第六章應用案例

應用案例 6.1

專利分析的文字探勘

專利是由國家授予發明人在有限時間內的專屬權利，以交換發明的揭露（請注意，授予專利的流程、對專利所有人的要求以及執行權利程度都因國而異）。這些發明的揭露是未來科技進展的關鍵。如果仔細分析，專利文件可以協助辨識新興技術、鼓勵新穎解決方案、促進共生夥伴關係以及增進企業能力與限制的整體意識。

專利分析為使用分析技術從專利資料庫抽取寶貴知識。維護專利資料庫的國家或國家團體（例如，美國、歐盟、日本）每年都會新增數千萬個新專利。如此大量的半結構化資料幾乎不可能有效地處理（專利文件通常包含部分結構化與部分文字資料）。使用半自動化軟體工具的專利分析是可以舒緩處理這些超大量資料庫的唯一方式。

專利分析的代表範例

伊斯曼柯達在全世界各地僱用了 5,000 多位科學家、工程師與技術人員。在 20 世紀，這些知識工作者與他們的前輩申請了近 20,000 個專利，讓公司成為全世界十大專利所有人之一。在持續改變的企業中，公司了解成功（或僅是存活）仰賴應用一個多世紀來的知識能力，想出科學與技術的新運用以及確保新運用的專利。

了解專利價值，伊斯曼柯達不僅創造新專利，也分析別人創造的新專利。利用努力工作的分析師與最新的軟體工具（包括 ClearForest Corp. 的特殊化文字探勘工具），柯達持續深入發掘原始資料（專利資料庫、新的研究檔案以及產品發表），以發展全面的了解。專利的適當分析帶給柯達這類公司許多助益：

- 它產生競爭性情報。了解競爭對手的現況可以協助公司制定因應措施。
- 它協助公司制定關鍵企業決策，像是推出或併購者或購買者想要購買的新產品、產品線以及／或技術。
- 它可以協助判斷與招募最佳與最聰明的新進人才，那些出現在公司成功關鍵專利上的人名。
- 它可以協助公司判斷其專利的未經授權使用，讓公司可以採取行動保護其資產。
- 它可以判斷創新的互補性以建立夥伴關係或促成併購或購買。
- 它避免來自建立類似產品的競爭以及保護公司免於捲入專利侵權訴訟。

使用專利分析做為豐富的知識與策略武器來源（攻防皆宜），柯達不僅可以存活，還可以利用創新與持續改變所定義出的市場區隔。

資料來源：P. X. Chiem, "Kodak Turns Knowledge Gained About Patents into Competitive Intelligence," *Knowledge Management,* 2001, pp. 11-12; Y. H. Tsenga, C-J, Linb, and Y-I, Linc, "Text Mining Techniques for Patent Analysis," *Information Processing& Management,* Vol. 43, No. 5, 2007, pp. 1216-1247.

應用案例 6.2

文字探勘幫助 Merck 更加了解和更能夠進一步服務顧客

Merck Sharp & Dohme (MSD) 為位於德國的全球性研究導向製藥公司，致力於解決全世界的健康照護需求。成立於 1891 年， MSD 發現、開發、製造與行銷疫苗與藥劑解決健康照護需求所面對的挑戰。

身為全世界最大的製藥廠之一， MSD 極度仰賴來自醫師的意見，協助提供更佳的服務給病患。預期的結果為提供罹患 AIDS、骨質疏鬆、心臟衰竭、偏頭痛與氣喘等病患更佳照護。

了解到知識發掘的重要性，許多年前， MSD 開發出分析程式，利用資料與文字探勘應用，更佳提升其資料與資訊資產價值。 MSD 使用 SPSS 的文字探勘技術分析它收集的不同原始資訊，然後利用資訊建立有效程式，更能解決醫師與病患需求。

