發布攸關的資訊。Eccles and Mavrinac (1995) 研究分析企業與投資者之互動行為，以公司經理人、財務分析師與專業投資人此三類使用者為受訪對象，發現此三類使用者最喜歡的溝通媒介中，年報的使用頻率最低，卻被列為第三重要的管道，主要原因為年報內容雖然是所有可獲取企業資訊中最為豐富及完整的，卻限於它的表達形式（以 pdf 或 html 為主的文件），導致無法擴散其應用價值。Plumlee (2003)、Engelberg (2008) 則皆曾提出資訊複雜度與處理成本會影響市場參與者解析公開資訊的效果，例如複雜的稅報資訊或以文字為主的法說會內容等，會導致資訊分析師預測的不效率，甚至產生所謂的「盈餘宣告後股價持續反應」（post-earnings announcement drift, PEAD）。

正如以上諸多學者的研究，過去的企業報告揭露實務，多數僅偏重於要求企業提供遵循性資訊的供給面議題，而忽略需求面如何運用這些封閉式或無法延伸的紙本報告，在愈來愈重視企業資訊供應效率的機構或個別投資人眼中，現行揭露機制的確存在許多企業資訊品質不良的問題（周濟群，2009）。

基於適當地運用資訊科技可以協助改善資本市場的資訊透明度、資訊不對稱、並能大幅降低遵行這些法令的成本、時間和風險的前提（Logan and Mogull, 2003），近年來，為了促進企業報告由封閉式或無法延伸的紙本、文檔的內容（content）與資料格式（data format），同時升級為以開放式或可延伸的企業報告架構，一個名為可延伸企業報告語言（eXtensible Business Reporting Language, XBRL）的國際企業報告共通標準，已為世界各經濟大國所陸續採用，希望藉由此技術標準發展出更符合使用者需要的企業報告揭露模型，以提昇企業報告提供內容的攸關性、即時性、可靠性、格式可比較性以及資訊的可取得性等品質（Bonsón, Cortijo, and Escobar, 2008; 周濟群，2009）。

儘管 XBRL 標準已受到全球各國企業報告規範者的矚目，而學術證據也顯示它確能改善企業報告品質（Hodge et al., 2004; 周濟群，2009）。然而，如何將這些新興互動式資料（interactive data）之應用，擴散至所有決策攸關的企業資訊，卻仍是

一個亟待研究的重要議題（Matherne and Coffin, 2001; Cunningham, 2005; Lara, Cantador and Castells, 2006）。具體而言，XBRL 透過將重要企業報告資訊有系統地貼上標記（tag）的技術，雖然可使得傳統報表被切割成為可搜尋（searchable）、可剖析（parsable）的資料單元，提升企業資訊的資料顆粒度（data granularity），然而非屬於 XBRL 分類標準所包含與定義的資料集合，卻仍無法納入自動化分析的範圍。以臺灣證券交易所 2010 年定版的分類標準架構而言，即僅包含四大財報資料，並不包含任何財報附註或管理階層討論與分析（MD&A, Management Discussion and Analysis）的資訊，使用者仍需由 pdf 格式的財報或年報電子書中，以人工的方式來擷取資訊、進行財務與非財務資訊的整合。

由非財務資訊的資訊擷取（information retrieval）效益來看，以全球發展最完整的美國分類標準架構（US GAAP Taxonomy Framework 2009）為例[1]，在MD&A這個分類標準套件之中，僅僅定義 68 個項目元素（element），且盡皆為文字項目型態（string item type）資料，僅能提供管理階層對於營運結果、財務狀況相關分析之描述性文字資訊，無法進一步地加以分析比較其中的重要資訊（Engelberg, 2008）。而由資訊整合（information integration）的效益來看，無論是IFRS taxonomy 2009 或 US-GAAP 2009 分類標準中[2]，均包含大量的財務報表附註（footnotes）資料，多數亦皆為文字項目型態。這些附註與MD&A等資訊，它們與財報科目之間可能均具有決策相關性，應可構成資訊的相互關聯、整合，但此種關聯卻無法由XBRL現有機制建立連結[3]。例如：在附註或MD&A中，重要會計科目說明下之閒置資產揭露，可讓報表使用者了解企業資產使用或預期使用之情形產生重大改變，經由附註揭露閒置資產之明細項目，可輔助投資人評估未來該資源是否具備處分、出租、或其他重新使用之可能性。

針對國內年報電子書等公開非財務資訊的資訊擷取問題（資訊內容無法充份以 XBRL 格式來儲存與表達），本研究的首要目標即為建立「企業年報非結構化文字資訊分類系統」，以年報資料中的「句」為單位，運用文字探勘技術，並應用企業策略分析中常用的分類方式，輔助使用者更有效率地分類擷取、閱讀、分析、整合這些非結構化資訊。

另一方面，針對資訊整合問題，基於年報或財報附註中的某些文字資訊與財報科目之間可能具備共通的屬性，因而構成資訊的相互參考價值。然而，由於國內現行揭露機制應用兩種不同格式：「pdf年報或財報附註」與「XBRL財報」，因而無

[1] 關於US GAAP分類標準的詳細資料，請參考：http://xbrl.org/TaxonomyRecognition/US%20GAAP%202009/XBRLUS-USGAAP-Country-Summary-2008-10-31.htm

[2] 關於IFRS分類標準的詳細資料，請參考：http://www.iasb.org/XBRL/IFRS+Taxonomy/IFRS+Taxonomy+2009/IFRS+taxonomy+2009.htm。

[3] 現行 XBRL v2.1 規格中，僅提供呈現、計算、標籤、參考來源、定義等五種單一報表上各元素的連結庫（linkbase）機制，無法擷取或表達使用者整合不同報表資訊的需求。

法進行此種參考性關聯，導致這些具有互補性（complementary）、附加性（augmentative）或解釋性（explanatory）資訊間的整併，僅能依賴專業使用者的人工判斷，無法發展成自動化的機制。為了證實非財務資訊與財務資訊間可利用資訊技術加以整合，本研究先根據XBRL技術規範，以臺灣證券交易所公布之「一般行業XBRL財務報表分類標準」為藍本[4]，建置XBRL企業財報案例文件（instance document），再透過連結機制的設計，將「企業年報非結構化文字資訊分類系統」與結構化XBRL財務資訊相互連結，希望協助使用者有效地瞭解並印證標的公司之產業發展與競爭策略，提升其決策品質。

本文後續各節依次為：貳為相關研究，首先探討國外目前有關 XBRL 如何改善企業報告資訊品質之研究或成功案例，其次則整理文字探勘、知識概念架構等相關技術與學術研究；參為研究方法與系統設計，說明系統設計架構，以及如何進行企業年報內容的文字探勘與知識概念分類；肆則以雛型系統說明文字探勘與知識概念分類的實作結果，以及如何建立非結構化資訊與結構化 XBRL 財報間的連結；伍為本文結論與後續研究建議。

貳、相關研究

本節將針對本研究所應用的 XBRL 相關技術、文字探勘與概念分類方法，分別整理其相關概念與重要研究，以利相關研究活動的後續說明。

一、XBRL 相關研究

(一) XBRL 提升企業報告品質的相關研究

由當前世界各經濟大國皆已陸續採用 XBRL 的實例而論，XBRL 能夠提升全球企業報告品質的事實不言而喻（各國採用 IFRS 與 XBRL 的現況調查請參考鄭丁旺、周濟群、周伯彥與廖育輝，2010）。而在學術證據方面，Hirst et al. (2002) 指出投資者認為企業報告中將資訊表達於附註內或是財務報表內，會造成投資者認知差異的風險。Maines and McDaniel (2000) 則發現非專業投資者（nonprofessional users）對於不同的資訊呈現處所給予的權重有別，例如將同一衡量指標置於損益表等績效報告、附註或是非財務資訊報表，非專業投資者較會關心損益表等績效報告中的指標。Hodge et al. (2004) 利用實驗室研究法，試圖探討 XBRL 對非專業投資者於企業報告資訊取得（information acquisition）與資訊整合（information integration）上之助益，認為藉由剖析品質較佳的 XBRL 資訊，使用者將可改善其投資決策品質。

具體而言，Hodge et al. (2004) 提出：「廣泛採用 XBRL 可以加強企業管理當局選擇財務報導方式（或會計方法）的透明度（the transparency of managers' choices of

[4] 下載連結：http://www.twse.com.tw/ch/listed/XBRL/standard.php。

reporting），亦即將會降低管理當局玩弄財務報表數字的可能性。」換言之，這項可強化企業報告透明度的技術，將使得管理當局於會計估計或假設的選擇，更容易地為一般使用者所發現，因而會促使管理當局以更為中立的方式來選擇會計方法與揭露資訊。Hodge et al. (2004) 的研究提出了一項重要的資訊意涵，由使用者之資訊需求觀點，財報附註或年報電子書等公開的非財務資訊，對於投資決策確實十分攸關，但卻可能因為仍使用 pdf、html 等非結構化（unstructured）的格式資訊，而導致資訊取得與資訊整合之不效率。

