運用公開來源文件於商業情報探勘

楊新章
國立高雄大學資訊管理學系
yanghc@nuk.edu.tw

黃怡翔
國立高雄大學資訊管理研究所
dream788@hotmail.com

摘要

情報蒐集與分析一直以來皆在企業成長扮演重要的角色，傳統的情報管理流程大多為隱密且需要大量的人力處理，同時其蒐集之情報量通常頗為稀少且蒐集過程可能具危險性，因此公開來源情報(open source intelligence)近年來逐漸成為情報蒐集與分析的主流，目前公開來源情報的處理大仍仰賴人力進行，必須耗費大量的人力與時間，自動化的公開來源情報處理對於現今之公開來源情報而言勢必無法避免，本文將基於一文本探勘過程來進行公開來源情報管理，主要的目的為自動的偵測對企業有益的事件資訊，本文的主要貢獻為提出了一高階的公開來源情報探勘方法，其結果將有利於國家安全、個人知識管理、與企業成長，

關鍵詞：公開來源情報、文本探勘、自我組織圖、商業情報

運用公開來源文件於商業情報探勘

壹、緒論

情報(intelligence)之蒐集與分析一直以來被認為在軍事與商業戰爭中具有關鍵的角色。歷史上多有獲得關鍵情報而獲得軍事或商戰勝利的案例。情報之用途，主要是能「料敵機先」，收「知己知彼，百戰百勝」之效。故情報之蒐集，應具有可靠性與廣泛性，求來源之充足與正確，以利分析。情報之分析，則應具有預測性，求關鍵事件之預知，而能防患未然。若能持續的、廣泛的、且可靠的獲得情報，再加以進行準確分析，則不論在戰爭行為或商業利益上，皆可獲得鉅大的進展。故自古以來，情報的蒐集與分析一直備受（政治或企業）當權者重視。

情報管理為一循環過程，完整的情報循環(intelligence cycle)包含下列步驟：

1. 規劃與指引(planning and direction)：決定要監測與分析何事物。
2. 蒐集(collection)：蒐集原始資料。
3. 處理(processing)：精煉與分析資訊。
4. 分析與產出(analysis and production)：將經處理後的資料轉換成情報完成品，包含對情報進行整合、校堪、評估、與分析。
5. 發佈(dissemination)：將處理結果提供予客戶。

上述流程為一循環，當情報發佈後可再進行下一階段之情報循環。其中也可以在完整循環中建立次循環，例如第 2-4 步驟。事實上，若將情報循環視為一資訊系統，則規劃與指引步驟可視為系統之輸入，發佈步驟可視為系統之輸出，其間之蒐集、處理、分析與產出步驟則可視為資料處理過程。以下便針對此處理過程作一探討。

在情報蒐集上，傳統以來，不論在軍事或商業上，情報大都藉由秘密、隱蔽的管道獲得。主要的原因是具有價值的情報通常具有機密性與敏感性而不對外公開。機密情報的取得因而時常經由不合法的方式取得，造成情報蒐集過程具有極大風險。另一方面，秘密情報蒐集因管道之特殊與稀少性，加上反情報(counter-intelligence)蒐集之可能，使得情報之可靠性亦時無保障。由於這些因素，使情報蒐集之過程一直以來皆蒙上一層隱晦的面紗，情報蒐集者（情報員）之故事亦在坊間多所流傳。由傳統管道所蒐集之情報，則具有資料量稀少、資料需驗證、通訊管道難以建立等缺點。

傳統之情報分析通常為專業人員進行。情報分析師依據其執行準則與流程[1]對情報進行分析。其間需運用大量之分析技巧與個人經驗與智慧。此類技巧之運用與經驗之累積需耗費極大的時間與心力，故無法大量為之。因此各國政府或企業莫不把具備高度分析能力之情報分析師視為重要資產。誠然，一具有高度情報分析能力之人員可為機構帶來巨大利益，然終就人力有限且訓練曠日費時，要依賴此方式進行情報分析僅國家及大型企業具有能力進行。

