DOI: 10.6245/JLIS.2017.432/734

網路不當內容過濾之研究調查

李龍豪
國立臺灣師範大學圖書資訊學研究所博士後研究員
E-mail: lhlee@ntnu.edu.tw

關鍵詞：對抗式資訊檢索；兒童保護；網際網路審查；資訊過濾；網路內容分級

【摘要】

隨著網際網路的蓬勃發展，網路成為最大的資訊來源，越來越多的不當內容，例如：色情、賭博、暴力、毒品、種族歧視、恐怖的、猥褻的或攻擊性的內容等等，可以輕易被存取。基於兒童保護的考量，網路內容分級已經成為重要的議題。是否採用網際網路過濾軟體屬於館藏決策，圖書館有權可以決定提供哪些資訊給讀者閱覽，理解過濾軟體的技術與其影響會是值得注意的決策參考。因此，本文從網路探勘的觀點出發，調查近二十年來的網路不當內容過濾研究，歸納出發展趨勢，建議未來可能的研究方向。最終，期許本文可以成為網路內容過濾軟體建置與技術研發的參考資料。

緒論

由於網際網路的普及，資訊的傳播非常迅速，網路上的資訊也逐漸趨於多元且良莠不齊，越來越多的不當內容（objectionable content），例如：色情（pornography）、賭博（gambling）、暴力（violence）、毒品（drug）、種族歧視（racism）、恐怖的（horror）、猥褻的（obscene）或攻擊性的（offensive）內容等散佈於網際網路中，其中，網路色情問題最為普遍且受到重視（Gilliat-Smith, 2001; Döring, 2009; White, 2001）。網路內容的存取方式，最常見的是透過搜尋引擎輸入關鍵字，然後從搜尋結果藉由超連結（hyperlinks）連結至網頁。Beitzel、Jensen、Chowdhury、Grossman 與 Frieder（2004; 2007）分析搜尋引擎的查詢日誌（query logs），調查用來作為資訊存取的查詢種類，發現頻率最常使用的前三名依次為色情、娛樂、和音樂。Jansen、Spink 與 Tefko（2000）研究發現約莫四分之一的查詢項目與性議題相關。Chau、Fang 與 Yang（2007）分析一個香港的中文搜尋引擎日誌，結果發現最常被使用的查詢前 100 名（包含 54 個英文查詢和 46 個中文查詢），超過一半以上都跟性慾相關。在缺乏完善的網路內容過濾機制之下，使用者可以輕易存取網站內容，網路內容管理已經成為刻不容緩的議題。

全球資訊網協會（World Wide Web Consortium, W3C）最早提出 PICS（Platform for Internet

Content Selection）的機制（Resnick & Miller, 1996），制定標籤（labels）作為網頁內容的詮釋資料（Metadata），由網頁提供者對內容先加上合適的關聯標籤，再行出版網頁內容。Lee、Hui 與 Fong（2002）實證發現只有 11%的網頁內容使用 PICS 標籤。網路內容分級協會（Internet Content Rating Association, ICRA）基於 PICS 的機制，提出一套內容描述的系統，包含裸露、性、褻瀆不敬的語言（profanity）與暴力等種類，讓數位內容創建者，可以對產生的內容自我規範（self-regulation）標記對應的種類（Machill, Hart, & Kaltenhäuser, 2002）。然而，Lee 與 Luh（2008）調查 1 萬個中英文雙語的色情網站，發現只有 5.4%的網站使用 ICRA 標籤。由此可知，內容自我審查機制幾乎不可行，ICRA 也於 2010 年 10 月，因為無法獲得廣泛的接受而宣告終止。

近年來，網際網路審查（Internet Censorship）對於網站及其承載的內容做過濾逐漸受到關注（Leberknight, Chiang, & Wong, 2012a, 2012b; Rosenberg, 2001）。網路作為一種大眾傳媒和新興的資訊媒介，不需要傳統的出版發行編審，就可以直接傳播給受眾，但內容可能包含與傳統道德觀念和現行法律相違背的部分。因此，網路內容的審查種類，包含涉及國家安全、侵權資料、違法網站、不當網站等。為了防範青少年接觸網路色情資訊，以及保護兒童心身健康的考量，各國制定了相關法律限制網路使用，但引發了兒童保護與言論自由的之間的討論與爭議（Akdeniz, 2010; Weitzner, 2007）。美國公共圖書館的經營，主要依賴政府經費挹注，故需配合政府政策。美國於西元 2000 年通過「兒童網際網路保護法案」（Children’s Internet Protection Act; CIPA），要求公共圖書館必須安裝網路過濾軟體，防止兒童接觸色情資訊等不當內容，否則不能獲得聯邦的預算補助或優惠，引發是否違反憲法第一修正案的言論自由權利的法律訴訟。後來，美國聯邦最高法院最終認為：圖書館是否使用網路過濾軟體屬於館藏決策（collection decision）範疇，與限制言論自由無關（ALA Intellectual Freedom Committee, 2000; Bell, 2000; Laughlin 2002; 張郁蔚，2002; 黃國正、黃玫溱，2004）。既然圖書館使用過濾軟體限制網路存取的內容，沒有違背法律規範且依法有據，如何選擇建置合適的網路過濾軟體，以及瞭解其背後過濾技術的發展與影響，對於圖書館決定哪些資訊將被納入館藏提供給讀者，將會是密切相關的重要議題。

綜合以上所述，本文將專注於網路不當內容過濾的相關技術，調查近二十年來的學術研究成果，整理出技術的發展趨勢和未來挑戰。最終，期許本文可以成為過濾軟體的技術研發與系統建置的參考資料。

文獻探討

網路探勘（Web Mining）是一門應用資料探勘的技術，從網頁資料中擷取出知識的研究領域（Etzioni, 1996）。這個研究領域大致上可以分成網路內容探勘（web content mining）、網路結構探勘（web structure mining）與網路使用行為探勘（web usage mining）三個子領域。

其中，網路內容探勘針對網頁的內容，包含文字、影像、視訊等等進行資料探勘；網路結構探勘探討個別網頁的文件結構，或者是網頁之間的結構組成關係；網路使用行為探勘則是從使用記錄檔中挖掘使用者的行為關係。