國內研究則提出所謂「開放式企業報告」資訊架構，利用 XBRL 的「可尋獲的分類標準集合」技術將「開放式企業報告」中所包含的重要資料項目與文字加以標記化、模組化，並建置成分類標準，同時依據企業報告資訊品質之相關研究與理論，提出四個研究假說，並利用一個單因子受試者間的實驗設計，藉由研究發展出的「開放式企業報告」分類標準與案例文件等，於實驗室進行各項假說之測試；實驗結果顯示，對於一般非專業使用者而言，無論是資料的可靠性、可比較性、可取得性、攸關性與即時性等，可剖析、標準化與可延伸定義的 XBRL「開放式企業報告」資訊，均較傳統非格式化、無系統性之企業報告具有較高的品質（周濟群，2009）。

在非財務資訊使用 XBRL 格式來呈現的研究方面，Debreceny, Chandra, Cheh, Guithues-Amrhein, Hannon, Hutchison, Janvrin, Jones, Lamberton, Lymer, Mascha, Nehmer, Roohani, Srivastava, Trabelsi, Tribunella, Trites, and Vasarhelyi (2005) 整理美國證券管理委員會 SEC 在推動美國上市公司的 XBRL 自願性揭露計畫（voluntary filing program, VFP）時，曾提出 SEC 準備利用 XBRL 技術來建構所謂「開放式的資訊揭露」（open-ended disclosure）方式，認為此一新方向將成為投資人未來評量企業營運概況、公司治理等問題的重要依據。然而，雖然強調非財務資訊對於投資決策過程中的重要性，該研究卻建議 SEC 於制定分類標準（taxonomy）時，勿將非財務資訊的標籤（tag）訂得太細，只需定義最簡單、最上層的資訊標籤，唯恐此類的資訊揭露若規範過細，反而無法反應企業真實經濟實況，甚至降低企業申報之意願。

(二) XBRL 技術架構

目前最新 XBRL 第 2.1 版規格標準，與一般 XML 綱要（XML schema）文件不同，並非將報告元素關係直接定義於分類標準文件中，而是利用 XLink 技術的「連結庫」（linkbase），將各報告元素間的連結關係，包括元素呈現關係、定義關係、計算關係、參考來源與標籤定義，使用許多 XLink 延伸連結，並獨立定義在 XBRL 連結庫中。因此，分類標準並非單一的 XML 綱要，而是併入獨立出來的連結庫後，成為一個分類標準套件（taxonomy package），套件包含所有相關綱要、各種連結庫等規格文件（specification document），統稱為「可尋獲的分類標準集合」DTS（discoverable taxonomy set）。包含在 DTS 內的文件必須以資料字典為核心，向外建立各種關係連結庫，以便產生與驗證一份符合 XBRL 規範的案例文件（instance document）。

XBRL 設計 DTS 機制的主要目標乃是藉由具備良好延伸性與標準化的格式，解決過去封閉式企業資訊擷取與交換的障礙，透過 XML Schema 和 XLink 等產業公認的網路技術，強化企業報告的延伸性、連結性、流動性和正確性等，尤其是強調「模組化」（modulize）、「匯入」（import）、「連結」（link）與「參照」（reference）等機制，使得 XBRL 企業報告資訊不侷限於傳統的財報，亦可納入各種非財務資訊，例如：企業重大訊息中所蘊含的數字、企業定期揭示的公司治理資訊等等。這些非財務資訊透過 XBRL 技術的應用，將能更即時地產生更多有用的分析價值，提供投資人在進行投資決策時的參考。

現行 XBRL 連結庫中雖然應用 XLink 技術來連結 XBRL 元素間的各種基本關係，但卻無法擷取具有互補性、附加性或解釋性資訊間的關聯，導致這些資訊的整合仍必須依賴使用者自行以人工方式或開發額外的應用程式來處理。

二、文字探勘

傳統的資料探勘（data mining）主要針對資料庫中結構化的數值型資料進行分析，試圖從中找出隱藏在資料中的趨勢、特徵及相關性的過程。而文字探勘則是針對文件型資料進行分析，自非結構化或半結構化的文字中，發掘出先前未知或者是隱含而有用的資訊。Sullivan (2001) 定義文字探勘為一種編輯、組織及分析大量非結構化文件的過程[5]，以符合使用者的特定資訊需求及發現某些特徵間的關聯，以挖掘出文件中的關鍵資訊，包括人、事、時、地、物、關鍵字等關鍵概念之間的階層關係，並加以分類、呈現。文字探勘相關之應用層面相當廣泛（Cimiano, 2006），包括資訊檢索（information retrieval）、資訊擷取（information extraction）、自然語言處理（natural language processing, NLP）、計算語言學（computational Linguistics）等，將大量文件，依據其特徵與屬性區分為許多分群，使性質相似的文件被分為同一群當中，讓使用者能快速區分文件類別，並迅速找到需要的文件。另一方面，對所有文件的分布提供一個綜覽，以提升文件的搜尋效益，並自動建立文件的分類架構，辨識文件中的字詞與關聯性，以減少文件檢索和查詢的誤判（謝家蘭，2006）。

在會計相關領域中，近來亦有若干學者運用文字探勘技術於財務會計研究中，例如 Engelberg (2008) 即曾定義所謂硬資訊（hard information）與軟資訊（soft information）之差異，利用 NLP 方法，針對企業法說會內容進行語意分析，將法說會報告內容區分為以文字表達的質性資訊（軟資訊，例如：管理當局討論與分析）與可數量化的資訊（硬資訊，例如：財務報表資訊），該研究發現軟資訊確實具有增額資訊內涵，但其反應時間較長，顯示資訊處理成本確實會影響資本市場效率。Antweiler and Frank (2004) 則使用簡單貝氏分類法（Naïve Bayes classifier），分類分析 Yahoo Finance 上討論區的文字性資料，發現這些討論區文章確實會影響股價

[5] 所謂非結構化文件，則是指以自由形態方式呈現的自然語言內容，像是新聞、會議紀錄、電子郵件、手冊、公司章程等，其內容並無一定格式，組成元件不易明確切割、命名（Zhou, 2007）。

與成交量，因此建議分析師應偵知這些網路上的情緒性資訊（sentiment detection）。Tetlock (2007) 則利用 General Inquirer 文字分析軟體，針對 S&P 500 的企業盈餘發布前的華爾街日報報導內容進行內容分析，以偵知投資人的情緒，並依哈佛心理字典（*Harvard IV-4 Psychological Dictionary*）進行內容的分類，結果發現負面報導確實具有異常報酬的解釋力。

在處理技術方面，英文文本不需要使用斷詞相關技術，而在中文文本中則必須加以處理。由於本研究之研究文本為中文年報資料，因此，以下將以中文文本為例，依次介紹中文文字探勘之主要工作：中文斷詞、N 元詞（N-gram）、特徵擷取及向量空間模型建構。

(一)中文斷詞

詞是自然語言處理上最基本的單位，所謂的詞是指語言學家所定義的「能夠獨立運用，具有完整語意的最小語言成分」。英文的每個單字都可以成為詞，具有自己的意義，且每個詞間都有明顯的空白作為分隔，因此沒有所謂斷詞的困擾。相反的，中文在書寫時，詞與詞之間無空白做為區分，也就是說，單就文字的表現形式來看，中文並沒有詞這個單位。

中文詞可由單詞之字元數來分類，因此中文自然語言處理領域以 N 元詞表示該詞內之字元數。例如二元詞（bi-gram）有「董事」、「金融」、「經營」等詞，三元詞則有「投資者」、「金融業」、「獨立性」等詞，依此類推。本研究在中文斷詞部分採用中研院中文斷詞系統，此系統由中央研究院詞庫小組（CKIP）開發，具有自動抽取新詞、建立領域用詞及線上即時分詞的功能，為一具新詞辨識能力且附加詞類標記的選擇性功能系統。

(二)特徵擷取及向量空間模型

向量空間模型（vector space model, VSM）的概念最早由 Salton and McGill (1983) 所提出，基本上，向量空間模型是一種由特徵詞與文件所組成的向量空間，具有擷取文件內資訊特徵以增強文件檢索效能的功能。為協助檢索進行，必須在檢索前對資訊本身進行分析。此分析過程被稱為建立索引，索引的主要目的在建立文件內容特徵，亦即透過賦予索引詞權重，以顯示詞彙在文件中之重要性。建立索引的方法為針對系統中整體文件集合 D，找出一組屬性為（$W_1,W_2,\ldots,W_k$），並在文件集合 D 中找出某一文件 D_i 能有一組屬性值為（$W_{i1},W_{i2},\ldots,W_{ik}$），使得文件 D_i 具有足夠的資訊以代表文件集合 D。該組屬性值稱為文件 D_i 的索引向量元素，即所謂的權重。以文件查詢為例，每份文件可以計算出代表該文件之特徵向量，該向量之每個維度都是文件中的一個詞。這些特徵向量可以表示成如圖 1 所示：