挑戰

與其他專業一樣，健康照護產業的從業人員有其信念與意見。這也是 MSD 所面對的挑戰：它必須確實了解在其領域中醫師所說的話，然後將資訊傳達給其產品開發團隊，才能夠產生更佳的藥物與有效的藥物行銷活動。必須考慮 MSD 的目標群眾，這是個一點也不簡單的任務。一方為「先驅」醫師，對新的見識與研究結果抱持非常開放態度，並且很快地將科學結果實際應用。另一方為擁有「保守性格」醫師，遵照傳統方法並且想要照著書本來行事，花很多時間在研究治療方法，他們的意見都是遵照專家報告中的詳盡研究或與同僚交換意見。要成功， MSD 必須接觸所有類型的醫師。若要這麼做， MSD 必須使用來自不同來源的資訊，包括內部資訊與外部提供資訊。

解決方案

MSD 決定使用文字探勘與 SPSS 量化分析工具，以更佳了解有些資料來自各種通訊講座的調查資料，然後提供此寶貴資訊給行銷團隊。這些調查中衡量的特性包括醫師執業的年資、病患人數以及開放文字回應的問題。一旦取得必要的資料後，必須進一步進行資料特別分析，以了解顯著性以及廣泛特性之間的關係。 MSD 也使用收集來的資料進行側寫。此分析工具讓 MSD 可以將醫師分成幾個類型。根據行銷部門引進的指標來區隔醫師， MSD 決定出最能夠描述相關目標群體特性的行動目錄。

結果

對 MSD 而言，文字探勘，無結構化的文字資料分析是不可或缺的。文字探勘的功能性根據文字自然文法分析。它不完全仰賴關鍵字搜尋，而是分析語言的語法，並且「了解」內容。這麼一來，它成為發掘改善公司競爭位置不可或缺知識。

MSD 與 Gesellschaft fur Konsumforschung 委員會（消費者研究協會 (Association fo Consumer Research, GfK)）合作，使用委員會內醫師的每日「日記」記錄來了解有哪些製藥代表曾經到訪、傳達哪些產品訊息以及是否醫師在未來的處方中會包括這些產品。醫生記錄的產品對話文字探勘分析顯示出與數種處方行為有關的語音模式。這讓 MSD 可以強化產品與

行銷活動以及改善銷售代表的溝通技巧。由於 SPSS 與其文字探勘工具， MSD 了解哪些藥品的特性與資訊可以透過與醫師對話即可充分了解，以及哪些在行銷活動中使用的字詞必須再定義。

資料來源：SPSS, "Merck Sharp&Dohme," http://www.spss.com/success/template_view.cfm?Story_ID=185 (accessed May 15, 2009).

應用案例 6.3

謊言探勘

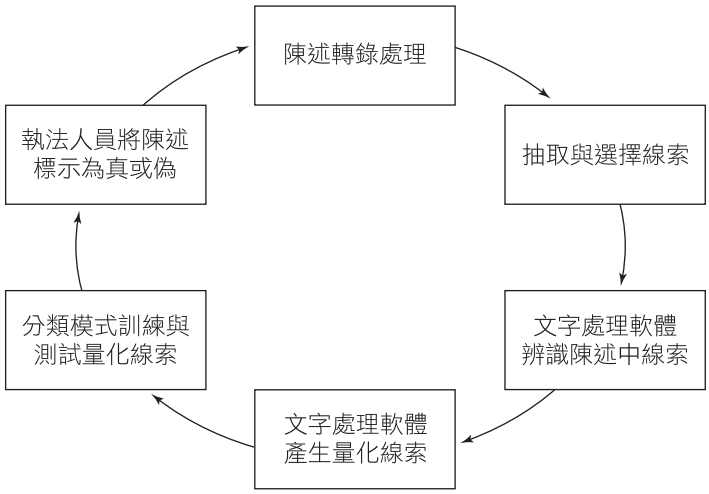
因為網路型資訊技術的進步以及全球化，以電腦為媒介的通訊持續深入每日生活中，形成新的詐騙場合。線上社群操作產生文字型聊天、即時訊息、文字訊息與文字增加的非常快。甚至電子郵件的使用持續增加中。隨著文字型通訊的大幅成長，人們透過電腦為媒介通訊來欺騙他人的可能也跟著增加，這些詐騙可以造成災難性的結果。

