基於網路拓撲的聚焦爬蟲研究
Research on Focused Crawler Based upon Network Topology

熊菲 Fei Xiong, 劉雲 Yun Liu, 李勇 Yong Li
北京交通大學通信與資訊系統北京市重點實驗室
xf1985@eyou.com

摘要

聚焦爬蟲面向主題，過濾無關鏈結，只抓取相關的網頁資訊。通用的聚焦爬蟲，無法處理冗餘鏈結，因此本文提出了一種基於網路拓撲的聚焦爬蟲。從搜索引擎獲取初始網頁集，使用向量空間模型計算文本相似性。對抽取出的URL先進行鏈結分析，再根據無標度網路特徵，修正URL的權值。同時回饋不相關的主題區域，並通過URL與種子集合的距離設置不相關URL的緩衝區長度。仿真結果表明基於網路拓撲的爬蟲比通用爬蟲具有更高的查準率。

關鍵字：聚焦爬蟲、鏈結分析、無標度網路、向量空間。

Abstract

Subject-oriented focued crawler, skips irrelevant links, and receives only relevant information. However, general fouced crawler couldn't deal with redundant links. This paper presents a kind of focused crawler based upon network topology. The crawler gets original URL sets from search engine, then calculates content similarity by the model of vector space. It analyzes link structure of websites, moreover modifies weight of URL according to the characteristic of scale-free network. Relevance feedback is used to disengage irrelevant regions, and the length of buffer is set for irrelevant URL by the distance between URL and seed pages. Experiments results prove that the precision of this focused crawler is higher than general crawler.

Keywords: Focused Crawler, Link Analysis, Scale-free Network, Vector Space.

1 緒論

隨著資訊網路化的日益普及，互聯網上的資訊與日俱增，巨大的潛在價值蘊含在這些海量異構的Web資訊資源中。互聯網方便快捷的資訊發佈方式以及受眾互動的交流平臺，使得網路已經超越傳統媒體，成為即時資訊獲取的主要方式。新聞事件通常最早出現在互聯網上，並在網路中引起討論。

如何有效地提取並利用網路資訊成為一個巨大的挑戰。搜索引擎通過查詢的方式為用戶提供快捷有效的資訊獲取途徑。網路爬蟲為搜索引擎從萬維網上下載網頁，是搜索引擎的重要組成部分[1]。通用的寬度優先搜索（BFS）爬蟲，在抓取過程中，完成當前層次的搜索後，才進行下一層次的搜索。該類爬蟲覆蓋面廣，加權僅根據鏈結出入度，與內容無關，往往包含用戶不關心的資訊。基於特定話題聚焦爬蟲很好地解決了這個問題。聚焦爬蟲根據既定的抓取目標，採用網頁分析演算法，有選擇的訪問相關鏈結，過濾與主題無關的鏈結，獲取所需要的資訊[2]。

主題網路存在著隧道現象，從一相關網頁到另一相關網頁的路徑中可能包含不相關的區域。文獻[3]在寬度優先搜索的基礎上，使用了相關回饋並合理設置隧道長度，使聚焦爬蟲儘早脫離不相關區域，挖掘隱藏不相關鏈結後的相關內容。文獻[4]結合鏈結結構及文本相似性處理URL，將URL權值定義為網頁相似性與URL鏈入鏈出度之和的乘積，是後效性的URL處理方式。文獻[5]在使用向量空間模型的基礎上，通過對鏈結進行物理結構及邏輯結構分析過濾URL，以略微降低查全率為代價，追求更高的查準率。

互聯網呈現出無標度網路的特徵，網站或網頁的入度出度服從冪率分佈，度高的節點吸附力強，形成「富者越富」的現象[6][7]。網路的特性對聚焦爬蟲的研究具有重要意義。

通用的聚集爬蟲，將同一網頁中提取的URL不加區分地對待，保留了較多不相關的鏈結，且無法降低相似性判決錯誤帶來的影響。針對上述問題，本文提出的聚集爬蟲根據互聯網拓撲處理URL，對URL進行鏈結分析、優先吸附加權及相似性回饋。最後驗證了該聚焦爬蟲與通用爬蟲相比在主題採集方面的優越性。