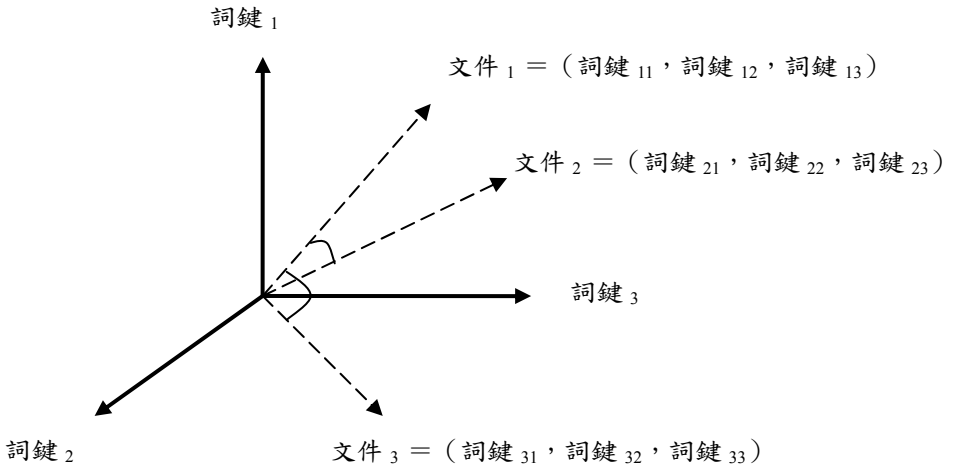


圖 1　向量空間模型

算出每個詞的向量之後，接著衡量詞向量與某文件或分類之向量距離比較近，距離越短者相似度越高，表 1 列出較常使用的相似度公式（Salton and Buckley, 1988）：

表 1　相似度計算公式表

衡量方式	兩文件間之向量計算式：以二元向量表示	兩文件間之向量計算式：以不同權重表示向量
向量內積（Inner Product）	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Dice 係數（Dice Coefficient）	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2}$
Cosine 係數（Cosine Coefficient）	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \cdot \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}}$
Jaccard 係數(Jaccard Coefficient）	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i \cdot y_i}$

利用向量代表各個文件，不但可以清楚呈現各個文件間的關係，且彼此間的相似度也較易計算，當文件意義相近時，可能會有很多相同的詞彙，若利用向量空間作表達時，這些向量會較接近。

在一份文件中，每個索引字詞都代表空間中的一個維度，維度上的值代表該文件在此維度上的重要程度，此值稱為該索引詞彙的權重值。而權重值之計算方式則有，IDF（inverse document frequency）加權模型、TF（term frequency）加權模型、

TFIDF（term frequency / inverse document frequency）加權模型及 TFITF（term frequency / inverse total term frequency）加權模型等。

其中，TFIDF 加權模型常用於資訊檢索與文字探勘，主要用來計算特徵詞於文件中之權重。本研究採用此模型計算詞項定義中特徵詞權重，以產生詞項特徵向量。過去 TFIDF 的研究主要針對文件分類方面，協助使用者有效的擷取與過濾網頁文件等資訊，而本研究則將分類單位縮小，嘗試將 TFIDF 應用於概念（concepts）上的詞鍵權重值計算，也就是年報非結構化文字內文分類階層中各分類項目。本研究修改原始 TFIDF 計算公式如下：

1. *TF* (term frequency)：計算字頻，某一詞鍵在概念中的出現頻率。

$$TF_{ij} = c_j / c_{all}$$

TF_{ij} ：詞鍵 j 在概念i中之出現頻率；
c_j ：詞鍵 j 在概念i中之出現次數；
c_{all} ：概念 i 中，所有具有意義的總詞頻。

2. *ICF* (inverse concept frequency)：計算反頻率，詞鍵 j 在所有概念裡出現頻率的倒數。

$$ICF_j = \log_2(N/cf_j)$$

ICF_j ：代表詞鍵 j 在所有概念裡出現頻率的倒數；
N ：代表所有概念的總數；
cf_j ：代表詞鍵 j 所出現的概念總數。

3. *Weight* (權重)：詞鍵在概念中的權重值。

$$Wi_j = Tfi_j \times ICF_j = \left(c_j / c_{all}\right) \times \log_2\left(N / cf_j\right)$$

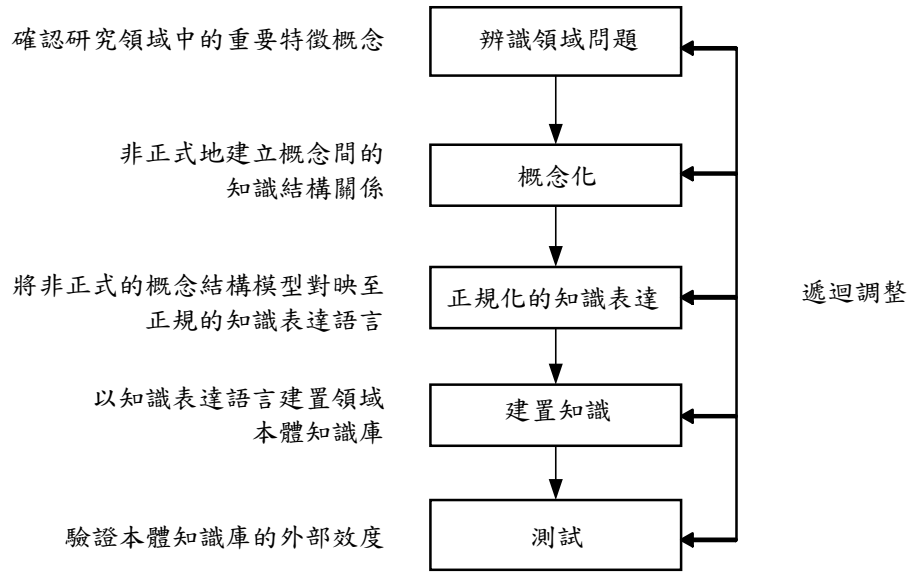
W_{ij} ：即為詞鍵 j 在概念 i 中的權重值；
TF_{ij} ：詞鍵 j 在概念 i 中之出現頻率；
ICF_j ：詞鍵 j 在所有概念裡出現頻率的倒數。

以當前的資訊科技而言，此類非結構化文件資料之處理，乃是以文字探勘（text mining）為主要方向，除文字探勘技術以外，雖然有許多研究嘗試將非結構化文件轉換為如 XML 或 XBRL 等結構化的格式，以利自動化分析，但由於 XML 或 XBRL 等結構化格式文件需要嚴謹的資料綱要（data schema）或元資料（metadata)）來定義與描述文件，以現有的文字探勘或資訊擷取（information retrieval）技術而言，尚無法完美地產生高品質的結構化文件。

三、應用本體的知識概念分類方法

本體技術發展的主要目的為使各領域知識與資訊能夠以相通之架構來呈現及描述，以便能配合智慧型或自動化系統支援知識分享及重複利用，目前早已廣泛地應用於知識工程（knowledge engineering）、人工智慧（artificial intelligence）、電腦科學（computer science）、商業及社會科學等領域（Gómez-Pérez, Fernández-López, and Corcho, 2004）；在建構某一特定領域的知識體系和基本內容時，採用本體技術可針對領域概念、專門術語及其間相互關係進行規範化的描述，呈現出提供一個讓人與人之間及不同的應用系統之間，可以彼此分享、溝通，進而達成共識的一個關於某個領域知識內容的媒介。由於本研究的主要研究目標之一，乃是建立企業年報資訊的概念分類系統，基於目前並無可明確參考的年報內容分類，因此本文乃嘗試引用 Fujihara, Simmons, Ellis, and Shannon (1997)「知識概念階層形成過程」的本體建置方法來進行。

Fujihara et al. (1997) 改良並簡化Uschold and King's Method (Uschold and Gruninger, 1996) 方法，認為高品質的本體知識擷取過程應為一個循環的生命週期模型（life cycle model），共包含辨識問題、概念化、正規化、建置知識、測試等五個步驟（如圖 2），由於本研究並非建置正式本體定義的大型研究，基於此法的簡便性與可行性，應較適用於本研究。



資料來源：整理自 Fujihara et al. (1997)

圖 2　知識概念階層之形成步驟

參、研究方法與系統設計

本研究採取設計科學研究方法，以建置企業資訊整合分析系統的概念（construct）、模型（model）、模型建構方法（method）、實體案例（instantiation）等四個產出為主要研究目標（March and Smith, 1995; 周濟群，2009）。系統設計則以林東清（2002）之決策支援系統架構理論為核心，將企業資訊整合分析系統架構將分為四部份：資料來源、資料儲存、資料分析與資訊顯現，如圖 3 所示，並分述如下：