遺憾的是，一般來說人類對於詐騙偵測不在行。此現象在文字型通訊會更加嚴重。大部分詐欺偵測的研究（也稱為**可信度評估**(credibility assessment)）涉及面對面會議與訪談。然而，隨著文字型通訊的成長，文字型詐騙偵測技術變得很重要。

成功詐騙（謊言）偵測技術可以廣泛應用。執法者可以使用決策支援工具與技術調查犯罪、在機場執行安全篩檢以及監控可疑恐怖份子的通訊。人資專家可以使用詐騙偵測工具篩選應徵者。這些工具與技術篩選電子郵件，發覺公司主管的詐欺或其他不實行為。雖然許多人認為自己可以馬上辨識不老實的人，一份詐欺研究摘要指出平均只有 54%的人能夠準確判斷真實性 (Bond and DePaulo, 2006)。而在人們試圖偵測文字詐欺時，比率更低。

使用文字探勘與資料探勘技術的組合， Fuller 等人 (2008) 分析涉及軍事犯罪中警方感興趣對象的陳述。在這些陳述中，嫌犯與目擊者都必須用自己的話以書面再現事件的當時情況。軍法執法人員搜尋檔案資料來判斷陳述的真偽。這些決定是根據佐證與已偵結案件為基礎。一旦判定為真實或不實時，執法人員會取出辨識資訊並且將陳述交給研究團隊。總共， 371 筆可用陳述交由分析。 Fuller 等人 (2008) 使用的文字型詐欺偵測方法是根據**訊息特色探勘**(message feature mining)，仰賴資料與文字探勘技術中要素。圖 6.1 為此流程的簡單說明。

首先，研究人員準備處理資料。原始手寫陳述必須轉變為文字處理檔。第二，必須辨識特性（即線索）。研究人員辨識 31 種代表語言類別或種類特性，相當獨立的文字內容，可以使用自動化方法分析。例如，第一人稱代名詞像是 I 或 me 可以判斷為不需要周圍文字分析。表 6.1 列出在此研究中使用的特性分類與範例。



資料來源：C. M. Fuller, D. Biros, and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," *41st Annual Hawaii International Conference on System Sciences (HICSS)*, Big Island, HI, IEEE, pp. 80-89.

圖 6.1　文字型詐欺偵測流程

表 6.1　詐欺偵測中使用語言特性的類別與範例

編號	建構（類別）	範例線索
1	數量	動詞量、名詞片語量等
2	複雜性	平均線索數、平均句子長度等
3	不確定性	修飾語、情態助動詞等
4	非直接性	被動語態、物化
5	表達性	情緒化
6	多元性	詞彙多元性、贅詞等
7	非正式	印刷錯誤比率
8	特定性	時空資訊、感知資訊等
9	影響	正面影響、負面影響等

從文字陳述中抽取特性並且輸入一般檔案中進行進一步處理。使用幾種特性選擇方法以及 10 摺交叉驗證，研究人員比較三種常用資料探勘方法的預測準確度。他們的結果顯示類神經網路模式績效最佳，有 73.46%的測試資料樣本預測正確率；決策樹居次，有 71.60%正確率；以及邏輯迴歸最差，正確率為 67.28%。

結果顯示自動化文字型詐欺偵測有潛能協助文字詐欺並且可以成功應用在實際資料中。即使受限於文字線索，這些技術的正確率超過大部分其他詐欺偵測技術。

資料來源：C.M. Fuller, D. Biros and D. Delen, "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection," *41st Annual Hawaii International Conference on System Sciences (HICSS)*, Big Island, HI, IEEE, pp. 80-89； C. F. Bond and B.M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Reports,* Vol. 10, No. 3, 2006, pp. 214-234.