2 種子網頁的獲取

根據聚焦關鍵字，訪問各搜索引擎，獲取前m條記錄，作為聚焦的初始鏈結。抓取初始鏈結的原始檔案，得到種子網頁集合$D=(D_1,D_2,D_3...D_m)$。對集合中的每篇網頁D_i，提取主題資訊進行分詞，去掉無意義的助詞、副詞和停用詞，表示成文檔向量形式。

在文檔向量模型中，每篇文檔以一個向量表示，向量的每一維分量對應文檔的一詞條。若文檔D_i包含的詞條為$(t_1,t_2,t_3...t_n)$，則對應的n維文檔向量為$(w_{i1},w_{i2},w_{i3}...w_{in})$，其中$w_{ij}$為詞條$j$在文檔$D_i$中出現的資訊，即詞條$j$的權重。$w_{ij}$採用經典的$TF\times IDF$定義。$IDF$根據文檔總數增量更新。

種子網頁集合被映射成了文檔向量集$W=(W_1,W_2,W_3...W_m)$。從種子網頁集中解析出的URL，賦予初始權值1，加入聚焦爬蟲的搜索佇列，採集時優先搜索較高權值的鏈結。

3 相關度計算

文檔間的相似性使用文檔向量夾角的餘弦來度量。餘弦相似性表徵了文檔在投影方向上分量的相對大小。

新抓取到的網頁，經預處理及分詞後，轉化成詞條向量，計算新網頁與種子網頁集合的相似性。種子網頁集合D，新網頁V，新網頁與種子網頁集的相似性為該網頁與網頁集合所有網頁相似性的平均值$sim\langle V,D\rangle=\frac{1}{m}\sum_{k=1}^{m}sim\langle V,D_k\rangle$。相似性較高的網頁，在向量空間中的夾角越小，傾向於描述同一話題，反之，相似度越低的網頁，屬於不同話題的概率越大。若網頁與種子網頁集合的相似度高於門限值，則把該網頁加入種子集合。

4 互聯網拓撲分析

通用爬蟲將互聯網作為獨立網頁的集合，忽略網路的鏈結結構和拓撲關係，不同URL被相同的處理。然而互聯網的組織結構中蘊含了豐富的資訊。從局部看，網頁之間的鏈結揭示了父網頁與子網頁的關係，它們的目錄層次體現了內容的相關性。宏觀上，互聯網表現出無標度網路的特徵，網頁鏈結的度服從冪率分佈。因此，分析網頁的鏈結關係，並根據網路組織結構對URL加權，以充分利用網路拓撲資訊。

4.1 鏈結分析

網頁的URL是網頁在網站伺服器的目錄層次。URL中，功能變數名稱或IP「http://([a-z0-9]+\.)*[a-z0-9]+/」後，查詢符「？」前的字串即網頁在伺服器的存放路徑，各級目錄以「/」分隔。網站的目錄安排通常都是有條理的，同一檔夾下的網頁屬於同一分類，一個專題下的主題具有相似的目錄結構，包括標識其時間及內容的字串。

父網頁中解析出的URL，其結構反映了子網頁與父網頁的關係，視以下幾種情況分配不同權值（權參量為t，$0.4<t<0.6$）：

(1)子URL包含父URL，則子網頁處於父網頁的下級目錄中。子網頁的主題是父網頁主題的擴展和延伸，子URL分配較高的鏈結權t。

(2)子URL與父URL具有相似的路徑。子網頁與父網頁目錄深度和檔夾長度相同，新主題是前期或跟蹤報導，給予較高權值t。

(3)父URL包含子URL，子網頁主題是對父主題的匯總和聚焦，也分配一定權值t^2。

(4)背景插圖、廣告等冗餘鏈結，權值較低，$t/10$。

同時，較多與主題相關的URL，鏈結附近一段文本中都包含聚焦關鍵字。因此網頁i對父網頁j的鏈結加權係數為：$link_{ji}=path_{ji}+freq_i$，其中，$path_{ji}$為上述四種不同的URL路徑權值，$freq_i$為歸一化的錨文本關鍵詞頻率。