一般蒐集情報之方法，分為人類情報(human intelligence, HUMINT)、信號情報(signal intelligence, SIGINT)、度量與簽章情報(measurement and signature intelligence,

MASINT)、影像情報(imagery intelligence, IMGINT)、地理空間情報(geospatial intelligence, GEOINT)、財務情報(financial intelligence, FININT)、技術情報(technical intelligence, TECHINT)、與公開來源情報(open source intelligence, OSINT)等。以上之各種情報蒐集方法各有其應用領域與限制及其優劣點。一般而言，上述各種情報蒐集方法皆需運用專業技術且需花費大量財力與人力，非一般個人或企業所能負荷，因此多由國家級專門機構進行。然而其中之公開來源情報卻具有成本低、資訊即時、資訊量充足等優勢，對情報蒐集而言成為一新興且具吸引力之管道。

公開來源情報係指自公開來源，如新聞、政府公報、企業財報、機關網頁等，所能獲取之情報。「公開」代表大眾可以付費方式或免費方式自由取得，乃相對於隱匿或機密而言。公開來源情報大約來自下列各來源：

1. 媒體：報紙、雜誌、廣播、電視、網路媒介等。
2. 網路社群：社交網站、視訊分享網站、維基百科、部落格、社會性書籤網站等。
3. 公開資料：政府公報、官方文件如財務報告、戶政資料、公聽會、國會質詢、記者會、演講、環境影響評估等。
4. 觀察報告：業餘無線電監聽者、飛機觀測者、公開之衛星圖與地圖等。
5. 學術專業：學術會議、學術組織、學術論文、專家等。

公開來源情報在蒐集過程中和傳統方法最大的不同在於傳統情報中，蒐集過程是最困難的一部份，特別是常常要自不合作的目標獲取情報。這部份對於公開來源情報卻是最簡單且花費了最少成本的部份。公開來源情報最大的困難在於自大量的資料中偵測出相關且可靠的來源。以往對公開來源情報的分析需仰賴具備高度能力，足以立即處理這些資訊的專家為之。事實上，直至目前為止，公開來源情報分析絕大部份仍需仰賴人力進行。

如前述，公開來源情報之分析目前皆由人力進行。然而人力之涵蓋範圍與即時性皆有所限制。目前來自上述之公開來源資料量極為鉅大，早已超乎人類可以處理的極限。自動化的公開來源情報處理成為一必要且逐漸熱門之議題。然而自動化處理需克服下列困難：

1. 大量資料處理：公開來源之資料量十分鉅大且格式不一致。因此不論是線上或離線的處理皆十分困難。尤其是當資料來源是即時產生的（如新聞），如何快速的擷取與處理大量資料是一頗為困難的問題。
2. 情報分析：對於公開來源情報而言，情報來源不虞匱乏，甚至可以說是多到氾濫。除了上述之資料處理困難外，如何自如此大量之資料中過濾、偵測、摘錄出重要情報成為一十分困難之課題。傳統依人類智慧之分析方法顯已不可行，一套半自動乃至全自動之情報分析方法將是不可或缺的。惟該自動化方法必須克服大量資料與即時性要求，另必須針對不同決策需求運用不同的分析方法，亦需滿足可調整性(scalability)之要求。
3. 使用者介面：自動化情報處理和傳統情報處理方式最大的不同便是情報使用者（決策者）不需仰賴人力來蒐集與分析情報。因此必須建立一適當之人機

介面以利該等決策者制定目標與檢視分析過程與結果。在建立此類使用者介面時，需參卓各項質化與量化指標以利呈現。

本論文將發展一自動化程序來處理與管理公開來源情報以應用於商業情報探勘應用上。我們將使用來自於公開來源之文字文件，首先將其進行分群以獲得文件間之關聯。而後再應用一情報偵測過程在分群結果上以發掘有用的事件與主題。所偵測的情報將可應用於企業決策支援上。