應用案例 6.4

飛躍過文字

文字探勘經證實為從數位形式儲存書面文件抽取組織知識的寶貴工具。分析師使用文字探勘軟體，透過模式辨識，鎖定關鍵問題領域。例如，在航空產業中的公司可以應用意外報告中文字探勘提升組織知識品質。透過使用文字探勘，他們可以及時研究機械、組織與行為問題。

航空公司使用詳盡與系統化操作分析。在造成問題的事件發生時，就必須準備意外報告。文字探勘技術可以用來自動偵測大量意外報告中關鍵議題。航空公司維護的大資料庫在人類可解之意義的解釋上受限，而專有名詞的解釋也不盡相同。

Aer Ligus (aerlingus.com) 檢視 1988 年 1 月至 2003 年 12 月產生的意外報告找出可能的模式與相關性。 Aer Lingus 使用 Megaputer 的全面資料與文字探勘軟體 PolyAnalyst (megaputer.com)。研究目的為開發調查人員定期用來辨識與意外種類、地點、時間與其他詳細資料有關模式與關聯。

在意外報告中判斷出最常發生的項目。 PolyAnalyst 中的字詞雖然不完整，但是提供寶貴的文字分析起點。它也產生資料中關鍵字詞清單（或他們的語義相關）。建立常見字詞報告，含括已辨識字詞與頻率。目標為辨識有趣的叢集。描述性摘要包括將描述性說明分成至的有意義字詞組。例如，關鍵字詞 spillage（溢出）可以與四個其他關鍵字詞有關：food（食品）、 fuel（燃油）、 chemical（化學）與 toilet（廁所）。從字義上來看， food 與 coffee、 tea 以及 drink 相關，因此， food 成為分類節點，以及報告中原本歸屬 spillage 食品字詞就與 food 配對。

航空公司意外報告的文字探勘可以辨識出改善安全的重要根本原因。文字探勘也可以與大批意外報告資料一起使用，驗證之前定義理論與常識知識以及收穫與新增知識的新模式。

資料來源：J. Froelich, S. Ananyan and D. L. Olson, “Business Intelligence Through Text Mining,” *Business Intelligence Journal,* Vol. 10, No. 1m 2005, pp. 43-50.

應用案例 6.5

使用文字探勘進行研究文獻調查

研究人員在進行相關文獻調查與回顧時，面對著越來越複雜與繁重的任務。要擴大相關知識，必須努力收集、組織、分析與吸收文獻中既有資訊，特別是專有領域部分。相關領域的顯著文獻大幅增加，以及甚至在傳統上認為不相關的重要文獻也在增加中，如果要做好詳盡工作，研究人員的任務更重大。

在新研究中，研究人員的任務更沉悶與複雜。如果是傳統上需要許多人力檢視發表文件，要找出其他人的相關研究很困難。即使有一群辛勤的研究生或提供協助的同儕，要找出所有潛在相關發行研究還是個問題。

許多學術會議都是每年舉行。除了擴展研討會目前焦點的知識外，主辦單位通常希望提供其他小型討論會或工作坊。在許多例子中，這些額外活動的目的為介紹相關領域重要研究的參與者並且試圖確認相關研究興趣與焦點的「下一件大事」。確認該類小型討論會與工作坊的主題通常很主觀，而不是客觀地由既有與新研究中決定。