以下文獻探討將從網路探勘的角度切入，分別說明不同的網路內容過濾方法，並將技術優缺點整理如表 1。

表 1　**網路不當內容過濾技術的比較整理**

方法領域	分析標的	優點	缺點
網路內容探勘（web content mining）	1. 文字（text） 2. 影像（image） 3. 視訊（video） 4. 標題（title） 5. 網址（URL）	1. 內容蘊含資訊量豐富，分類效果顯著。 2. 純文字分析的資訊量小且複雜度較低。 3. 影像分析不需要語言知識，都是型樣識別問題。	1. 文字分析需具備該語言的專業背景知識。 2. 影像處理比文字分析花費更多的系統資源與運算時間。 3. 皮膚比率較高的圖片（如：證件照和泳裝照），容易造成誤判。
網路結構探勘（web structure mining）	1. 超連結（hyperlinks） 2. 詮釋資料（Metadata） 3. 圖片說明文字（image tooltip）	1. 連結分析可以提升網頁分類效果 2. 透過連結找到更多的潛在目標	1. 通常要搭配內容分析，無法單獨使用。 2. 容易有誤差傳播現象，導致誤判接連發生。
網路使用行為探勘（web usage mining）	1. 查詢日誌（query logs） 2. 點擊資料（click-through data）	1. 當行為明顯且具有一定特性時，可以獲得較高成效。 2. 相對於內容與結構分析，耗費較少計算資源。	1. 行為不明顯或是改變迅速時，無法有效判斷。 2. 行為資料不容易取得，且須事先徵求使用者同意。

網路內容探勘

根據內容分析標的不同，各自說明如下：

一、文字分析

最常採用的技術是純文字分析，因為文字資訊量小，且為網頁內容提供者最容易使用的表現形式。直覺的做法是關鍵字比對，過濾軟體依照事先定義好的關鍵字清單，只要內容提及預設的關鍵字，則該網頁會被過濾軟體封鎖，導致無法存取內容。這種方式看似簡單且字串比對較為容易，但因為沒有文字語意上的判斷，經常造成種類誤判的問題。例如：性（sex）這個字非常高頻率的出現在色情網頁，所以被列為過濾的關鍵字，但卻連帶導致有關性教育的網站也被封鎖。相似地，胸部（breast）這個關鍵字，會造成有關乳癌的內容被阻擋。

Caulkins、Ding、Duncan、Krishnan 與 Nyberg（2006）提出統計式分類過濾方法（Filtering by Statistical Classification, FSC），分析網頁的文字內容，將文字特徵歸類成多個統計自變數，

利用多變數分析的概念，計算預測值作為是否過濾一個網頁具有色情內容的依據。

Lee 與 Luh（2007; 2008）分析網頁的文字內容，先求出個別字詞的色情傾向（porn tendency），然後透過卡方檢定為基礎的統計方法，計算整個網頁內容的色情指標值（indicator value），根據這個數值將網頁分成色情、未確定與非色情三個類別。Lee、Luh 與 Yang（2008）基於這個方法，提出一個早期決策啟發原則（early decision making heuristics），針對線上即時網頁分類過濾時，在時間和計算資源有限下，探討如何選擇部分網頁內容做分析，實驗結果顯示分析網頁的標題（title）和前 10%的網頁內容，考量成本效益權衡之下，可以達到最佳結果。

除了分析網頁內容，連結到網頁的網址（Uniform Resource Locator, URL），也被用來當作判斷是否需要過濾的依據（Ding, Chi, Deng, & Dong, 1999）。如果一個網址存在於一個建置好的黑名單（blacklist）中，則該網址的內容請求會被拒絕；相反地，如果不在黑名單的網址列表中，則網頁內容的存取將被允許。

Baykan、Henzinger、Marian 與 Weber（2009）提出只用網址來判斷網頁種類的做法，將網址視為由字母構成的字串序列，然後依照非字母和標點符號切割成多個記號（token），這些記號構成的特徵向量，藉由支援向量機（Support Vector Machine, SVM）訓練得到的分類器，可以只從網址判斷是否屬於成人（adult）網頁類別。

Ma（2008）將網址視為字串，用類神經網路（neural network）中字串編碼器（string encoder）的方法，區分一個網址是否導向色情網頁，只對網址做分類比分析內文本身來得快速，計算複雜度較低，需要較少計算資源，特別適合線上的即時過濾。

Zhang、Qin 與 Yang（2006）提出網址為基礎的過濾方式，先將網址表示成由多個連續詞組成的詞組（n-gram）序列，然後應用機器學習的演算法訓練一個分類器。實驗結果顯示，網址為基礎的方式在辨識色情網頁上，效果還不錯，進一步與內容為基礎的方式整合之後，可以達到更好的過濾成效。

二、影像分析

在判斷某個網頁是否為色情時，影像分析的做法是先對單一圖片利用影像處理技術判斷是否為色情圖片，然後根據網頁中色情圖片的數量比例來判別是否為色情類別。

最常見的方式是基於皮膚顏色模型（Hammami, Tsishkou, & Chen, 2004; Kelly, Donnellan, & Molloy, 2008; Lee, Kuo, Chung, & Chen, 2007; Mofaddel, & Sadek, 2010; Wu, Zuo, Hu, Zhu, & Li, 2008; Zhu, Wu, Cheng, & Wu, 2004; Zuo, Hu, & Wu, 2010），找出圖片中的可能與人體皮膚區塊有關的部分，描繪出皮膚的統計特徵，再經由大量圖片的機器學習訓練，找到色情圖片的樣式做形樣識別（pattern recognition）。這類型的分析方法，假設如果一張圖片有過高比例的皮膚區塊，表示可能過於裸露，有較高的可能性是色情圖片，但這個假設容易對證件照或

是泳裝照造成誤判。除了皮膚顏色之外，形狀特徵（shape features）也被整合來偵測和過濾色情圖片（Drimbarean, Corcoran, Cuic, & Buzuloiu, 2001）。