4.2 權值分配

萬維網屬於小世界網路，兼有無標度網路的特徵，同時擁有大的聚集係數和小的平均路徑長度。網頁或站點作為網路的節點，鏈結作為網路的邊，網路具有冪率的度分佈。無標度網路的成長和優先吸附原則，使得那些包含鏈結數較多的網頁越可能獲得新鏈結。新加入網路的主題鏈結到已存在主題的偏向概率與主題包含的鏈結數成正比。無標度網路有著另一個顯著特徵，少數結點具有較大的度，這些結點是網路的中心，它們傾向於彼此相連，形成「富人俱樂部」。在主題網路中，這些結點通常是門戶網站的專題網頁。

相似性判斷失誤，會將不相關的網頁歸入聚焦話題，僅靠鏈結分析不能有效地過濾掉這些網頁中抽取出的鏈結。但這些錯誤聚焦的網頁，通常包含的有效鏈結數較少。因此使用無標度網路的偏向概率對URL加權，降低了這些判決失誤網頁中抽取出的鏈結的權值，提高了查準率。

在google的pagerank演算法基礎上，兼顧相關性計算，鏈結分析及無標度網路特性的影響，修正後的加權值如下：$score(i)=\sum_{t=1}^{n}link_{ti}\cdot\eta(k_t)\cdot sim(V_t,D)$。其中，$n$為網頁$i$的入度，$sim(V_t,D)$是父網頁與種子集合的相關性，$link_{ti}$是網頁$i$對父網頁的鏈結加權係數，$\eta(k_t)=k_t/\sum_j k_j$為主題網頁的偏向概率，$k_t$為父網頁引用的有效鏈結數。$sim(V_t,D)$是URL權值的重要部分，相似性的閾值決定著子URL的取向。

若詞條總數為m，捕獲網頁總數為n，網頁包含的平均鏈結數為l。則相似性計算的複雜度為$O(mn)$，偏向概率的複雜度為$O(nl)$。由於$l<<m$，可見，使用偏向概率加權並沒有帶來較大的額外計算量。

5 相似性回饋和原子模型

一篇網頁的內容與聚焦話題相關，其解析出的子URL與話題相關的可能性較大，反之，子URL傾向於描述新的話題。相關性較高的網頁，可能引用了大量的推薦鏈結，使得抽取出的子URL並非屬於聚焦話題，但由於其繼承了父網頁的相關性而權值較高。抓取這些子URL，將會得到過多不相關的內容，影響系統性能。因此，抓取由同一父網頁解析出的子URL時，若連續數次均得到話題無關的資訊，則減小所有子URL權值。

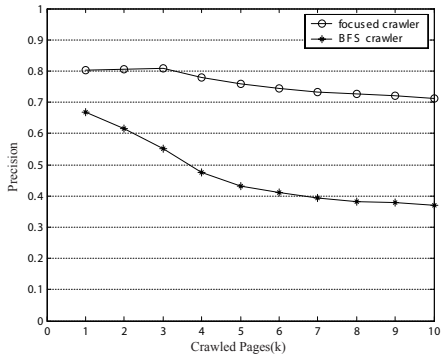
主題網路的隧道效應，使得相關的主題網頁隱藏在無關的鏈結中。與種子集合相似性較低的網頁，在經過多級鏈結後，可能鏈結到相似性高的網頁。忽略相似性低於閾值的網頁中抽取的URL，將會失去隱藏