在最近的研究中， Delen 與 Crossland (2008) 提出大力協助與提升研究人員能力的方法，透過文字探勘應用，進行大量已發表文獻的半自動分析。使用標準數位圖書室與線上著作搜尋引擎，作者下載並且收集所有在管理系統領域的三大主要期刊論文： MIS Quarterly (MISQ)、 Information Systems Research (ISR) 與 Journal of Management Information Systems (JMIS)。為了讓三份期刊的時間一致（為了進行比較時間研究），此研究使用電子版發行日做為研究開始日期（即 JMIS 其電子期刊日期自 1994 年開始）。在每一篇論文中，他們抽取出論文名稱、摘要、作者名單、發表關鍵字、冊數、期數以及發表年份。他們之後下載所有文資資料至一個簡單資料庫檔案中。為了識別分析，也將每篇論文的期刊種類結合至資料集檔案中。收集中省略編輯意見、研究員意見與本文摘述。表 6.2 以表格顯示出資料的呈現方式。

在分析階段中，他們選擇只使用論文摘要做為資訊抽取的資料來源。他們選擇不包括關鍵字，基於兩個主要的原因： (1) 在常態情況下，摘要中都已經含括關鍵字，所以在分析中含括列出的關鍵字可能意味著重複資訊以及給予不當權重；與 (2) 列出的關鍵字可能是作者想要提供的論文關聯字詞 (可能沒有真正出現在論文中)，所以可能會將不可量化的誤差引進分析內容中。

此第一份探究性研究檢視三份期刊的時間觀點（即研究主題的時間演進）。為了執行時間研究，他們將每一份期刊的 12 年時間（ 1994 年至 2005 年）分成三年期。此架構產生有 12 個互斥資料集的 12 個文字探勘實驗。此時，在每一個 12 個資料集中，他們使用文字探勘抽取摘要中最能代表論文的字詞。就三份期刊的發行時間，檢視時間改變結果並且製成表格。

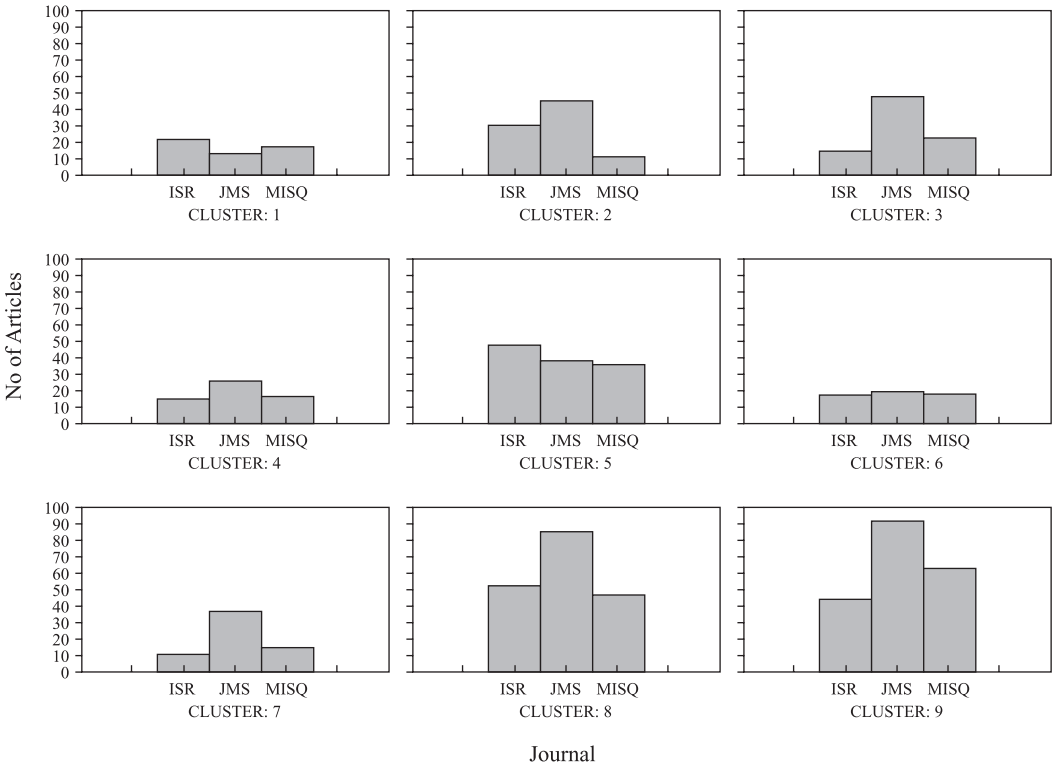
第二份探索論文使用完整資料集（包括所有三份期刊與四個時期），他們執行分群分析。分群是最常見的文字探勘技術。此研究中使用分群來辨識論文的自然群組（放入不同個別叢集），然後列出最能代表叢集的敘述字詞。他們使用奇異質分解減少字詞 ── 文件矩陣維度以及最大期望演算法建立叢集。他們進行幾個實驗判斷叢集的最佳數目，最後決定為 9 個。在建立 9 個叢集後，他們從兩個觀點分析這些叢集內容： (1) 期刊種類的代表（參考表 6.2）以及 (2) 時間的代表。此概念為探索三份期刊的潛在差異以及／或共同處與