Deselaers、Pimenidis 與 Hey（2008）基於影像的紋理特徵（texture feature），提出視覺詞袋模型（Bag-of-Visual-Words Model），將影像表示成沒有順序性的特徵組合，然後運用機器學習演算法訓練分類器，用來辨識一張影像是否屬於色情圖片。

Arentz 與 Olstad（2004）則不同於一般色情圖片建構特徵向量的方式，對於任何一張影像給定某個機率值，用來表示該張圖片包含色情內容的可能性，再透過基因演算法來演化計算，藉此建構色情特徵向量，達到色情圖片辨識的目的。

Lienhart 與 Hauke（2009）採用機率式潛在語意分析（Probabilistic Latent Semantic Analysis, PLSA）的標題模型（topic model），應用在色情圖片分類上，可以達到不錯的辨識效果，而且誤判率的情形也很少。

Li、Xiong、Wu、Hu、Maybank 與 Yan（2015）提出一個情境感知的多實例學習（context-aware multi-instance learning）演算法，先用隨機漫步（random walk）的方式擷取影像的鄰近區域特徵，然後是模糊支援向量機（Fuzzy Support Vector Machine, FSVM）最佳化分類器，最後應用在恐怖影像辨識。

三、視訊分析

視訊（video）是由一連串的影像構成的動態串流，偵測一部視訊是否含有不當內容，通常是將每一張影像圖框（frame）抽取出來，對個別影像做類別判斷，如果超過一定比例的影像都屬於特定類別，則該視訊則歸屬於某個分類。

Kim、Kwon、Kim 與 Choi（2008）將視訊切割成多個圖框，然後基於圖片的皮膚顏色、紋理與形狀特徵，對某一個圖框做是否屬於色情圖片的識別，然後彙整作為整個視訊是否為色情影片的依據。Lee、Shim 與 Kim（2009）提出一個階層式系統架構用來偵測色情影片，第一階段事先運用雜湊值（hash）做早期預測，第二階段根據單一圖框做即時偵測，第三階段整合所有圖框做事後整體的預測。Akbulut、Patlar、Bayrak、Mendi、Hanna（2012）擷取每一個圖框的膚色特徵，構成整個視訊的觀察序列，然後運用隱馬可夫模型（Hidden Markov Model, HMM）做視訊分類，用以判斷與過濾色情影片。

Jansohn、Ulges 與 Breuel（2009）提出動作資訊（motion information）結合靜態影像特徵的視訊分類作法，運用在色情影片偵測。Xing、Liang、Cheng、Dang 與 Huang（2011）提出 SafeVchat 系統架構，針對網路隨機視訊服務，建立基於動作為基礎的皮膚偵測模型，用來辨識視訊是否含有猥褻的內容的不當行為。

Wang、Li、Hu 與 Wu（2011）提出基於視訊、聲音（audio）與情感（emotion）特徵的

多實例學習（Multiple-instance Learning）方法，用來偵測電影中是否存在恐怖場景。

Li、Huo、Jin 與 Xu（2016）對 MediaEval 2015 暴力影片資料集，進一步標記子類別，包含血腥（blood）、槍（gun）、強迫（force）、死亡（death）、武器（weapon）、繩索（rope）、搏鬥（fight）、襲擊（hit）、捆綁（bind）與瞄準（aim）共 10 個小類，然後利用這些更細緻的資訊來改進暴力偵測的效果。

四、綜合討論

網頁內容包含文字、影像、視訊等多種不同載體的數位內容，除了從單一媒體預測不當內容的類別可能性，也可以採取兩種以上媒體做綜合判斷，目的是希望充分利用網頁的內容來輔助分類，避免只有純文字或是影音多媒體時，單一種內容分析無法辨識的限制。

內容探勘方法的優點是資訊量豐富，如果能夠有效利用並擷取特徵，通常可以達到不錯的分類效果，缺點則是計算複雜度較高，需要較多的計算資源，即時內容分析需要耗費較多時間沒效率。

網路結構探勘

結構探勘分析標的是個別網頁的文件結構，或是超連結結構的組成關係，依照連結的不同又可分為網頁文件內連結（intra-document hyperlinks）或是網頁間連結（inter-document hyperlinks）。在不當內容網頁分類與過濾的研究上，結構探勘的技術通常與內容分析搭配，企圖增進內容分析的成效。

Ho 與 Watters（2004; 2005）調查發現色情網頁比一般網頁，含有較多的連結和圖片內容，因此，結合內容和結構特徵發展一個預測模型，利用關聯法則（Association Rules）和貝氏分類（Bayesian Classification）的方法，辨識和過濾色情網頁。

Chau 與 Chen（2008）將網頁表示成內容與連結為基礎的特徵集合，網頁內容部分有標題的關鍵字數目，以及內容中關鍵字的 TF-IDF（Term Frequency－Inverse Document Frequency）數值的加總，連結特徵部份包括網頁的對外連結（outgoing links）、指向該網頁的內連結（incoming inks）、和指向同一個網站的兄弟連結（sibling links）這三種連結類型的同樣兩個量化數值。有了特徵數值集合後，可以套用不同的機器學習方法，判斷網頁的對應類別。

Lee、Hui 與 Fong（2002; 2003; 2005）從訓練網頁的文字內容中，挑選常用的色情關鍵字，除了網頁內容本身，還包含網址、網頁詮釋資料中關鍵詞（keywords）和描述（description）部分，以及圖片的說明文字（image toolkit），建構網頁的特徵向量，然後透過兩種類神經網路，分別是自組織對映（Self-Organizing Maps）和模糊自適應共振理論（Fuzzy Adaptive Resonance Theory），計算關鍵字的權重，最後用來對測試網頁做分群，用以判別網頁是否屬於色情類別。

Agarwal、Liu 與 Zhang（2006）從不同資料來源擷取網頁特徵，包括網址、連結文字（anchor text）和指向的連結、圖片和說明文字、標題、詮釋資料、內容正文等，套用支援向量機的機器學習法，比較不同來源特徵的效能差異，實驗結果指出結合不同來源有助於提升網頁分類效能。