本論文的組織如下：在第二節中我們將進行文獻探討。第三節將介紹文件的前置處理與分群過程．第四節將介紹本文所提出的情報探勘方法．在第五節中將呈現實驗結果．最後在第六節中為結論與討論．

貳、文獻探討

公開來源情報之相關研究，早期大多僅限於其作業規範，並由國防單位進行。例如美國國家情報總監下設公開來源中心，另北約組織(NATO)於 2001 年出版了一有關公開來源情報之操作手冊[2]，其中詳述了有關公開來源情報之各層面，如資料來源、可用軟體、可用服務、及處理循環中各步驟之說明等。手冊中也提供了廣泛的參加資料，如公開來源情報相關網站與訓練教材。本手冊是瞭解公開來源情報之重要啟始知識。北約組織也另外出版了一本有關公開來源情報之文選[3]。另針對網際網路之公開來源情報處理，北約亦發表了相關著作[4]。這一系列著作可以說是瞭解公開來源情報的踏腳石。然而其中所敘述的，大多為描述公開來源情報之規範與如何進行公開來源情報管理，對於自動化處理模式幾無著墨。

近年來自動化公開來源情報處理吸引了來自於學術界，尤其是計算機領域學者，的關注。和情報界不同的是，這些學者較關心的是如何設計一程序以取代人力進行公開來源情報處理。由於情報分析可以說是要從資料中發掘出可用之情報，與資料探勘之目的相近，因此很自然的可以引用資料探勘技術在公開來源情報之處理機制上。然而這個想法直至近年才真正開始吸引一些學者投入並發表其研究成果。事實上，應用資料探勘方法於公開來源情報之應用直至目前仍甚少被探索。國際上也很少發表這方面的論文，證明此領域仍在嬰兒期。目前較相關之會議為已經舉辦了三屆的 International Symposium on Open Source Intelligence and Web Mining (OSINT-WM)。此會議第一屆是於 2008 年於英國倫敦舉行，其後 2009 年於西班牙巴塞隆納舉行，今年(2010)年則於丹麥歐恩塞(Odense, Denmark)舉行。這個會議主要便是希望藉由資料探勘方法，尤其是網路探勘(Web mining)技術與社交網路分析，進行公開來源情報之分析，可以說是目前世界上少數聚集這方面研究之重要會議。除了此會議之外，另有一些論文散見於不同的國際會議中。整理後得知，此方面之研究大多以歐洲國家為主。以下則擇其重要之論文進行探討。

義大利的 SYNTHEMA 公司之 Baldini 等人於 2007 年開始發表了數篇論文描述其所發展的公開來源情報處理平台 SPYWatch[5-7]。他們提出一架構來進行公開來源情報之蒐集、處理、分析、產出、與發佈。此系統的核心技術在使用 K-means 演算法將文件分群後再進行分類。特點是本系統可處理來自不同語文文件之情報。

奧地利的 Sail Technology 之 Pfeiffer 等學者[8]則發表了基於 MPEG-7 之處理平台 Media Mining System。此系統之輸入可以為不同型式之公開來源資料，如衛星影像、電視影像、網頁與 RSS 輸入等。這些原始輸入隨後被處理以萃取其內容。產出的內容則可在該公司之 Media Mining Server 中被檢索、分析與檢視。Media Mining System 可用於早期預警、資訊分享、與風險評估上。

英國 Innovation Works 公司之 Vincen 等人[9]則提出一集中式架構以融合來自不同來源的資訊以提供緊急服務所用。他們的技術主要的核心是使用機率加強知識本體(probabilistic enhanced ontology)並配合多功能的服務介面與使用語意特徵。本系統於發表時尚未成熟，其主要目標是要能夠達成情境認知(situation awareness)與影響評估(impact assessment)。

Badia 等學者[10]分析文字文件中的語句以獲得文件之時空資訊以提供公開來源情報使用。他們先將語句轉換成主詞－動作－受詞的型式，再依據剖析程式所提供之提示資訊與特定的語法樣式來找出語句中的時、空資訊。