這些叢集的潛在改變，即回答「特定單一期刊是否有代表不同研究主題的叢集？」與「這些叢集是否有隨時間變化的特性」等問題。使用代表結果的表格與圖示，他們發現並且探討幾個有趣的模式（更詳細訊息，請參考 Delen and Crossland, 2008）。

資料來源：D. Delen and M. Crossland, "Seeding the Survey and Analysis of Research Literature with Text Mining," *Expert Systems with Applications,* Vol. 34, No. 3, 2008, pp. 1707-1720.

表 6.2　結合資料集中含括欄位的代表表格

期刊	年份	作者	標題	卷期	頁碼	關鍵字	摘要
MISQ	2005	A. Malhotra, S. Gossain, and O. A. El Sawy	Absorptive capacity configurations in supply chains: Gearing for partner-enabled market knowledge creation	29/1	145-187	management supply chain absorptive capacity interorganizational information systems configuration approaches	knowledge continual value innovation is driving supply chains to evolve from a pure transactional focus to leveraging interorganization partnerships for sharing	The need for
ISR	1999	D. Robey and M. C. Boudtreau	Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications		165-185	organizational transformation impacts of technology organization theory research methodology intraorganizational power electronic communication misimplementation culture systems	Although much contemporary thought considers advanced information technologies as either determinants or enablers of radical organizational change, empirical studies have revealed inconsistent findings to support the deterministic logic implicit in such arguments. This paper reviews the contradictory
JMIS	2001	R. Aron and E. K. Clemons	Achieving the optimal balance between investment quality and investment in self-promotion for information products		65-88	information products internet in advertising product positioning signaling signaling games	When producers of goods (or services) are confronted by a situation in which their offerings no longer perfectly match consumer preferences, they must determine the extent to which the advertised features of



資料來源：D. Delen and M. Crossland, "Seeding the Survey and Analysis of Research Literature with Text Mining," *Expert Systems with Applications,* Vol. 34, No. 3, 2008, pp. 1707-1720.

圖 6.2　在九個叢集中三份期刊論文數的分佈

應用案例 6.6

網站逮捕

我們通常都在所處的環境外搜尋問題答案。然而，通常在問題都來自內部。在採取行動對抗全球恐怖份子部分，美國國內極端份子團體通常都被忽略。但是，國內極端份子對美國安全帶來顯著威脅，因為透過網路使用，他們所擁有的資訊以及越來越提升的能力都可以與全世界的極端份子團體互通有無。

要監控網路上的內容很困難。研究人員與有關當局需要優越的工具來分析與監控極端份子的活動。亞利桑納州大學的研究人員，在國土安全部門與其他機構的贊助下，開發出網站探勘方法尋找並且分析美國國內極端份子經營網站，以了解這些團體的網際網路使用。極端份子使用網際網路通訊、取用私人訊息以及線上募款。

此研究方法開始時收集相關極端份子與恐怖份子網站的基準資料。執行超連結分析，連結至其他極端份子與恐怖份子網站。與其他網站的相互連結對於評估不同團體的目標相似性很重要。下一步驟為內容分析，可以進一步根據不同屬性，分類網站，像是通訊、募款以及理念分享等。