Hammami、Chahir 與 Chen（2003; 2006）提出 Web Guard 系統架構，結合文字內容分析、關鍵字字典與影像皮膚顏色分析，同時探索網頁的結構，包含已知在黑名單中的連結數量、連結文字包含禁用詞的數量、網頁詮釋資料中關鍵詞部分，和網址字串中有色情關鍵字的數量等建構出特徵向量，然後利用資料探勘技術，偵測與過濾色情網頁。

Lee 與 Luh（2008）觀察到色情網頁有較高的比例會互相連結，物以類聚具有群聚現象。有鑑於此，定義具有集線器（hub）特徵的色情網頁，實驗結果顯示藉由追縱這類型的網頁，更容易找到其他存在的色情網頁，可以用來有效更新現有的色情黑名單。

Guermazi、Hammami 與 Hamadou（2007; 2008）針對含有暴力內容的網頁，提出基於關鍵字字典的分類方法，網頁特徵擷取自、網址、標題、網頁內容文字、對外連結和影像數量等，運用並且評估不同機器學習方法，包含支援向量機、類神經網路與決策樹（Decision Trees），實驗結果反映整合所有方法可以達到最佳的過濾效果。

一般來說，網路結構探勘的優點是在內容分析效能還不錯的前提之下，可以藉由連結分析進一步提升網頁分類的效果，並且找到更多的潛在目標。缺點則是連結分析通常無法單獨使用，而且如果搭配的內容特徵不佳的情形下，則容易有誤差傳播現象（error propagation），導致誤判接連發生。

網路使用行為探勘

行為分析探索使用者的行為特徵，為了獲得相似的資訊，存取網路內容的行為通常相似且具有一定特性，使用行為探勘著重於挖掘大量使用者的行為紀錄檔，希望在沒有網路內容分析的前提之下，只分析行為紀錄就能判別存取網頁的類別。

Szummer 與 Craswell（2009）將使用者在搜尋引擎的查詢詞（queries）和點擊的搜尋結果，建構出使用者點擊圖（click graph），圖上的每一個節點代表查詢詞或者是搜尋結果，每一條邊表示該搜尋結果被某個查詢點擊過，然後在這個點擊圖上做相似度分群，最後在影像搜尋的使用者行為紀錄上驗證成效，結果顯示可以有效找出被使用者點擊過的色情圖片。

使用者存取網路內容最常見的方式是透過搜尋引擎的協助。根據不同的資訊需求，轉化成不同的查詢詞彙，再從搜尋引擎回傳排序過的搜尋結果中，挑選點擊滿足需求或是較接近的網址，然後探索被點擊網址指向的網頁內容，這種資訊尋求的過程，直觀上將網頁內容與查詢詞彙做關聯，某種程度上隱含間接對網頁內容做類別標註（tagging）的可能性。

Lee 與 Chen（2011b; 2012）分析 MSN 搜尋引擎的查詢日誌檔，先找到可能具有色情搜尋意圖的查詢詞，然後對被這些查詢詞點擊的網址，做多數決投票（majority voting），通過票數門檻值（threshold）的候選網址，最後再做網址的詞組檢查。當使用者意圖和點擊行為相似，藉由探索群體智慧，可以答案過濾網路色情的效果。類似地，Lee 與 Chen（2011a）改從被點擊的網址出發，將網址表示成查詢詞構成的集合，然後透過卡方檢定為基礎的統計方法（Lee & Luh, 2008），判斷一個網址是否指向色情網頁，並將這些網址彙整在一個黑名單中，供網路色情過濾使用。

除了透過搜尋引擎存取網路內容，使用者還可以選擇在瀏覽器的網址列直接鍵入網址，或者選擇一個過去存取過的網頁書籤（bookmark），或是藉由某個網頁內容中的連結連到外部其他網頁等等可能的方式，這些使用者網路瀏覽行為產生的點擊資料（click-through data），蘊含重要的行為資訊，但沒有包含在搜尋引擎查詢紀錄檔中。

Lee、Juan、Chen 與 Tseng（2013）探索趨勢科技（Trend Micro）提供的點擊資料，分析一萬兩千多位匿名使用者一個月的上網瀏覽產生的點擊序列，抽取每個點擊的代表性特徵，透過條件隨機域（Conditional Random Field, CRF）機器學習法，預測網址指向的內容屬於何種類別，總共有 83 個定義好的網頁種類，其中，包含 9 個不當內容的潛在類別，分別是墮胎（abortion）、酒精或菸草（alcohol/tobacco）、非法事務（illegal affairs）、毒品（drugs）、賭博（gambling）、大麻（marijuana）、色情（pornography）、暴力或種族歧視（violence/racism）、和武器（weapons）。Lee、Juan、Tseng、Chen 與 Tseng（2015）分析更大量的趨勢點擊資料，包括一千一百五十幾萬全球用戶的網路瀏覽行為，提出一個整合隱馬可夫模型和頂級域（Top-Level Domain, TLD）條件機率的聚合模型（aggregation model），對於 83 種定義類別，做類別可能性預測的機率排序，實驗結果指出前兩名的預測類別中，就會包含網頁正確類別。

整體來說，網路使用行為探勘優點是：當使用者行為特徵明顯時，容易構成群體智慧，可以獲得較高的類別判斷精確率，且相對於分析網頁內容和結構，行為分析通常耗費較少的計算資源。此外，相對於分析網頁內容和結構，行為分析通常耗費較少的計算資源。但伴隨的方法缺點是：當使用者行為不明顯或是改變迅速時，無法有效判斷。另外，使用行為資料非常難取得，且必須事先徵求使用者同意。

發展趨勢

從時間的推移觀察研究發展歷程，歸納整理如下：

從單一到多元

網路不當內容分類與過濾這個研究議題，近二十年的發展有從單一類別到多元類別的發展趨勢。早期被關注的種類是色情類別，主要原因是網路色情的日益氾濫，伴隨的兒童保護

的意識逐漸高漲，另一方面，色情的定義相對其他不當內容類別明確，標記資料做監督式機器學習（Supervised learning）較為容易。近年來，資訊社會的發展迅速，網路上的數位內容也趨於多元，開始出現許多暴力、種族歧視、恐怖的或猥褻的內容等等，過濾軟體的發展也因此朝多元不當類別發展。