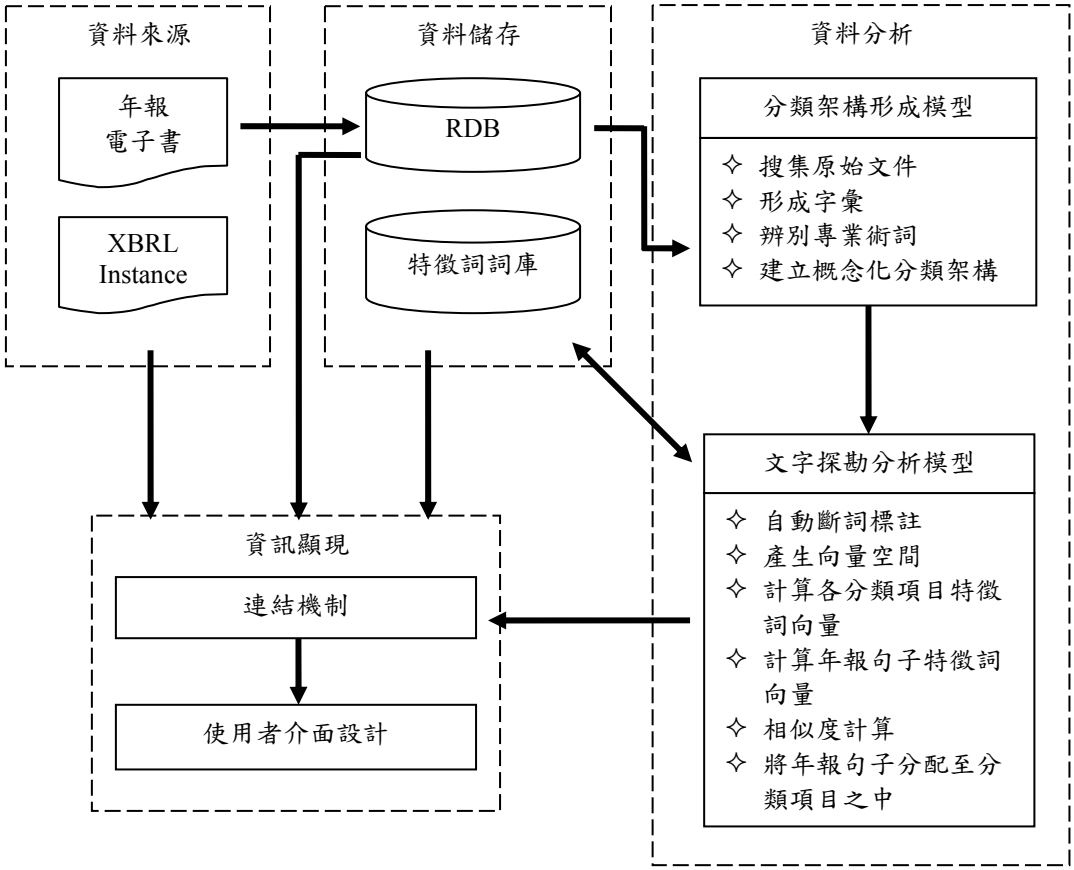


圖 3　系統架構圖

一、資料來源

本研究之主要資料來源有二，皆來自於臺灣證券交易所公開資訊觀測站：一是企業年報電子書，二則是將企業年報電子書中的財報資料依「一般行業XBRL財務報表分類標準」所轉換之XBRL案例文件[6]。

[6] 本研究採用人工方式建置 XBRL 案例文件，主要使用工具為 Fujitsu 的 XWand Instance Editor 軟體（周

表 2 年報內文自動分類訓練與測試資料表

手動對映訓練資料		半自動對映訓練資料(1)		半自動對映訓練資料(2)		測試資料	
公司	年度	公司	年度	公司	年度	公司	年度
友達	2003	華碩	2005	華亞科	2005	台積電	2003
台達電	2003	可成	2005	大立光	2005	亞泥	2004
友達	2004	華寶通訊	2005	聯發科	2005	中鋼	2005
宏碁	2004	仁寶	2005	南電	2005	台化	2005
中華電信	2004	長榮海	2005	聯詠科技	2005	台肥	2005
長榮海	2004	遠傳	2005	寶成	2005	瑞昱	2005
台積電	2004	台塑化	2005	廣達	2005	神達	2005
友達	2005	鴻準	2005	矽品	2005	研華	2005
宏碁	2005	宏達電	2005	台哥大	2005	正崴	2005
日月光	2005	鴻海	2005	台積電	2005	國巨	2005

本研究採抽樣方式選取樣本公司，以 2006 年 10 月 13 日臺灣證券交易所公佈之臺灣 50 指數成份股為母體，在剔除金融證券業後，計有 36 家上市公司，時間範圍將設定在 2003 年至 2005 年。此外，由於部分成份股之年報電子書存在著轉換編碼無法解開的問題，可能會造成年報自動分類訓練與測試樣本之不足，因此另自臺灣 100 中型指數取出部分成份股來補足。此部份並非主觀因素所造成，應不至於造成訓練或測試之誤差。且本研究所應用的中文斷詞技術，乃是由中央研究院詞庫小組（CKIP）開發，在處理企業年報中內含文字時，原則上不會因為時間別或產業別產生差異，因此本研究所建構之雛型系統，其可一般化程度應不會受到時間別或產業別之影響。

在「企業年報非結構化文字資訊分類系統」方面，需要進行兩階段的資料訓練：手動與半自動，其中半自動部份再分為兩個子階段，共抽取 30 份年報文件，而測試階段則抽取 10 份（見表 2），測試後的資料亦可再回饋至分類系統中，以增強分類效率。至於樣本量部份，所輸入之 40 本年報電子書共約超過 4,000 句，以此文字訓練總量應已足夠實驗之所需，不致影響訓練或測試之效度。

至於 XBRL 財報資料，則限於人工建置的成本過高，本研究僅選取母體中之 10 家公司，2003 年至 2005 年度，共計建置 30 份 XBRL 案例文件（見表 3）。

二、資料儲存

本研究以關聯式資料庫儲存企業年報資料，但由於年報電子書多為 pdf 格式，對程式而言並不容易處理，因此需先轉換成純文字格式，以方便資料儲存與分析。

濟群，2009），依據的分類標準版本為 2010 年度，因此若遇有不存在之科目，本研究已自行延伸，詳細的分類標準與案例文件建置方法不列為本研究之範圍，故將此部份視為資料來源。

表 3 XBRL 案例文件內容一覽表

股票代碼	公司名稱	股票代碼	公司名稱
2308	台達電	2353	宏碁
2311	日月光	2357	華碩
2317	鴻海	2490	友達
2324	仁寶	3008	大立光
2330	台積電	3045	台灣大哥大

此外，年報內文自動分類的訓練結果，以及連結機制的特徵詞詞庫也需要儲存於資料庫中。至於 XBRL 案例文件，則是以檔案型式另外儲存於作業系統中。

三、資料分析

模型導向之資料分析主要透過模型庫中各種不同模型擷取資料庫內之資料來加以分析，依本研究之需要，共需建立兩個資料分析模型：

(一)分類概念模型：

關於如何形成年報非結構化文字之分類階層，本研究利用企業策略分析的 SWOT 矩陣與平衡計分卡概念，並援引 Fujihara et al. (1997) 提出的知識概念擷取方法，將原本非結構化之純文字敘述，轉換成有系統的概念性階層。

(二)文字探勘分析模型：

完成概念性階層之後，再利用非結構化文字資訊之探勘方法，輸入已轉換成純文字格式之企業年報文字檔案，透過自動斷詞標註、產生向量空間、計算各分類項目之特徵詞向量（feature vector）、計算以句子為單位之年報內文特徵詞向量、計算分類項目特徵詞向量與年報句子特徵詞向量之相似度，最後將年報以句為單位分別對應至適當分類項目之中，完成整個年報內文自動分類之流程。

四、資訊顯現

經由使用者介面之設計，系統使用者除瀏覽分類後之企業年報內文外，也可透過連結機制，同時查詢 XBRL 案例文件中的財務資訊。

肆、研究成果

一、年報知識概念階層形成過程

目前企業的非財務資訊揭露，受限於法令之規範而須以制式化的方式呈現[7]，但這些制式的分類表達方式，對投資人來說可能難以快速、有效地應用。因此，若能適當運用資訊技術，將這些非結構化格式資訊加以處理、分析，以挖掘出文件中的關鍵資訊，並加以適當地分類，應可提升使用者的資訊解讀效能。

本研究所提出「企業年報知識概念階層」的主要分類內容範圍，乃是針對現行企業年報中，關於企業組織結構與營運相關事項的報導。本研究首先將 Fujihara et al. (1997)「知識概念階層形成過程」具體化為：解析年報文件、形成專業術詞、辨別概念化階層關係、建立分類概念、測試分類有用性，再依各步驟將企業年報中的重要資訊擷取並標記出來，將原本缺乏結構性的資訊，轉換成有系統的概念性階層。