根據連結分析以及內容分析，研究人員辨識出 97 個美國極端份子與仇恨團體網站。通常，社群之間的聯繫並不一定代表任何合作。然而，找出共同利益團體之間的數個聯繫幫助對共同旗幟下共同利益團體分群。使用資料探勘自動化流程的進一步研究有個全球目標，判斷國際仇恨與極端份子團體與美國國內團體的連結。

資料來源：Y. Zhou, E. Reid, J. Qin, H. Chen and G. Lai, "U.S. Domestic Extremist Group on the Web: Link and Content Analysis," IEEE Intelligent Systems, Vol. 20, No. 5, September/October 2005, pp. 44-51.

應用案例 6.7

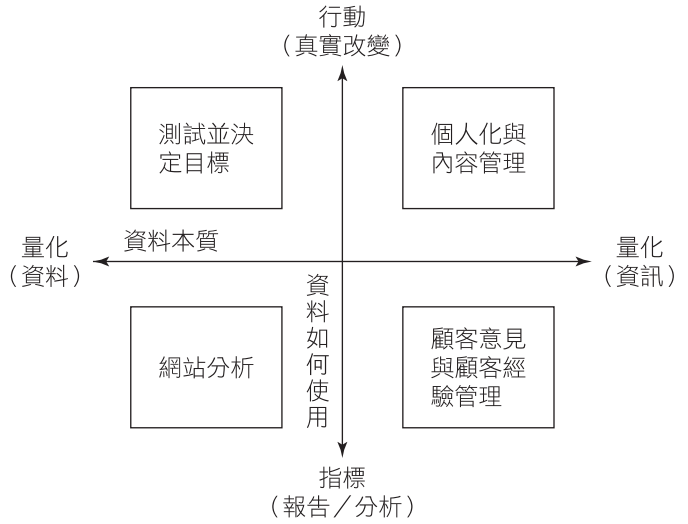
網站最佳化生態系統

似乎所有在網路上的每件事都可以測量，每次點擊都可以紀錄，每次瀏覽都可以捕捉，每次到訪都可以分析，以持續與自動最佳化線上經驗。遺憾地是，線上通路的「無限可測量性」與「自動最佳化」概念比所能理解的要來得複雜許多。假設任何網站探勘的任何單一應用可以提供了解網站訪客行為是否不實與有潛在風險的必要知識。理想上，要全面了解顧客經驗只能使用量化與質化資料來捕捉。有前瞻性想法的公司，像是之前章節討論過的（即， Ask.com、 Scholastic.com 以及 St. John Health System）已經採取步驟捕捉與分析全面的顧客經驗，產生顯著增加的財務成長與顧客忠誠度以及滿意度優勢。

根據 Peterson (2008)，網站最佳化的輸入可以分成兩個面向，說明資料本質與資料使用方法。其中之一為資料與資訊；主要是量化資料與質化資訊。另一個為指標與行動；指標為驅使行動的報告、分析與建議、網站持續進行流程中的改變以及行銷最佳化。這些維度建立的每個象限可以提升不同技術並且產生不同輸出，但是與生物生態環境系統非常類似，每個技術利基都與其他利基互動，以支援整個線上環境（參考圖 6.3）。

大部分的人認為網站最佳化生態系統以對網站訪客的瀏覽行為對數、語法分析以及報告功能來定義。此功能的基本技術為一般所稱的網路分析。雖然網路分析工具提供無價的知識，了解訪客行為是量化不同網頁點擊數的質化判斷。幸運地是有兩類應用，可以提供更質化的線上訪客行為，目的為報告整體使用者經驗以及由訪客與顧客提供的直接意見回饋：顧客經驗管理 (customer experience management, CEM) 與顧客意見 (voice of customer, VOC)：