美國德州的 Austin Info System 之 Palmer 早於 2005 年發表論文[11]則提出了一語意比較量度以進行事件分析(event analysis)。他使用 Lavalette 分布取代了較早所使用的語料庫分析[12]。他的系統主要的特點是可以偵測事件間之關聯。

歐盟執委會(European Commission)的聯合研究中心(Joint Research Centre, JRC)建立了一個二階段的事件萃取系統[13]。在第一階段中，他們建立了一稱為歐洲媒體監測器(Europe Media Monitor, EMM)的新聞報導蒐集平台。這些新聞報導會被分類與分群以供第二階段使用。在第二階段中，他們使用了兩個方法，其一是 JRC 所發展的 NEXUS 系統[14-15]。此系統以分群為中心，採取簡淺語言學方法來自某一主題群組中萃取資訊。其二是芬蘭赫爾辛基大學所發展了 PULS 系統[16-17]。此系統則較為深入的分析新聞之內容，因而允許使用者自未明之新聞中發掘事件。

Wiil 等人[18-19]以圖論的方法來來分析恐怖份子網路(Terrorist Network)。本文之特殊性在於其分析著重於連結(link)而非傳統該類網路所著重之節點(node)。Bartik[20]的研究試圖將文本資料進行分類。他除了使用傳統的 tf-idf 加權法來描述文件內容外，也採用視覺特徵，即文字所出現的位置作為分類的依據。Dawoud 等人[21]則結合數個社交網路分析常用的量度成為一全域性量度，以度量恐怖組織之組織強度。Liu 與 Sandfort[22]則針對公開源碼，分析其與公眾參與對社會創新的影響。雖然她們的研究亦與 Open Source 相關，但與此處之公開來源情報較無關聯。Neri 等人[23]則以義大利總理之性醜聞為例，探討如何分析、標示、分群新聞文件並發掘其隱含之關聯與情感方向。

參、文件前置處理與分群

一、 文件前置處理

為了將文件轉換為適合訓練使用，我們必須加以處理以轉換為向量型式。我們首先必須去除與內容無關之網頁語言標記，將網頁轉換為本文檔。而後進行斷詞(segmentation)，將本文轉換為字詞之集合。標準的字詞處理程序，如常用字去除(stopword

elimination)、字根還原(stemming)、關鍵字選取(keyword selection)等也被運用以降低關鍵字之數量，即字彙集(vocabulary)之大小。最後我們再利用向量空間模型(vector space model)將網頁 P_i 轉換為一向量 $\mathbf{P}_i$。

二、 文件分群：

文件經由轉換為向量後，接下來我們想對文件進行分群(clustering)，在此本研究使用的是自我組織圖。自我組織圖的主要概念是透過計算文件向量與神經元突觸權重向量的距離來映射文件至神經元上，藉此將文件做分群。在此將其訓練過程描述如下：

Step 1: 設定訓練所需參數

第一個步驟先設定訓練所需的參數，其中包含輸入層神經元數、輸出層神經元數、輸入文件(向量)筆數、學習速率 $\alpha(t)$、學習次數 t，並以亂數設定突觸權重向量 $\mathbf{w}_j$。

Step 2: 執行學習流程

執行學習流程表示將一文件向量進行學習步驟的過程，其步驟如下所示：

Step 2.1: 隨機輸入一文件向量 $\mathbf{v}_i$ 進行訓練。

Step 2.2: 尋找優勝神經元。

Step 2.3: 更新突觸權重的向量：

當一文件向量被映射至其優勝神經元上時，此神經元與其附近之神經元之突觸權重會受到新加入之文件向量的影響而改變，所以必須更新突觸權重的向量。

Step 2.4: 重複 Step 2.1~2.3 直到所有的文件向量都經過訓練一次。

Step3: 檢查停止條件

令 $t = t + 1$；假如 t 達到了預先設定的總學習次數 T 時，則訓練完成；否則就減少學習速率 $\alpha(t)$，並縮減鄰近區域的範圍，回到 Step 2 繼續執行訓練。