(一)解析年報文件

取得企業年報的主要管道乃是透過證交所公開資訊觀測站的電子書下載，通常內容約有一百多頁以上，其中除有表格化或結構化的資訊之外（例如四大財務報表），尚有相當數量的非結構化文字敘述，記載著各種企業現況，以及未來策略走向、整體產經環境趨勢等。本研究選擇年報中與企業策略資訊相關之「致股東報告書」、「營運概況」、以及「財務狀況及經營結果之檢討分析與風險事項」內的風險事項等三大章節，去除表格資料後，作為本分類階層之原始資料來源。

(二)形成專業術詞

原始文件解析完畢後，將文句中之停字詞（stop words）或干擾詞（noise words），例如「的」、「是」等字詞加以過濾刪除後，再經過該領域專家之判讀，去蕪存菁，從中條列出有意義之專業術詞。以下為專業術詞的簡單釋例：

XX 光電、顯示器產業、創新、研發、研發資金、前瞻技術、專利、新製程、產品改良技術、創新產品的開發

(三)辨別概念化階層關係

辨別出該領域之專業術詞後，接著必須找出各詞彙間之關係，並考慮詞彙在該領域中之影響程度，以階層的方式呈現其分類架構。Gómez-Pérez et al. (2004) 認為，將無階層結構的術詞集合加以概念化（或一般化）的過程，需要更深入的領域知識，因而建議若領域中已存在較成熟的知識架構，將可提升建立概念化階層的效度。據

[7] 如上市櫃公司之股東會年報需依《公開發行公司年報應行記載事項準則》進行編製。

此，本研究乃參考領域專家對於企業策略管理與價值之相關理論，如Pricewaterhouse Coopers於 1999 年提出之企業價值報告（business value reporting, Eccles, Herz, Keegan and Phillips, 2001）、管理領域中著名的SWOT分析矩陣，以及平衡計分卡（balanced scorecard, BSC）概念等，分析、整合尚未階層化的專業術詞，建構出完整之年報分類階層。透過重新分類後的年報資訊階層，使用者得以於短時間內瞭解企業於財務、顧客、內部程序及學習與成長四大構面中，並比較可能存在的內部優勢與劣勢，以及目前或未來將面臨之外在機會與威脅。依據上述三種領域分類年報中的知識，共完成五個層級，共計 92 個項目之企業年報概念階層，簡述如下（部份內容如表 4）[8]：

1. 優勢與劣勢、機會與威脅（第一層級：2 項）

PWC 的企業價值報告主要應用於企業外部因素之分析，並探討企業內部如何依據外在環境擬定目標價值（價值策略）、如何經由日常管理過程來執行策略步驟（價值管理）以及管理階層應採取哪些有助於長期企業價值成長之活動（價值平台）。另一方面，SWOT 分析則考量企業內部條件之優勢與劣勢，是否有利於在產業內競爭；機會和威脅是針對企業外部環境進行探索，探討產業未來情勢之演變。綜上所述，乃將分類階層第一層分為優勢和劣勢（內部因素）與機會和威脅（外部因素）。

2. 財務、顧客/產業、內部程序、學習與成長（第二層級：8 項）

平衡計分卡要求經理人由四個向度評估組織的表現，即「顧客」、「內部程序」、「學習與成長」及「財務績效」，並在願景和策略的引導和整合下達成績效目標。據此，乃將第一層之優勢與劣勢(S/W)、機會與威脅(O/T)加以進一步細分：S/W 細分為「財務」、「顧客」、「內部程序」、「學習與成長」；O/T 則細分為「財務」、「產業」、「內部程序」、「學習與成長」。

3. 財務表現、顧客/市場區格、品質控管、創新流程等（第三層級：26 項）

根據第二層四個構面之分類，並實際分析選定之年報非結構化文字資訊內容，再予以細分，例如 S/W 之「內部程序」再細分為「創新流程」、「營業流程」與「售後服務流程」；O/T 之「產業」則細分為「產業景氣與供需狀況」、「產業未來發展趨勢」、「企業因應未來發展之對策」、以及「產業競爭情勢」等。

[8] 本研究之主要目的乃是提供一種以文字探勘技術來輔助文件內容分類的方法，並不能宣稱本研究所訂定的分類方式（如表 4）為一般公認的分類。由資訊技術的角度而言，若能依使用者不同的分類需求，彈性地調整文件內容分類方法，以客製化的方式來滿足不同人、時、地下的不同使用者需求，以求符合實際的決策情境，即已達到資訊技術發展之目的，因此本研究所提出的分類方式應可視為諸多分類中的一個案例。

表 4 年報章節內容與分類階層對映表（部份）

年報章節	年報章節內容	分類階層
致股東報告書	營運績效	S/W：財務
	未來公司發展策略	S/W：內部程序
	外部競爭環境	O/T：產業
	總體經營環境	O/T：財務
營運概況	業務內容	S/W：顧客
	市場及產銷概況	O/T：產業
	研究發展	S/W：內部程序
	從業員工	S/W：學習與成長
	環保支出	O/T：內部程序
風險事項	總體經營環境	O/T：財務
	科技改變及產業變化	O/T：產業
	從事高風險之投資事項	O/T：內部程序

4. 營收表現、主要業務內容、研發能力等（第四層級：33 項）

根據第三層的分類，並斟酌選定之年報非結構化文字資訊內容，繼續往下細分，如「創新流程」細分為「研發能力」、「製程技術」與「未來研發展望」；「經濟因素」則細分為「總體經濟表現」、「物價水準」、「利率波動」、「匯率波動」與「貨幣供給」五項。

5. 市場占有率、研究發展投入、研發成果等（第五層級：23 項）

根據第四層的分類，並參考選定之年報非結構化文字資訊內容，繼續往下細分，如「主要業務內容」細分為「營收比重」與「市場占有率」；「研發能力」則細分為「研究目標與願景」、「研究發展投入」與「研發成果」。

(四)建立分類概念

建立上述各階層的分類概念架構主要是為用於後續的資料訓練，故不需使用本體正規語言來建立分類知識，僅需將其分類結果依階層建置於資料庫中即可。

(五)測試分類有用性

本研究後續將以建立雛型系統的方式來驗證上述各階層的分類概念。

二、年報內文與分類概念階層相似度之運算方法

圖 4 為年報內文與分類概念階層相似度之運算流程。以下針對特徵詞向量空間的建立、特徵詞向量值之計算以及相似度計算三項關鍵方法加以說明：

(一)建立特徵詞向量空間

利用 CKIP 斷詞系統，將訓練樣本中所有詞彙建立成一個文字表（word table），以此作為特徵詞向量（feature vector）空間。

(二)特徵詞向量之計算

根據文獻探討所提到之修改後 TFIDF 計算公式，透過將年報內文逐句對映至各分類概念，即可建立各每個概念（分類項目）的詞鍵權重列表，亦即各個概念之向

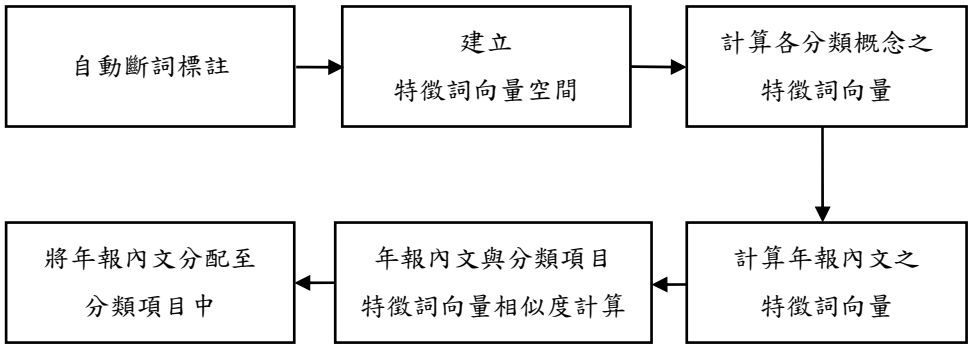


圖 4　年報內文與分類概念階層相似度之運算流程

量值。另一方面，年報內文也可以利用此方式計算出每一句之特徵詞向量，以進行最後之相似度之比對。

(三)相似度計算

由於本研究的特徵向量空間較大，且 cosine 係數較常用於計算兩文件間的向量夾角，因此採用 cosine 係數來計算兩個特徵向量間之相似度。cosine 係數值介於 0 至 1 之間，當 cosine 係數值越接近 1，代表兩向量夾角越小（如圖 5），即代表此二特徵向量相似度越高，否則相似度越低。