- 透過已經定義的多重步驟流程，整合、探勘與視覺化大量資料、報告線上行銷與訪客購買記錄、總結頁面上訪客互動資料以及總結訪客流，網路分析應用著重於「何處與何時問題」。
- 藉由收集與報告網站訪客的直接回饋意見、與其他網站與離線通路以及支援未來訪客行為的塑模，顧客意見應用鎖定「誰與如何」問題。
- 藉由偵測網路應用議題與問題、追蹤與解決企業流程與使用性阻礙、報告網站績效與可用性、執行即時警告與監測以及支援已觀察訪客行為的深度診斷，顧客經驗管理應用鎖定「什麼與為什麼」問題。

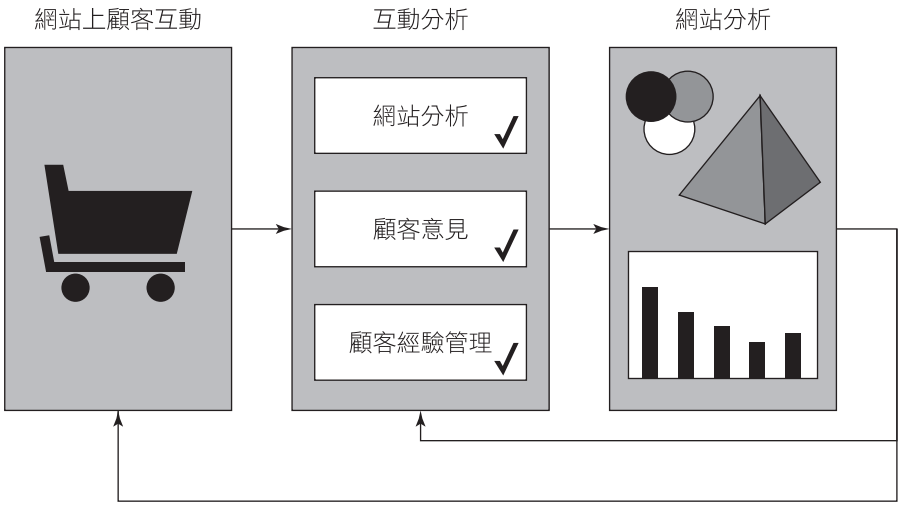


資料來源：E.T. Peterson， The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization (2008), foreseeresults.com_peterson_WebAnalstics.html (accessed on May 22, 2009).

圖 6.3　網站最佳化輸入二維圖

三個應用都需要有完整的顧客行為圖，每個應用都扮演獨特與寶貴角色。以網站最佳化生態系統為基礎的網站分析、CEM 與 VOC 應用，支援線上企業正面影響期望結果（圖 6.4 顯示網站最佳化生態系統流程圖）。此類似但是獨特應用可以讓網站經營者確認、回應與因應個別網站所有者正面臨的挑戰。最佳化流程的基本為測量，收集可以轉變為有形分析的資料與資訊以及建議使用網站探勘工具與技術的改善。適當使用時，這些應用可以進行聚合驗證，結合從相同對象所收集的不同資料集，提供更豐富與深入的行為了

解。聚合驗證模式當中的多種原始資料說明可以增加結果分析深度與豐富性的整合群體，形成網站最佳化生態系統的架構。一邊為 VOC 應用的主要質化輸入，另一邊則為 CEM 的主要量化輸入，以資料發掘的支援關鍵要素縮減落差。適當執行時，三個系統都是從同樣的對象取樣。這些資料的組合，不論透過資料整合專案或簡單透過執行良好分析的流程，比其他個別生態系統成員更可以支援更可以行動的想法。



資料來源：E.T. Peterson, "The Voice of Customer: Qualitative Data as a Critical Input to Web Site Optimization," 2008, foreseeresults.com/From_Epeterson_webAnalysis.html (accessed on May 22, 2009).

圖 6.4　網站最佳化生態系統流程圖