在這些網頁後的相關鏈結，獲得的主題減少。對不相關網頁設定適當的緩衝，提高聚焦爬蟲的查全率。這裏採用原子模型來分配緩衝的鏈結跳數。種子集合是話題聚焦的依據，作為原子核，原子核不斷更新。從種子集合中提取的URL，作為核外電子。同一URL可能有多個入鏈，與種子集合的最小距離作為其的鏈結深度。處於同一深度的URL視為相同能級。URL的深度越小，越靠近原子核，受到原子核的束縛力越大，通過其獲得相關主題網頁的概率較大，設置較大的緩衝級數。相反，URL深度越大，掙脫原子核束縛逃逸的概率也較大，不易獲得聚焦主題。緩衝跳數的設定為 $step(i) = floor\left(\frac{\sigma}{n(i)}\right)$，其中$step(i)$為網頁$i$的緩衝網頁跳數，$floor$是向下取整，$\sigma$為初始深度參數常量，$n(i)$為種子集合至網頁$i$的鏈結深度。

6 系統仿真

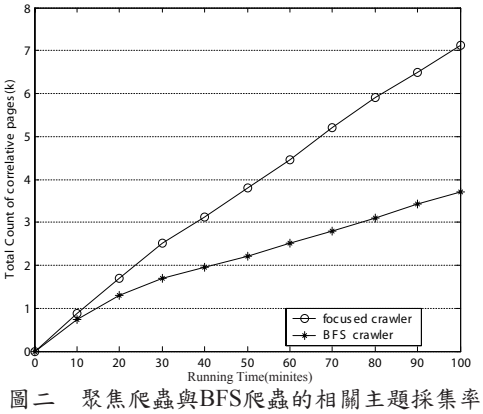
聚焦爬蟲的性能從兩方面來衡量：查全率及查準率。查全率反映了爬蟲的聚焦主題覆蓋面，查準率是獲得相關主題網頁精度的度量。若C為網路的相關主題總數，M為捕獲主題總數，T為所獲取網頁中的相關主題數，則查全率為$recall=T/C$，查準率為$precision=T/M$。由於C通常不易獲得，因此這裏使用查準率作為度量聚焦爬蟲性能的指標。

通過對關鍵字「北京奧運」聚焦，抓取了上百個網站的多份主題網頁，比較聚焦爬蟲與通用爬蟲的相關主題採集效率及不同系統參數對聚焦爬蟲性能的影響（鏈結權參量取0.5）。圖中的結果均為多次仿真數據的平均值，為了保證可比較性，聚焦爬蟲的種子集合與BFS爬蟲的初始佇列一致。



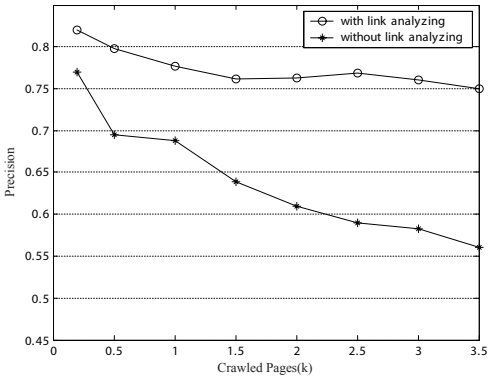
圖一　聚焦爬蟲與BFS爬蟲的查準率

圖一為聚焦爬蟲與BFS爬蟲的查準率隨抓取網頁總數的變化。從搜索引擎獲取聚焦爬蟲的種子網頁集及BFS的初始URL佇列，連續抓取10000多份網頁。圖二為對應的相關網頁採集率。可以看出，本文所提的聚焦爬蟲查準率明顯高於BFS爬蟲，且隨著收集網頁數的增多，BFS相關主題採集緩慢，查準率降低至40%以下，而聚焦爬蟲仍維持在70%以上。