另外，從技術面來看，也是從最簡單的單一技術，到複雜的多重技術整合。研究議題的濫觴，通常都是最直覺的方式，例如：文字分析的關鍵字比對和影像分析的皮膚顏色模型，等發現到方法的不足之處之後，才開始導入其他特徵擷取方式，然後整合更多的先進技術，多元的模型整合往往可以增進系統效能，但同時需要更多的計算資源耗費。

從靜態到動態

另外一個研究發展趨勢是從靜態內容到動態資訊。早期的內容分析標的不是純文字就是單一圖片構成的靜態內容，然而，隨著網路世代的高速發展，有越來越大的網路頻寬（bandwidth）和越來越小的網路延遲（latency），動態的視訊可以直接在網路上串流播放，如何擷取動態的內容特徵，如何偵測到不當的場景或是不洽當的行為，近幾年來開始受到關注。

此外，過去這個研究議題都是將網頁視為靜態的資料集，然後在資料集上做網頁分類，但是不當內容的本質變化快速的，靜態的分類作法經常是治標不治本，研究實驗的當下成效顯著，但隨著內容變化往往每況愈下。近期開始有研究關注如何從已知的不當內容黑名單，更新黑名單中網址和找到更多不當內容網站，企圖在高度動態的網路環境中，達到令人滿意的過濾成效。

未來挑戰

未來的研究挑戰，建議可從以下兩方面思考：

意圖驅動的行為內容整合

不當內容過濾議題本質上屬於對抗式資訊檢索（Adversarial Information Retrieval）。從內容提供者的角度，希望不當內容可以能見度越高，越容易被存取越好。但是，從內容過濾軟體的觀點，卻是希望可以阻擋不當內容被看見。兩者之間存在敵對立場，彼此相互對抗的關係，內容提供者迅速改變網頁的內容呈現方式、網址、網站的結構等等，希望可以避免被過濾技術辨識到，而過濾軟體則想要趕上變化，盡可能找到越多不當內容越好，尤其是那些經常被高度存取的不當內容。

回歸問題本質上來看，因為使用者有內容存取的意圖（intent），所以不當內容的存在才有意義，如果沒有這類型的需求，即使不當網站存在於網路上，沒有資訊存取的話，內容屬於何種類別其實無關緊要。所以，從使用者意圖出發，對於使用者點擊的網址及其內容，如

何整合行為和內容分析的技術，追縱動態的變化，達到不當內容辨識和過濾效果，將會是一個值得研究的發展方向。

新興平台的內容過濾

網路科技蓬勃發展，現今有許多新興的資訊傳遞平台，例如：Facebook 社群網站、Line 即時通訊軟體等，在這些平台上可以透過加入朋友關係，形成個人的社交網路（social network），屬於特定對象的封閉資訊傳播系統，在這些平台上藉由群體訊息散佈不當內容，或是網路直播方式從事不當行為，技術上如何達到內容識別與過濾，將會是一個值得挑戰的難題。

結語

圖書館使用過濾軟體限制網路存取的內容，隸屬於館藏決策範疇，圖書館可以決定哪些資訊將被納入於館藏中供讀者存取，如何選擇合適的過濾軟體，將決定涵蓋的網路資訊類別，這部分與圖書館的資訊取得習習相關。

本文從網路探勘的觀點切入，調查網路不當內容過濾技術的相關研究，整理過濾技術的的優缺點，歸納出研究發展趨勢，根據具挑戰性的議題，建議未來可能的研究方向。最終，期許可以成為網路內容過濾軟體建置與技術研發的參考資料。

參考文獻

Agarwal, N., Liu, H., & Zhang, J. (2006). Blocking objectionable web content by leveraging multiple information sources. *ACM SIGKDD Explorations Newsletter*, *8*(1), 17-26.

Akbulut, A., Patlar, F., Bayrak, C., Mendi, E., & Hanna, J. (2012). Agent based pornography filtering system. *In Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications* (pp. 1-5). Hoboken, NJ: IEEE Press.

Akdeniz, Y. (2010). To block or not to block: European approaches to content regulation, and implications for freedom of expression. *Computer Law and Security Review*, *26*(3), 260-272.

ALA Intellectual Freedom Committee (2000). Statement on library use of filtering software. Retrieved from http://www.ala.org/Template.cfm?Section=IF_Resolutions&Template=/ContentManagement/ContentDisplay.cfm&ContentID=13090.

Arentz, W. A. & Olstad, B. (2004). Classifying offensive sites based on image content. *Computer Vision and Image Understanding*, *94*(1-3), 295-310.

Baykan, E., Henzinger, M., Marian, L., & Weber, I. (2009). Purely URL-based topic classification. *In Proceedings of the 18th International World Wide Web Conference*, 1109-1110. New York: ACM Press.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. *Journal of the American Society for Information Science and Technology*, *58*(2), 166-178.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 321-328. New York: ACM Press.

Bell, B. W. (2000). Filth, filtering, and the first amendment: ruminations on public libraries' use of Internet filtering software. *Federal Communications Law Journal*, *53*(2), 191-238.

Caulkins, J. P., Ding, W., Duncan, G., Krishnan, R., & Nyberg, E. (2006). A method for managing access to web pages: Filtering by Statistical Classification (FSC) applied to text. *Decision Support Systems*, *42*(1), 144-161.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, *44*(2), 482-494.

Chau, M., Fang, X., & Yang, C. C. (2007). Web searching in Chinese: A study of a search engine in Hong Kong. *Journal of the American Society for Information Science and Technology*, *58*(7), 1044-1054.

Deselaers, T., Pimenidis, L., & Hey, H. (2008). Bag-of-visual-words models for adult image classification and filtering. *In Proceedings of the 19th International Conference on Pattern Recognition*, 1-4. Hoboken, NJ: IEEE Press.