三、年報內文自動分類

本階段主要目的為利用文字探勘作為年報內文分類之技術基礎，並根據案例式學習（case-based learning）的概念，設計四階段的訓練及驗證過程：準備工作、手動對映、半自動對映與全自動對映：

(一)準備工作

經由年報蒐集器自公開資訊觀測站下載年報電子書，並將選定之非結構化文字資訊內容由 PDF 格式轉換成純文字 txt 格式，並剔除因特殊編碼而造成無法正確進

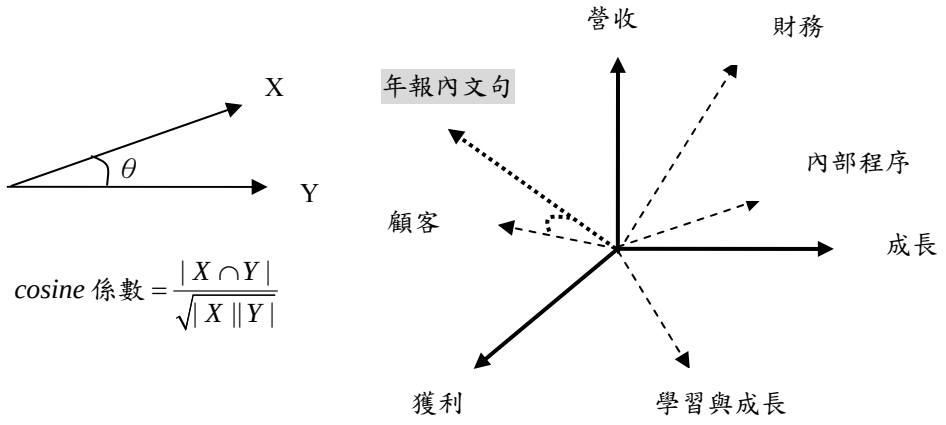


圖 5 相似度計算

行格式轉換之公司與年度後，共完成 40 份年報資料，並以一個句子為一筆記錄列的方式，存於資料庫中。

(二)手動對映－訓練階段

完成第一階段之資料蒐集與轉換後，自資料庫中讀取 txt 檔，解析該檔之年度、公司與報告名稱，並運用領域專家之專業知識，將年報內文以句為單位對映至適當之分類階層項目（節點），存入資料庫。手動對映樣本累積至相當數量後，即停止人工對映之方式，並將手動對映之全部案例結果進行斷詞、關鍵字與統計值之計算。本階段手動對映之驗證方式，係透過隨機抽取分類樣本，由領域專家進行年報內文分類正確性評估，並依據評估結果進行分類之修正，存入關鍵詞詞庫。

(三)半自動對映－訓練階段

所謂之「半自動對映」，顧名思義即以上一階段人工對映為基礎，自案例庫中取出經過處理的關鍵詞與統計值，再由程式將新輸入之年報內文進行斷詞、關鍵字與統計值之計算，交由領域專家驗證調整，並重覆上述之對映動作。藉由機器對映與人工驗證之比較，計算目前程式分類之正確率[9]。由於人工與半自動化對映之後續作業皆包含領域專家驗證調整過程，因此可確保存入關鍵詞詞庫中的結果是完全正確的。

(四)全自動對映－測試階段

當年報內文分類正確率達到滿意水準之後（本研究設定滿意水準為：任一階層>80%），便完全交由機器進行自動對映，並記錄未達滿意水準之異常對映。由於自

[9] 正確率=正確分類之句數/所有句數，正確率結果可參考表 5。

表 5 半自動對映與全自動對映正確率一覽表

		分類階層				
		1	2	3	4	5
分類項目數（累計）		2	10	36	69	92
正確率	半自動對映訓練資料 (1)	77.83%	58.36%	46.77%	39.94%	37.42%
	半自動對映訓練資料 (2)	87.12%	69.94%	60.11%	52.66%	51.30%
	全自動對映測試資料	87.45%	74.71%	65.21%	58.37%	56.46%

動對映完全交由程式分類，領域專家僅進行分類正確性評估，因此本質上已為測試作業。表 5 彙整半自動對映與全自動對映階段之程式分類正確率。正確率的高低問題，較難客觀評斷，過去較多文獻是以整份文件為分類單位（De Bruijn and Martin 2002; Hui and Yu 2005; Chi 2007），且其第一階層正確率亦均介於 40% 至 80% 之間，而本研究採用字詞量相對較少的「句」為單位，當單一句子包含字詞量較少時，特徵值會較模糊，必將更難以準確分類。

至於分類愈細時，可能有些分類彼此之間有些模糊空間存在，所以即使是人類專家，都不容易區分，因此正確率必然會隨著分類愈細而降低。此外，因類別越多，訓練資料會被切割（字詞量越少），所以也會導致訓練資料不足，進而造成正確率下降。

四、連結機制

(一)特徵詞詞庫

連結機制係將年報非結構化文字內文分類階層中各分類項目之特徵詞向量排序取出前 20 名，剔除無意義或不合理之單詞後[10]，從中篩選出與財務報表科目有關詞彙，建構分類標準各項目元素之特徵詞詞庫，並以新增延伸性標籤的方式儲存於 XBRL標籤連結庫之中。系統使用者於閱讀已分類之年報內文同時，可透過句子中某些特徵詞連結至XBRL案例文件，此即本特徵詞詞庫之主要功能。另一方面，除本特徵詞詞庫內建置完成之 14 組項目元素外，使用者亦可自行新增其他特徵詞，以擴充特徵詞詞庫之內容[11]。

(二)連結內容

完整的投資決策流程除企業策略分析之外，尚需探討盈餘品質之會計分析、以及比率分析為核心之財務分析，因此連結機制除原來對映之主要財務數字資訊外，

[10] 例如：原本「資本」一詞亦為「應付公司債」的特徵詞，究其因，乃是由於這兩個詞項同時出現於年報中的頻率甚高所致，但因其不合常理，故業已將「資本」刪除。

[11] 特徵詞詞庫部份內容如附錄 1。

並同時提供使用者年報分類資訊相關的其他財務性輔助資訊，如相關會計科目、財務比率以及重要會計政策之說明[12]，並自XBRL案例文件中抓取上述之實際值元素與實際值。

以「營業收入」為例，其相關會計科目為營業成本、營業毛利，以及扣除營業費用後的繼續營業單位稅前淨利、應收帳款等，透過數字與趨勢分析圖之呈現，協助系統使用者評估企業其收入品質，以及針對企業銷售策略進行收入相關資訊之驗證與整合。

五、企業年報非結構化文字資訊分類雛型系統

本節將以台積電 2003 年度年報為模擬範例，展示本研究部份實作成果。使用者可在雛型系統中進行年報非結構化文字內文分類閱讀，瞭解企業過去與未來之競爭策略走向及產業發展情況，同時可經由連結機制查詢財務性資訊，以印證該企業策略之執行成果。

(一)年報非結構化文字資訊分類閱讀

透過本研究分類閱讀的設計，將年報中非結構化文字資訊的部分（圖 6 左方），以企業策略分析之 SWOT 與 BSC 等理論加以分類（圖 6 右方），以進行年報非結構化文字資訊分類閱讀，例如在「營收表現」分類項目下，使用者可以閱讀到原本位於致股東報告書的內文：「民國九十二年，本公司財務表現優異，全年營收為新台幣 2,019 億 4 百萬元，較前一年成長 25%；稅後淨利達新台幣 472 億 5 千百萬元，較前一年成長 119%；每股盈餘則為新台幣 2.33 元，較前一年成長 122%。」，透過此種分類閱讀方式，可提升年報之可閱讀性。