肆、情報探勘

經過自我組織圖的訓練後，我們將對神經元進行標記處理(labeling process)，並產生文件分群圖(document cluster map，DCM)。所謂的標記處理即將先前文件於訓練完成之自我組織圖之優勝神經元標示出來，如此便可以知道那些文件與文件間是相似的。我們將文件分群圖之標記方法敘述如下：

在 DCM 中，概念上每一個神經元即代表一些文件的集合，且標記於此神經元內的文件具有高度字詞同時出現(co-occurrence)的特性，因此被標記再同一或鄰近神經元上的文件彼此間有一定程度的語意相似程度。

產生 DCM 所使用的方法為計算文件向量與各神經元突觸權重向量的距離。我們將第 i 筆文件向量 $\mathbf{P}_i$ 與所有神經元的突觸權重向量進行比較。假設第 i 筆文件向量與第 j 個神經元的突觸權重向量距離為最小，則將此文件向量標記至此神經元上。亦即滿足下式：

$$\left\|\mathbf{P}_i - \mathbf{w}_j\right\| = \min_{1\le k\le J}\left\|\mathbf{P}_i - \mathbf{w}_k\right\| \qquad (1)$$

其中 J 為神經元總數 $\mathbf{w}_j$ 為神經元 j 之突觸權重向量。我們將所有文件向量之標記神經元記錄下來，便可得到 DCM。先前提過本文的文件向量是依各文件所包含的關鍵字來表示，因此具有多數相同關鍵字的文件在理論上表示其相似程度很高，所以在標記的過程

當中有很大的機會會被標記在同一個神經元上，也就是說被標記在同一個神經元上的文件在語意上具有較高的相似程度。因此包含相同字詞的文件會被標記在同一個或相鄰的神經元上。此外，由於神經元數目通常會小於文件數目，所以會有多份文件被標記在同一神經元上。因此一個神經元便構成一文件群集。

獲得文件分群圖後，我們便可據以進行情報探勘。本文將進行三種情報探勘技術，分述如下：

（一） 文件分群主題偵測：

我們可以發掘文件群組之重要關鍵字詞作為該分群主題，經由一字詞標示過程，我們可以獲得一關鍵字詞分群圖（keyword cluster map, KCM），與 DCM 不同的是，在 KCM 中每一神經元內所包含的是一些字詞的集群，且這些字詞為其對應之文件中的常用字詞，換句話說這些字詞在其被標記的神經元之突觸向量中具有一定程度的權重值。

本文中的文件向量是以二元向量來表示，因此經由自我組織圖訓練後的突觸權重向量理論上最好的情況應是 0 或 1，向量 0 代表字詞對於此神經元完全不重要，相反的，向量 1 表示字詞對於此神經元具有很大的重要性；但事實的情況並不會只出現 0 或 1 兩種極端情況，因為在訓練過程中每一神經元可能都會受到其鄰近神經元之修正。所以本文設計以下方法來產生 KCM：檢視第 j 個神經元內的突觸權重向量 $\mathbf{w}_j$，若某一字詞所對應之元素其值超過一預先設定之臨界值，則將此字詞標記在此神經元上。這裡所提到的臨界值即一介於 0 到 1 之間的數值，其中越接近於 1 表示此字詞在神經元中所代表的重要性越高，因此通常都設定為一接近於 1 的數值，但具有彈性，可根據樣本資料的特性來調整臨界值的大小。

在標記處理之後，一個神經元會被數個字詞所標記，如此即形成了一個字詞群集。在 KCM 裡，在文件中常常同時出現的字詞會被標記到相同或鄰近的神經元上。例如，「微軟」與「比爾蓋茲」經常在一份文件中同時出現，所以它們會被標記到相同或鄰近的神經元，因為它們所對應的元素在轉換成文件向量時都會同時設定為 1，因此，神經元便會試著同時去學習這兩個字詞。相反的，不同時存在相同文件中的字詞在圖中就會被標記在距離較遠的神經元。如此我們就可以依據兩個字詞在 KCM 中所對應的神經元來發掘它們彼此之間的關係。