Ding, C., Chi, C.-H., Deng, J., & Dong, C.-L. (1999). Centralized content-based web filtering and blocking: How far can it go? *In Proceedings of the 1999 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 115-119). Hoboken, NJ: IEEE Press.

Döring, N. M. (2009). The Internet's impact on sexuality: A critical review of 15 years of research. *Computers in Human Behavior*, *25*(5), 1089-1101.

Drimbarean, A. F., Corcoran, P. M., Cuic, M., & Buzuloiu, V. (2001). Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition. *In Proceedings of the 2001 International Conference on Consumer Electronics*, 278-279. Hoboken, NJ: IEEE Press.

Etzioni, O. (1996). The world-wide web: Quagmire or gold mine? *Communications of the ACM*, *39*(11), 65-68

Gilliat-Smith, M. (2001). It's time to take porn seriously. *Network Security*, *2001*(8), 20.

Guermazi, R., Hammami, M., & Hamadou, A. B. (2007). Using a semi-automatic keyword dictionary for improving violent web site filtering. *In Proceedings of the 2007 International Conference on Signal Image Technologies and Internet Based Systems*, 337-344. Hoboken, NJ: IEEE Press.

Guermazi, R., Hammami, M., & Hamadou, A. B. (2008). WebAngels filter: A violent web filtering engine using textual and structural content-based analysis. *Lecture Notes in Artificial Intelligence*, *5077*, 268-282.

Hammami, M., Chahir, Y., & Chen, L. (2003). WebGuard: Web based adult content detection and filtering system. *In Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, 574-578. Hoboken, NJ: IEEE Press.

Hammami, M., Chahir, Y., & Chen, L. (2006). WebGuard: A web filtering engine combining textual, structural, and visual content-based analysis. *IEEE Transactions on Knowledge and Data Engineering*, *18*(2), 272-284.

Hammami, M., Tsishkou, D., & Chen, L. (2004). Adult content web filtering and face detection using data-mining based skin-color model. *In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo*, 403-406. Hoboken, NJ: IEEE Press.

Ho, W. A., & Watters, P. A. (2004). Statistical and structural approaches to filtering Internet pornography. *In Proceedings of the 2004 IEEE International Conference on Systems, Man, and Cybernetics*, 4792-4798. Hoboken, NJ: IEEE Press.

Ho, W. A., & Watters, P. A. (2005). Identifying and blocking pornographic content. *In Proceedings of the 21st International Conference on Data Engineering*. Hoboken, NJ: IEEE Press.

Jansen, B. J., Spink, A., & Tefko, S. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management*, *36*(2), 207-227.

Jansohn, C., Ulges, A., & Breuel, T. M. (2009). Detecting pornographic video content by combining image features with motion information. *In Proceedings of the 17th ACM International Conference on Multimedia*, 601-604. New York: ACM Press.

Kelly, W., Donnellan, A., & Molloy, D. (2008). Screening for objectionable images: A review of skin detection techniques. *In Proceedings of the 2008 International Machine Vision and Image Processing Conference*, 151-158. Hoboken, NJ: IEEE Press.

Kim, C.-Y., Kwon, O.-J., Kim, W.-G., & Choi, S.-R. (2008). Automatic system for filtering obscene video. *In Proceedings of the 10th International Conference on Advanced Communication Technology*, 1435-1438. Hoboken, NJ: IEEE Press.

Laughlin, G. K. (2002). Sex, lies, and library cards: The first amendment implications of the use of software filters to control access to Internet pornography in public libraries. *Drake Law Review*, 51, 213-282.

Leberknight, C. S., Chiang, M., & Wong, F. M. F. (2012a). A taxonomy of censors and anti-censors: part I – impacts of Internet censorship. *International Journal of E-Politics*, *3*(2), 52-64.

Leberknight, C. S., Chiang, M., & Wong, F. M. F. (2012b). A taxonomy of censors and anti-censors: part II – anti-censorship technologies. *International Journal of E-Politics*, *3*(4), 20-35.

Lee, L.-H., & Chen, H.-H. (2011a) Collaborative blacklist generation via searches-and-clicks. *In Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 2153-2156. New York: ACM Press.

Lee, L.-H., & Chen, H.-H. (2011b). Collaborative cyberporn filtering with collective intelligence. *In Proceedings of the 34th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 1153-1154. New York: ACM Press.

Lee, L.-H., & Chen, H.-H. (2012). Mining search intents for collaborative cyberporn filtering. *Journal of the American Society for Information Science and Technology*, *63*(2), 366-376.

Lee, L.-H., Juan, Y.-C., Chen, H.-H., & Tseng, Y.-H. (2013). Objectionable content filtering by click-through data. *In Proceedings of the 22nd ACM Conference on Information and Knowledge Management*, 1581-1584. New York: ACM Press.

Lee, L.-H., Juan, Y.-C., Tseng, W.-L., Chen, H.-H., & Tseng, Y.-H. (2015). Mining browsing behaviors for objectionable content filtering. *Journal of the Association for Information Science and Technology*, *66*(5), 930-942.

Lee, J.-S., Kuo, Y.-M., Chung, P.-C., & Chen, E-L. (2007). Naked image detection based on adaptive and extensible skin color model. *Pattern Recognition*, *40*(8), 2261-2270.

Lee, L.-H. & Luh, C.-J. (2007). Classifying pornographic web pages using a chi-square based statistics method. *Journal of Information Management*, *14*(2), 225-246

Lee, L.-H., & Luh, C.-J. (2008). Generation of pornographic blacklist and its incremental update using an inverse chi-square based method. *Information Processing and Management*, *44*(5), 1698-1706.

Lee, L.-H., Luh, C.-J., & Yang, C.-J. (2008). A study on early decision making in objectionable web content classification. *In Proceedings of the 6th IEEE International Conference on Intelligence and Security Informatics*, 35-39. Hoboken, NJ: IEEE Press.

Lee, P. Y., Hui, S. C., & Fong, A. C. M. (2002). Neural networks for web content filtering. *IEEE Intelligent Systems*, *17*(5), 48-57.