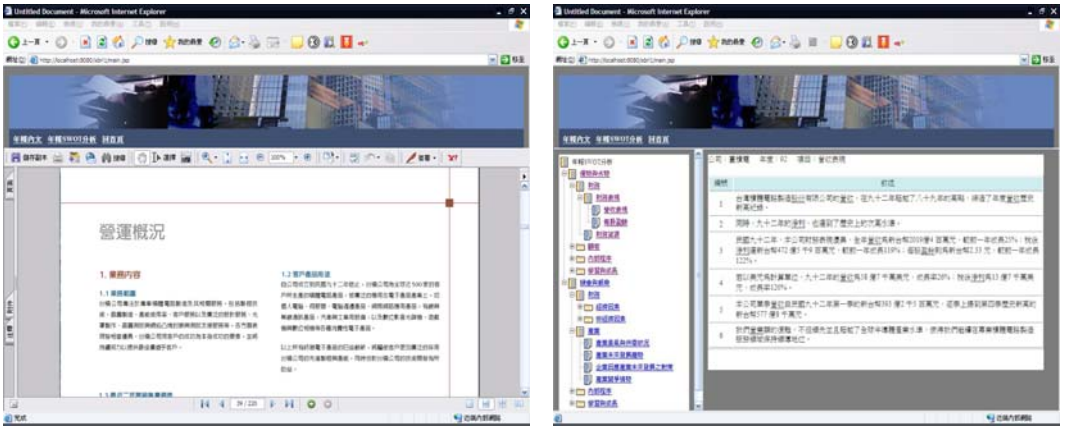


圖 6　年報內文分類閱讀

[12] 請參閱附錄 2「財務性輔助資訊內容表－相關會計科目」、附錄 3「財務性輔助資訊內容表－財務比率」、以及附錄 4「財務性輔助資訊內容表－重要會計政策」。

(二)非結構化文字資訊與財務性資訊之連結

在閱讀企業策略資訊的過程中，投資人可能需要財務資訊的輔佐以印證標的企業其策略方針之落實情況，每一句被分類的年報內文中，可能或多或少隱含著這些訊息。在過去，投資人只能自行翻閱企業的財務報表進行對照，而且需要額外成本才能解決檔案格式上的不一致。然而，透過本研究的連結機制（如圖 7），將年報內文中與財務資訊有關詞彙，將自動以加底線方式註明，並超連結至 XBRL 案例文件中抓取對應之實際值，讓使用者可以在最短的時間內，取得與企業策略相關之財務資料。以前述之年報內文為例，使用者於點閱「淨利」與「盈餘」這兩個詞彙後，將會出現台積電 2003 至 2005 年本期淨利之金額，並呈現與本期淨利有關之科目金額，例如投資人最關心的每股盈餘數字，以及財務比率分析結果，如衡量企業經營績效之資產報酬率（ROA）、股東權益報酬率（ROE）與毛利率等。

圖 7　年報資訊與財務資訊之連結

伍、結論與建議

一、研究結果彙總

本研究利用概念階層形成之理論，建置年報內容分類階層，並以文字探勘技術為基礎，透過案例式學習方式訓練年報內文之分類能力，提供系統使用者進行分類閱讀。另一方面，根據目前主管機關對於上市櫃公司財務報表之相關規範，延伸建置 XBRL 財務報表分類標準，並編製 XBRL 案例文件，以標準化財報資料格式。最

後，透過連結機制之設計，建置分類標準項目元素之特徵詞詞庫，從已分類之年報句子中與財務相關特定詞彙，連結至 XBRL 案例文件內之財務數字、會計政策以及財務比率之計算，以提供系統使用者輔助查詢閱讀之用。經由研究結果的呈現，證實建立年報中非結構化資訊的分類系統，以及具決策關聯資訊間之連結，確實可以達到機器協助人類使用者（包含財報分析專家/非專家）進行自動化資訊擷取與整合的目的。

二、本研究之會計意涵

相較於過去大海撈針無效率之人工閱讀方式，本研究善用文字探勘技術與 XBRL 之優勢結合，顯著地提升決策制定者搜尋與分析之效率，系統使用者可能從現有資料中挖掘出新的事實及可能發現專家尚且不知的新關係。採用本研究開發的可連結財務/非財務資訊整合平台，對於會計專業來說，具有以下意涵：(1) 資訊擷取的效益：經由文字探勘技術的協助，可以將非結構化的年報資訊以具有知識結構的方式呈現，不需再以人工方式進行全文瀏覽、搜尋，可降低自資訊取得至其達到可用狀態之準備時間，提升資訊擷取的準確率與速度；(2) 資訊整合的效益：使用者能同時瀏覽到財報、財報附註及企業年報中原本並無直接連結之資訊，增加了單一資訊無法提供的整合價值，可以彈性整合不同類別的企業報告，協助使用者取得最具攸關性的資訊來進行研判或預測；(3) 非財務資訊或軟資訊處理技術的重要性：以往會計研究較為忽視的非財務資訊，在近來諸多財務與會計研究中，逐漸受到重視，主要原因即為過去此類資訊的處理成本過高所致，但透過如文字探勘、XBRL 等資訊技術，可將解析非財務資訊的成本降低，增加此類資訊的效益；(4) 新的企業報告技術架構：在資訊技術進步的同時，未來企業報告的發布與流通形式應重新考量，應思考如何善用技術來提升市場效率，讓軟、硬資訊皆能充份得到使用，例如：目前年報這類具有資訊內涵的非財務資訊，仍以 pdf 等電子書格式來呈現，雖然較為美觀，但對於自動化處理則成為成本負擔，即使擁有本研究所提出之文字探勘方法，仍必須先將 pdf 轉換為純文字格式方能繼續處理，極為不便，未來若能透過 XBRL 文字標籤來加以呈現，除可解決資料處理問題以外，更能與財務資訊格式統一，增加企業資訊的可用度。

三、未來研究建議

在未來研究建議方面，以下將就 XBRL 於非財務性資訊之應用與平台之延伸應用兩方面進行說明：

(一) XBRL 於非財務性資訊之應用

以本研究之現階段成果而言，是以文字探勘技術暫時代替非財務性資訊 XBRL 標籤的制定，而且分類標準範圍尚未涵蓋所有的非財務性資訊，未來若能在資訊顆

粒切分與資訊揭露自由兩者間取得平衡點的話，訂定完整的企業非財務性資訊分類標準，相信對於企業資訊透明度的助益會相當顯著。

(二)平台之延伸應用

本研究所發展之企業年報非結構化文字資訊分類雛型系統為一可延伸性平台，目前應用層面僅限於企業策略分析與財務資訊之輔助閱讀，未來可延伸運用於各種分析預測模型，如會計盈餘品質分析、股價預測模型等，提供決策制定者詳盡之參考依據。此外，亦可依決策者之需求建構其他分類階層，分析不同領域之非結構化文字資訊，例如分析師報告、產業報告或法說會報告等，幫助決策者迅速、正確地擷取複雜報告中的重要資訊，對於促進資本市場效率預期可產生一定的幫助。

四、研究限制

本研究之企業年報非結構化文字分類階層，由於缺乏所謂「黃金標準」（golden standard）等級之公認本體（Brank, Grobelnik and Mladenić, 2005），因此僅能以研究者本身（皆為會計領域專家）先驗知識（prior knowledge）為基礎。所謂的先驗知識，泛指個人接受學習、受雇工作及自行創業等經驗中得到的知識與資訊，有助於對於某些事物之瞭解、推論與解釋，這些知識是無法複製於其他人身上的（Roberts, 1991）。因此，其他研究者亦可按照本研究之分類概念形成步驟或其他本體論建構過程，完成專屬於自己之年報分類階層。由研究信度而言，此一基於研究者個別認知的結果，雖然代表信度不足的風險，但由另一角度而言，本研究所發展的整體知識分類方法和連結機制，實可容納不同專家的先驗知識，依其個別需求來開發出個人化或客製化的決策資訊系統。

在分類正確率方面，由於本研究主要目標為驗證年報內文自動分類方法之可行性（proof of concept），並非開發出以正確度為依歸之產品，因此，受限於研究資源之不足，使得分類正確率於第一、二層分類項目雖達七至八成之正確率，但第三層之後卻降至五成左右。影響分類正確率的因素很多，可能測試文件中還有其他類別的關鍵詞，或者原始訓練資料錯誤分類等等，未來若能輸入更多訓練資料的話，相信分類正確率仍有相當之進步空間。

參考文獻

林東清，2002，資訊管理－E 化企業的核心競爭能力，台北：智勝文化事業有限公司。

周濟群，2009，利用XBRL技術設計可剖析的開放式企業報告架構，東吳會計學報，第 1 卷第 2 期：1-35。

鄭丁旺、周濟群、周伯彥與廖育輝，2010，全球化財報互通的兩項利器－看 IFRS 與 XBRL 如何整合，會計研究月刊 291 期：71-88。

謎家蘭，2006，淺介資料探勘與 XBRL，會計研究月刊，第 245 期：56-63。

Antweiler, W., and M. Z. Frank. 2004. Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance* 59 (3): 1259-1294.

Bonsón, E., V. Cortijo, and T. Escobar. 2008. The role of XBRL in enhanced business reporting (EBR). *Journal of Emerging Technologies in Accounting* 5: 161-173.

Brank, J., M. Grobelnik, and D. Mladenić. 2005. A Survey of ontology evaluation techniques. In SIKDD 2005 at Multiconference, Ljubljana, Slovenia.

Chi, Y. L. 2007. Elicitation synergy of extracting conceptual tags and hierarchies in textual document. *Expert Systems with Applications* 32 (2) 349-357.

Cimiano, P. 2006. Ontology *Learning and Population from Text: Algorithms, Evaluation and Applications* (1st ed.). Berlin: Springer-Verlag.

Cunningham, C. S. 2005. XBRL: A multitalented tool XBRL can save time and money and facilitate information analysis. *Journal of Accountancy* 199 (4): 70-71.

De Bruijn, B., and J. Martin. 2002. Getting to the (c)ore of knowledge: mining biomedical literature. *International Journal of Medical Informatics* 67 (1-3): 7-18.