（二） 特定事件偵測：

特定事件指使用者事件設定之事件，例如，一金融業者會關心是否發生了有關於歐元匯率之事件，他可以設定一些與此類事件相關之關鍵詞，我們則可以監看新進文件以偵測該類事件是否發生，令 $E = \{e_i\}$ 為使用者所設定用來偵測某事件的關鍵字詞集合，首先我們先發掘每一 e_i 所屬之文件集合，令 C_i 為 e_i 所屬之文件集合，若 e_i 為 KCM 中神經元 i 之關鍵字集合中之成員，C_i 即為該神經元所對應之 DCM 中之文件分群，若 e_i 出現在多個關鍵字詞群組中，我們選擇突觸權重向量中具有最高之對應成份之分群為 C_i，決定了各事件關鍵字詞之對應文件群組後，我們可以用下列方式來偵測特定事件，一新進文件 D_I 首先依第三節所述進行前置處理並轉換為一文件向量 $\mathbf{D}_I$，此輸入文件向量再與文件分群圖中之所有神經元比較以找出最近的文件分群 C_I，若 C_I 與任一事件分群（即 C_i）相同，則 D_I 會被視為使用者有興趣之特定事件文件，

（三） 新奇事件偵測：

新奇事件之定義為未能歸屬於任一文件群組之事件，當新進文件 C_I 出現時，它的文件向量會與自我組織圖中之所有神經元比較，即計算文件向量 $\mathbf{C}_I$ 與每一神經元之突觸權重向量 $\mathbf{w}_j$ 之歐氏距離(Euclidean distance)，亦即 $\frac{\|\mathbf{C}_I - \mathbf{w}_j\|}{|N|}$，若與所有神經元之距離皆超過一門檻值，則我們便認為 C_I 為一新奇事件，因其與任一文件群組皆不相似。

伍、實驗結果

為了驗證本文所提方法之效能，我們使用普遍的 Reuters-21578 資料集來進行實驗，此資料集將文件區分為 135 個類別，然而其中部份類別並不包含任何文件，我們使用其中的 Modified Apte Split 方法將其區分為訓練資料與測試資料，其中各包含 9603 與 3299 份文件，為了達到更好的效果，我們捨棄了包含少於 20 份文件的類別，我們也捨棄了字數過少（少於 20 個字）與字數過多（多於 300 個字）的文件，經上述處理後訓練資料與測試資料各包含 4825 與 1768 份文件，我們再將這些文件依第三節所述方法轉換為向量，我們在建立字彙集時會捨棄只出現一次的關鍵字與不是名詞的字，然後我們建立一自我組織圖來分群並標記文件以建立文件分群圖，表一為自我組織圖之統計資料，請注意此表所顯示的為獲得最佳結果之自我組織圖，而後我們針對特定事件與新奇事件偵測進行實驗，其結果分述如下，

表 1 自我組織圖統計資料

參數	值
自我組織圖大小	15×15
神經元突觸數量	3726
學習速率初始值	0.4
最大訓練週期	600

（一） 特定事件偵測實驗結果：

首先我們使用 Reuters-21578 資料集之類別關鍵字作為事件關鍵字，資料庫中超過 20 份文件的類別共有 57 個，因此我們便以這些類別的主題作為事件關鍵字來測試本文的方法，我們先以第四節所述方法找出這些事件關鍵字所屬的文件分群，而後我們從測試資料集中選取某一類別的文件，再將此文件之文件向量與所有神經元相比較，若最相近者即為該類別之主題所屬之文件分群，則我們將其視為一成功的偵測，表二顯示實驗的結果，