Lee, P. Y., Hui, S. C., & Fong, A. C. M. (2003). A structural and content-based analysis for web filtering. *Internet Research*, *13*(1), 27-37.

Lee, P. Y., Hui, S. C., & Fong, A. C. M. (2005). An intelligent categorization engine for bilingual web content filtering. *IEEE Transactions on Multimedia*, *7*(6), 1183-1190.

Lee, S., Shim, W., & Kim, S. (2009). Hierarchical system for objectionable video detection. *IEEE Transactions on Consumer Electronics*, *55*(2), 677-684.

Li, B., Xiong, W., Wu, O., Hu, W., Maybank, S., & Yan, S. (2015). Horror image recognition based on Context-aware multi-instance learning. *IEEE Transactions on Image Processing*, *24*(12), 5193-5205.

Li, X., Huo, Y., Jin, Q., & Xu, j. (2016). Detecting violence in video using subclasses. *In Proceedings of the 2016 ACM Multimedia Conference*, 586-590. New York: ACM Press.

Lienhart, R., & Hauke, R. (2009). Filtering adult image content with topic models. *In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, 1472-1475. Hoboken, NJ: IEEE Press.

Ma, H. (2008). Fast blocking of undesirable web pages on client PC by discriminating URL using neural networks. *Expert Systems with Applications*, *34*(2), 1533-1540.

Machill, M., Hart, T., & Kaltenhäuser, B. (2002). Structural development of Internet self-regulation: Case study of the Internet Content Rating Association (ICRA). Info, *4*(5), 39-55.

Mofaddel, M. A., & Sadek, S. (2010). Adult image content filtering: a statistical method based on multi-color skin modeling. *In Proceedings of the 2010 IEEE International Symposium on Signal Processing and Information Technology*, 366-370. Hoboken, NJ: IEEE Press.

Resnick, P., & Miller, J. (1996). PICS: Internet access controls without censorship. *Communications of the ACM*, 39(10), 87-93.

Rosenberg, R. S. (2001). Controlling access to the Internet: The role of filtering. *Ethics and Information Technology*, 3, 35-54.

Szummer, M., & Craswell, N. (2008). Behavioral classification on the click graph. *In Proceedings of the 17th International World Wide Web Conference*, 1241-1242. New York: ACM Press.

Wang, J., Li, B., Hu, W., & Wu, O. (2011). Horror video scene recognition via multiple-instance learning. *In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1325-1328. Hoboken, NJ: IEEE Press.

Weitzner, D. J. (2007). Free speech and child protection on the web. *IEEE Internet Computing*, *11*(3), 86-89.

White, B. (2001). Fighting the porn war: The rise of email pornography in the workplace. *Network Security*, 2001(11), 16-17.

Wu, O., Zuo, H., Hu, W., Zhu, M., & Li, S. (2008). Recognizing and filtering web images based on people's existence. *In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*

and Intelligent Agent Technology, 648-654. Hoboken, NJ: IEEE Press.

Xing, X., Liang, Y.-L., Cheng, H., Dang, J., Huang, S., Han, R., Liu, X., Lv, Q., & Mishra, S. (2011). SafeVchat: Detecting obscene content and misbehaving users in online video chat services. *In Proceedings of the 20th International Conference on World Wide Web*, 685-694. New York: ACM Press.

Zhang, J., Qin, J., & Yang, Q. (2006). The role of URLs in objectionable web content categorization. *In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 277-283. Hoboken, NJ: IEEE Press.

Zheng, Q.-F., Zeng, W., Wen, G., & Wang, W.-Q. (2004). Shape-based adult image detection. *In Proceedings of the 3rd International Conference on Images and Graphics*. Hoboken, NJ: IEEE Press.

Zhu, Q., Wu, C.-T., Cheng, K.-T., & Wu, Y.-L. (2004). An adaptive skin model and its application to objectionable image filtering. *In Proceedings of the 12th Annual ACM International Conference on Multimedia*, 56-63. New York: ACM Press.

Zuo, H., Hu, W., & Wu, Q. (2010). Patch-based skin color detection and its application to pornography image filtering. *In Proceedings of the 19th International Conference on World Wide Web*, 1227-1228. New York: ACM Press.

張郁蔚（2002）。美國公共圖書館與網路資訊過濾技術使用之初探。**國立中央圖書館臺灣分館館刊，8**(2)，36-50。

【Chang, Yu-Wei (2002). Měiguógōnggòngtútúshūguǎn yǔ wǎnglu zīxùn guòlǜ jìshù shǐyòn zhī chūtàn. In Proceedings of the 8-2th Journal of National Taiwan Library, 8(2), 36-50.】

黃國正、黃玫溱（2004）。公共圖書館網路使用政策的探討。**圖書與資訊學刊，49**，68-78。

【Huang, Guo-Jeng, & Huang, Mei-Chen (2004). On the Internet Use Policy in Public Libraries. In Proceedings of the 49th Journal of Library and Information Studies, 49, 68-78. 】

DOI: 10.6245/JLIS.2017.432/734

A Survey of Research on Web Objectionable Content Filtering

Lung-Hao Lee
Postdoctoral Fellow, Graduate Institute of Library and Information Studies,
National Taiwan Normal University, Taiwan (R.O.C.)
E-mail: lhlee@ntnu.edu.tw

Keywords: Adversarial Information Retrieval; Child Protection; Internet Censorship; Information Filtering; Web Content Rating

【Abstract】

With the proliferation of the Internet, the Web has become the largest accessible repository of information. Increasingly more objectionable web content, such as pornography, gambling, violence, drug, racism, horror, obscene or offensive content, and so on can be accessed easily. Based on child protection considerations, web content rating has been regarded as an important issue. In libraries, the use of Internet filtering software depends on collection decision. Librarians have the rights to provide selected information for their readers. To understand techniques behind a filtering software and its influences will be noticeable concerns for making the decision. Hence, from web mining perspectives, this paper investigates research studies on objectionable web content filtering in recent decades, induces the development trends, and suggests the possible research directions in the future. The hope is such survey can be a supplementary reference for development and implementation of objectionable web content filtering software.