Debreceny, R. S., A. Chandra, J. J. Cheh, D. Guithues-Amrhein, N. J. Hannon, P. D. Hutchison, D. Janvrin, R. A. Jones, B. Lamberton, A. Lymer, M. Mascha, R. Nehmer, S. Roohani, R. P. Srivastava, S. Trabelsi, T. Tribunella, G. Trites, and M. A. Vasarhelyi. 2005. Financial reporting in XBRL on the SEC's EDGAR system: a critique and evaluation. *Journal of Information Systems* 19 (2): 191-210.

Eccles, R. G., and S. C. Mavrinac. 1995. Improving the corporate disclosure process. *Sloan Management Review* 36 (4): 11-25.

Eccles, R. G., R. H. Herz, E. M. Keegan, and D. M. H. Phillips. 2001. *The Value Reporting Revolution: Moving beyond the Earnings Game* (1st ed.). New York: PricewaterhouseCoopers.

Engelberg, J. 2008. Costly information processing: Evidence from earnings announcements. 2009 American Finance Association (AFA) Annual Meeting, San Francisco. Working Paper, University of North Carolina.

Fujihara, H., D. B. Simmons, N. C. Ellis, and R. E. Shannon. 1997. Knowledge

conceptualization tool. *IEEE Transactions on Knowledge and Data Engineering* 9 (2): 209-220.

Gómez-Pérez, A., M. Fernández-López, and O. Corcho. 2004. *Ontological Engineering: with Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web*. London: Springer-Verlag.

Hirst, D. E., P. Hopkins, and J. Wahlen. 2002. Fair values, performance reporting, and bank analysts' risk and valuation judgments. Working Paper, The University of Texas at Austin and Indiana University.

Hodge, F. D., J. J. Kennedy, and L. A. Maines. 2004. Does search-facilitating technology improve the transparency of financial reporting? *The Accounting Review* 79 (3): 687-703.

Hui, B., and Yu, E. 2005. Extracting conceptual relationships from specialized documents. *Data and Knowledge Engineering*, 54 (1): 29-55.

Lara R., I. Cantador, and P. Castells. 2006. XBRL Taxonomies and OWL ontologies for investment funds. *Lecture Notes in Computer Science* 4231: 271-280.

Logan, D., and R. Mogull. 2003. Sarbanes-oxley: The role of technology. Retrieved from Gartner Research, http://www.gartner.com/DisplayDocument?id=412278.

Maines, L. A., and L. S. McDaniel. 2000. Effects of comprehensive-income characteristics on nonprofessional investors' judgments: The role of financial-statement presentation format. *The Accounting Review* 75 (2): 179-207.

Maines, L. A., E. Bartov, P. M. Fairfield, D. E. Hirst, T. E. Iannaconi, R. Mallett, C. M. Schrand, D. J. Skinner, and L. Vincent. 2002. Recommendations on disclosure of nonfinancial performance measures. *Accounting Horizons* 16 (4): 353-362

March, S. T., and G. F. Smith. 1995. Design and natural science research on information technology. *Decision Support Systems* 15 (4): 251-266

Matherne, L., and Z. Coffin. 2001. XBRL: A technology standard to reduce time, cut costs, and enable better analysis for tax preparers, *Tax Executive* 53 (Jan/Feb):68-70.

Pincus, T. H. 1989. Emerging from the dark ages: an overview of the investor relations art today. In Donald R. Nicholas ed., *The Handbook of Investor Relations* (1st ed.), pp.1-18. Illinois: Dow Jones and Company, Inc.

Plumlee, M. A., 2003, The effect of information complexity on analysts' use of that information. *The Accounting Review*, 78 (1): 275-296.

Roberts, E. B. 1991. *Entrepreneurs in High Technology: Lesson Form MIT and Beyond* (1st ed.). New York: Oxford University Press.

Salton G., and M. J. McGill. 1983. *Introduction to Modern Information Retrieval* (1st ed.). New York: McGraw-Hill.

Salton G., and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24 (5): 513-523.

Sullivan, D. 2001. *Document Warehousing and Text Mining: Techniques for Improving Business Operations, Marketing and Sales* (1st ed.). New York: John Wiley & Son, Inc.

Tetlock, P. C. 2007, Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance* 62 (3): 1139-1168.

Uschold, M., and M. Gruninger. 1996. Ontologies: principles, methods and applications. *Knowledge Engineering Review* 11 (2): 93-155.

Zhou, L. 2007. Ontology learning: State of the art and open issues. *Information Technology and Management* 8 (3): 241-252.

附錄 1　特徵詞詞庫部份內容

編號	項目元素名稱	項目元素標籤	特徵詞
1	NetIncomeLoss_9600	本期淨利	淨利、盈餘、純益、獲利
2	OperatingIncome_4000	營業收入	營業、營收、收入、銷售、營業額、總營收
3	OperatingCosts_5000	營業成本	成本、進貨
4	GrossProfitLossOperations_5910	營業毛利	毛利
5	BondsPayable_2410	應付公司債	公司債
6	CapitalStock_31xx	股本	資本額
7	CommonStock_3110	普通股股本	普通股、股份
8	TreasuryStock_3510	庫藏股票	庫藏股
9	ResearchDevelopmentExpenses_6300	研究發展費用	研發、開發
10	OperatingExpenses_6000	營業費用	費用
11	Patents_1720	專利權	專利
12	Inventories_120x	存貨	存貨、庫存
13	PensionReserve_2810	退休金準備/應計退休金負債	退休、退休金
14	LongTermInvestments_1420	長期投資	投資

附錄2　財務性輔助資訊內容表－相關會計科目

編號	主要財務數字資訊	財務性輔助資訊－相關會計科目
1	本期淨利	繼續營業單位稅前淨利、營業毛利、基本每股盈餘
2	營業收入	營業成本、營業毛利、繼續營業單位稅前淨利、應收帳款淨額、應收帳款淨額－關係人
3	營業成本	營業收入、營業毛利、應付帳款、應付帳款－關係人
4	營業毛利	營業收入、營業成本
5	應付公司債	一年或一營業週期內到期長期負債
6	股本	股東權益
7	普通股股本	股本、股東權益
9	研究發展費用	營業費用
10	營業費用	推銷費用、管理及總務費用、研究發展費用
11	專利權	無形資產
12	存貨	流動資產、存貨盤損、存貨跌價及呆滯損失
14	長期投資	累計減損－長期投資、基金及投資

附錄3　財務性輔助資訊內容表－財務比率

編號	主要財務數字資訊	財務比率名稱	公式（以元素名稱列示）
1	本期淨利	純益率	本期淨利/銷貨收入
		資產報酬率	[本期淨利+利息費用×(1-稅率)]/平均資產[13]
		股東權益報酬率	本期淨利/平均股東權益
2	營業收入	應收帳款週轉率	銷貨收入/(平均應收帳款額+應收帳款淨額－關係人)
		平均收現日數	365/應收帳款週轉率
		毛利率	營業毛利/銷貨收入
3	營業成本	應付帳款週轉率	銷貨成本/(平均應付帳款+平均應付帳款－關係人)
4	營業毛利	毛利率	同「毛利率」
		純益率	同「純益率」
5	應付公司債	負債占資產比率	負債/資產
		長期資金占固定資產比率	(股東權益+長期負債)/固定資產淨額
6	股本	股東權益報酬率	同「股東權益報酬率」
7	普通股股本	股東權益報酬率	同「股東權益報酬率」
9	研究發展費用	研究發展費用占營業費用比率	研究發展費用/營業費用
11	專利權	專利權占無形資產比率	專利權/無形資產
12	存貨	存貨週轉率	銷貨成本/平均存貨
		平均銷貨日數	365/存貨週轉率

[13] 此「平均」並非元素標籤，係指期初與期末元素實際值相加除以 2。

附錄4　財務性輔助資訊內容表－重要會計政策

編號	主要財務數字資訊	財務性輔助資訊－重要會計政策元素名稱	重要會計政策元素標籤
1	本期淨利	AccountingPolicieRevenueRecognitionSalesAllowanceReturns	收入認列及備抵退貨及折讓
3	營業成本	AccountingPolicieClaasificationCapitalExpenditureOperatingExpenditure	資本支出與收益支出的劃分
5	應付公司債	AccountingPolicieBondsPayable	應付公司債
8	庫藏股票	AccountingPolicieTreasuryStock	庫藏股
9	研究發展費用	AccountingPolicieClaasificationCapitalExpenditureOperatingExpenditure	資本支出與收益支出的劃分
10	營業費用	AccountingPolicieClaasificationCapitalExpenditureOperatingExpenditure	本支出與收益支出的劃分
11	專利權	AccountingPoliciesIntangibleAssets	無形資產
12	存貨	AccountingPoliciesInventories	存貨
13	退休金準備/應計退休金負債	AccountingPoliciesPensionCosts	退休金
14	長期投資	AccountingPoliciesLongTermInvestments	長期投資