表 2 特定事件偵測結果

描述	值／結果

測試文件數量	1768
測試主題（事件）數量	57
成功偵測數量	1354
準確率	76.6%
最佳單一類別準確率	96.3%
最差單一類別準確率	53.2%

（二） 新奇事件偵測實驗結果：

要評估新奇事件偵測之效能較為困難，主要是因為我們必須另於訓練文件外準備一組新奇文件．之前所使用的測試文件集並不能滿足新奇性的要求．一個簡單的策略為假設屬於不同類別的文件即屬不相關．根據這樣的假設，我們將訓練文件集依其類別切割．原始的資料集共包含 6593 份訓練與測試文件．我們將其依類別重新切割為訓練與測試文件集．訓練文件集包含 57 個類別中的 50 類別，共 5871 份文件．測試文件集則包含剩餘的 7 個類別中的文件，共 722 份文件．我們使用新的訓練文件集依表 1 的參數重新訓練自我組織圖．而後使用測試文件集中的文件來與自我組織圖中的神經元進行比較以辨識其是否為新奇文件．理想情況下所有測試文件集中的文件應該都是新奇的，然而因文件間存在部份關聯性，會影響其結果．表 3 為使用二個不同門檻值之實驗結果．實驗結果並無法到達高度的準確性，主要是因為測試文件和訓練文件間存在著部份關聯性．若能使用更精確分割的文件集，當可獲得較佳的結果．

表 3 新奇事件偵測結果

描述	值／結果	
測試文件數量	722	
門檻值	0.3	0.4
成功偵測數量	346	263
準確率	47.9%	36.4%
最佳單一類別準確率	65.2%	58.7%
最差單一類別準確率	21.7%	20.2%

陸、結論

本文中我們提出了一個新的方法來自公開來源文件中發掘商業情報．我們首先將文件進行分群並找出文件間之關聯．兩種新的偵測方法，即特定事件偵測與新奇事件偵測，被用來偵測有用的情報．初步實驗的結果驗證了本文所提方法之可行性．

參考文獻

1. Johnson, R. “Analytic Culture in the US Intelligence Community: An Ethnographic Study”. *Center for the Study of Intelligence, Central Intelligence Agency*. https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/full_title_page.htm. Retrieved 2010-12-9.
2. NATO. *NATO Open Source Intelligence Handbook*, Supreme Allied Commander Atlantic, Norfolk, VA, 2001.
3. NATO. *NATO Open Source Intelligence Reader*, Supreme Allied Commander Atlantic, Norfolk, VA, 2002.
4. NATO. Intelligence Exploitation of the Internet, Supreme Allied Commander Atlantic, Norfolk, VA, 2002.
5. Baldini, N., Neri, F. and Pettoni, M. “A Multilanguage Platform for Open Source Intelligence,” WIT Transactions on Information and Communication Technologies, Vol 38, 2007, pp. 325-334.
6. Neri, F. and Priamo, A. “SPYWatch, Overcoming Linguistic Barriers in Information Management,” LNCS, Vol. 5376, Intelligence and Security Informatics, 2008, pp. 51-60.
7. Neri, F. and Geraci, P. “Mining Textual Data to Boost Information Access in OSINT,” Proceedings of the 13th International Conference on Information Visualization, Vol. IV, 2009, pp. 427-432.
8. Pfeiffer, M., Avila, M., Backfried, G., Pfannerer, N., and Riedler, J. “Next Generation Data Fusion Open Source Intelligence (OSINT) System Based on MPEG-7,” *Proceedings of the International Conference on Technologies on Homeland Security*, Waltham, MA, 2008, pp. 41-46.
9. Vincen, D., Stampouli, D., and Powell, G. “Foundations for System Implementation for a Centralised. Intelligence Fusion Framework for Emergency Services,” *Proceedings of the 12th International Conference on Information Fusion*, Seattle, WA, 2009, pp. 1401-1408.
10. Badia, A., Ravishankar, J., and Muezzinoglu, T. “Text Extraction of Spatial and Temporal Information,” Proceedings of the 2007 International Conference on Intelligence and Security Informatics, 2007, pp. 381.
11. Palmer, J. “Textually Retrieved Event Analysis Toolset,” Military Communications Conference, Vol. 3, 2005, pp.1679-1685.
12. Jiang, J. and Conrath, D. W. “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” Proceedings of International Conference on Research in Computational Linguistics (ROCLING X), Taiwan, 1997.
13. Atkinson, M., Belayeva, J., Zavarella, V., Piskorski, J., Huttunen, S., Vihavainen, A., and Yangarber, R. “News Mining for Border Security Intelligence,” Proceedings of IEEE International Conference on Intelligence and Security Informatics, Vancouver, Canada, 2010, pp. 173.
14. Tanev, H., Piskorski, J., and Atkinson, M. “Real-Time News Event Extraction for Global Crisis Monitoring,” Proceedings of the 13th International Conference on Applications of Natural Language to Information Systems (NLDB 2008), Lecture Notes in Computer Science Vol. 5039, 2008, pp. 207-218.
15. Piskorski, J., Tanev, H., Atkinson, M., and Van der Goot, E. “Cluster-Centric Approach to News Event Extraction,” Proceedings of the International Conference on Multimedia & Network Information Systems, Wroclaw, Poland, IOS Press, 2009.