【Long Abstract】

Introduction

The web has become the largest accessible repository of information with the proliferation of the Internet. Objectionable web content, such as pornography, gambling, violence, and drugs, may cause disapproval, or protest during users' web surfing. The PICS (Platform for Internet Content Selection) specification was proposed by W3C (World Wide Web Consortium) to request content providers to associate their content with appropriate labels. Among several self-rating mechanisms that have been

built on PICS, the labeling system promoted by the ICRA (Internet Content Rating Association) is a popular one. Nevertheless, empirical evaluations have found that only about 10% of web content used these labels. Thus, the self-rating labeling mechanism still has a long way to go because objectionable content providers rarely fulfill this requirement to avoid being filtered easily.

Controlling access to the Internet through filtering has been increasingly recognized as a mandatory response to address the issue of inappropriate materials. Filtering spreading, objectionable web content also has attracted intensive attention to protecting children or anyone else from inappropriate materials. Users usually issue queries to search engines and click search results to meet their various information needs. Many searchers, especially children without adult supervision, however, do not expect to have sexually explicit content included in their results. Google's SafeSearch filter provides searchers with the ability to configure the browser settings to prevent objectionable content from appearing in search results. Nevertheless, objectionable-content providers can improve their content visibility to avoid being filtered, so keeping the filters as up to date as possible is a challenging issue with rapidly changing web.

From technical perspectives, objectionable web content filtering is a research topic of adversarial information retrieval. The content providers do their best to make their objectionable web content to be accessible as easy as possible. The filtering software focuses on detecting and identifying objectionable content categories automatically for blocking information accesses. Adversarial characteristics exist between each other. How to defeat the opposite side and achieve individual access purposes are main research objectives.

In libraries, the use of Internet filtering software depends on collection decision. Librarians have the rights to provide selected information for their readers. To understand techniques behind a filtering software and its influences will be noticeable concerns for making the decision.

Hence, in this paper, we investigate research studies on objectionable web content filtering from web mining perspectives to understand this important issue. The hope is such survey can be a supplementary reference for development and implementation of objectionable web content filtering software.

Research Survey

Web mining is the application of data mining techniques to discover data patterns from the World Wide Web. Web mining can be divided into three types: (1) Web Content Mining: the mining, extraction, and integration of useful data, information, and knowledge from the web page content. (2) Web Structure Mining: to use graph theories to analyze the node and connection structures of a web site. (3) Web Usage Mining: to discover interesting usage patterns from web data for understanding the needs of information accesses.

From the viewpoint of web content mining, traditional filtering techniques regard objectionable access identification as a categorization problem through content analysis, that is, crawling the content of URLs and analyzing the texts, images, video clips, and films by machine learning to distinguish normal web pages from objectionable ones. The keyword matching and filtering method intuitionally rejects any requests for web pages whose total number of objectionable keywords exceeds a predefined threshold. The major problem of this method is over-blocking, which results in blocking accesses to normal web content. Intelligent content analysis approaches attempt to understand the context in which discriminative features appear and make a classification decision. Overall, the strength of intelligent content analysis lies in its superior accuracy, but comparatively more training time and human intervention are needed for the claimed performance.

From the viewpoint of web structure mining, the hyperlinks, metadata, and images tooltips of objectionable content are included to enhance the categorization performance of intelligent content analysis. Web structure mining is usually cooperated with content mining rather than being adopted solely. Through structure mining, more potential candidates of objectionable content can be founded, but the weakness is error propagation problem when an incorrect classification is caused by content analysis.

From the viewpoint of web usage mining, users' search behaviors have been explored and found that the top-1 category in terms of popularity in information accesses is pornography. In addition to retrieving pornographic web content with search engines, users may access other kinds of objectionable web content relating to gambling, violence, and drugs. In practices, users have many alternatives to access web content, such as inputting the requested URLs into browsers directly, selecting specific bookmarks and accessing out-links within web pages. Besides searches and clicks, as manifested in sever-side query logs, click-through data that keep users' browsing behaviors on the client side may provide more information during real-life web surfing. Usage mining approaches have the advantage of consuming relatively fewer computational resources comparing with content and structure mining methods. However, the usage patterns are difficult to obtain due to insufficient users' intents and the limitations of behavioral data.

Research Trends and Challenges

We summarize the research trends of investigated studies following with the timeline. Two trends are induced: (1) from Single to Diverse: the main objectionable category is pornography in terms of popularity in information access. Recently, multiple objectionable categories consist of gambling, violence, and so on, are regarded as identified and filtered targets. Moreover, in addition to the single technique being used in this research topic, hybrid method integrated diverse techniques are used to

enhance the filtering performance. (2) from Static to Dynamic: in the earlier studies, traditional filtering techniques regard this research problem as categorization with static data. However, how to follow the changing trails of objectionable websites and maintain steady filtering performance is a challenging issue. In recent studies, behavioral mining approach is adopted to explore usage intents for keeping up with the dynamical variants of objectionable web content from users' perspectives.

We also suggest two possible research directions in the future. (1) Intent-driven content filtering: If users do not have information needs of accessing such objectionable content, the category of objectionable web content does not matter anymore. Starting from users' intents, how to integrate content and structure mining techniques to analyze users' clicked URLs and their content, and then keep up with dynamic changes of objectionable content to maintain steady performance that is a worthy of study to reflect adversarial natures. (2) Objectionable content filtering on the social media platforms: in recent years, social media platforms (e.g., Facebook and Line) are important information communication channels. Users are grouped into different social networks through friend relationships. In individual networks, users may distribute objectionable content or instant messages on such closed groups. How to identify and filter objectionable content in social media platforms will be a challenging research issue.

Conclusions

Automatically identifying the widespread objectionable web content, such as pornography, gambling, violence and drugs, is an important task for protecting children or anyone else from inappropriate materials during their web surfing. This paper investigates research studies on objectionable web content filtering in recent decades, induces the development trends, and suggests the possible research directions in the future. In libraries, how to use the Internet filtering software is a collection decision issue. We hope this survey can be a supplementary reference for development and implementation of objectionable web content filtering software.

【Romanization of Chinese references is offered in the paper.】