16. Grishman,R., Huttunen, S., and Yangarber, R. “Information Extraction for Enhanced Access to Disease Outbreak Reports,” Journal of Biomedical Informatics, Vol. 35, No. 4, 2003, pp. 236-246.
17. Yangarber, R., Jokipii, L., Rauramo, A., and Huttunen, S. “Extracting Information about Outbreaks of Infectious Epidemics,” Proceedings of the HLT-EMNLP 2005, Vancouver, Canada, 2005, pp. 22-23.
18. Wiil, U. K., Memon, N., and Karampelas, P. “Measuring Link Importance in Terrorist Networks,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp.225-232.
19. Wiil, U. K., Memon, N., and Karampelas, P. “Detecting New Trends in Terrorist Networks,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp.435-440.
20. Bartik, V. “Text-Based Web Page Classification with Use of Visual Information,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp. 416-420.
21. Dawoud, K., Alhajj, R., and Rokne, J. “A Global Measure for Estimating the Degree of Organization of Terrorist Networks,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp.421-427.
22. Liu, H. K. and Sandfort, J. “A Case Study of Open Source and Public Participation in Catalyzing Social Innovations,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp.428-431.
23. Neri, F., Geraci, P., and Camillo, F. “Monitor the Web Sentiment, the Italian Prime Minister’s Case,” Proceedings of 2010 International Conference on Advances in Social Networks Analysis and Mining, Odense, Denmark, 2010, pp.432-434.

Mining Business Intelligence Using Open Source Documents

Hsin-Chang Yang
Department of Information Management, National University of Kaohsiung
yanghc@nuk.edu.tw

Yi-Hsiang Huang
Institute of Information Management, National University of Kaohsiung
dream788@hotmail.com

Abstract

Intelligence collection and analysis always play a major role in a company's growth. Traditional intelligence management process was most concealed and required massive human effort. It also has the disadvantages of rarity and danger. Therefore open source intelligence (OSINT) emerged as a major intelligence collection and analysis approach. Differing from traditional approach, the sources of OSINT are publicly accessible and have the properties of openness and massiveness which may result in disadvantages of inconsistency and lack of validation. For now, most of the OSINT processing is conducted manually which requires massive human effort and time cost. Automatic processing of OSINT is then unavoidable for modern applications. Although there exists software services to aid such automatic processing, the functionality and degree of automation are still immature and limited. In this work we developed an automatic processing approach for OSINT based on proposed text mining techniques. This approach may automatically identify interesting events from various aspects from which business could benefit. The major contribution of this work is that we have developed high-order mining techniques for OSINT, which will benefit domains like national security, personal knowledge management, with emphasis on business growth.

Keywords: Text Mining, Business Intelligence, Open Source Intelligence, Self-Organizing Map