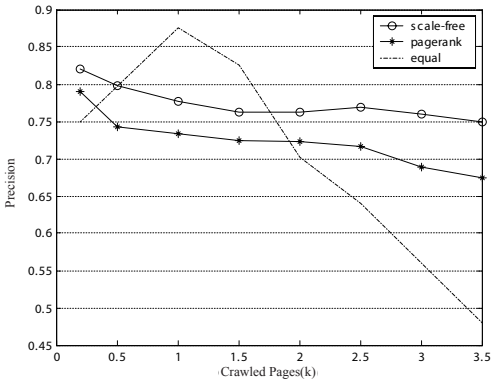
圖二　聚焦爬蟲與BFS爬蟲的相關主題採集率

圖三反映了鏈結分析對聚焦爬蟲查準率的影響。初始種子網頁集為10，抓取3500份網頁。可見，未進行鏈結分析，採集了大量主題網頁中的無效及冗餘鏈結而影響了系統的性能，在抓取2500網頁後查準率已降至0.6以下。



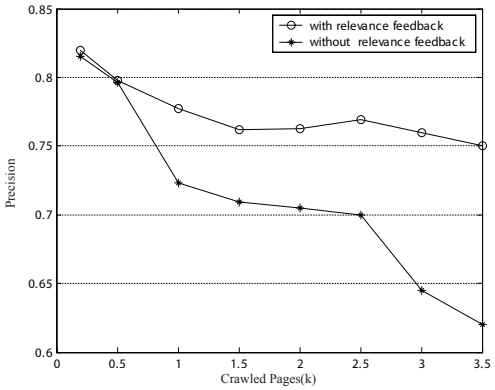
圖三　鏈結分析與未鏈結分析的查準率

圖四為查準率在不同加權方式下隨採集網頁數的變化。基於互聯網拓撲的權值分配方式查準率較pagerank權值方式高，降幅更平緩。而等增益權值分配，在採集初期查準率較高，但隨著系統的運行，片面追求最高的相似性，極易陷入局部最優，查準率迅速降低。



圖四　不同權值下的查準率

圖五是相似性回饋前後查準率的比較。相似性回饋使系統在陷入不相關區域後，能較快地退出抓取誤區，防止性能進一步惡化。



圖五　相似性回饋前後的查準率

下表是聚焦不同話題，選取10個初始鏈結，抓取5000網頁時，聚焦爬蟲與BFS爬蟲的查準率。

表1　不同話題的查準率

	聚焦爬蟲	BFS爬蟲
中國大飛機專案	0.38	0.071
神舟六號	0.65	0.12
汶川地震	0.76	0.28
筆記本電腦	0.73	0.13

7 結論

聚焦爬蟲是面向主題的，只關注某一話題及其演化的網頁內容，捨棄非聚焦的鏈結。本文提出的聚焦爬蟲，基於互聯網拓撲，通過鏈結的目錄層次篩選鏈結，根據優先吸附原則加權URL，使用相似性回饋避免陷入不相關區域，實驗證明該爬蟲的查準率及相關主題採集速率均高於通用聚焦爬蟲。但現今網路更新速度快，聚焦爬蟲的準確度並非長時間所得，而依賴於從搜索引擎獲取的初始鏈結，因此有必要對種子集合進行聚類以提高初始網頁品質。且相似性回饋是後驗的，需要連續監測同一父URL下的主題相關率。下一步的工作，採用啟發式的演算法（如遺傳演算法）更新待抓取列表，避免爬蟲陷入局部最優，迅速脫離不相關區域。

基金項目

教育部高等學校科技創新工程重大專案培育基金項目（707006），通信與資訊系統北京市重點實驗室資助項目，北京市教育委員會共建項目專項資助，北京交通大學科技基金項目（2007XM006）。

參考文獻

[1] J. Cho, *"Crawling the web: Discovery and Maintenance of Large Scaled Web Data,"* Computer Science, 2001.

[2] 周立柱、林玲，*聚焦爬蟲技術研究綜述*，電腦應用，Vol. 25，No. 9，2005，pp.1965-1969。

[3] Jyh-Jong Tsay, Chen-Yang Shih, and Bo-Liang Wu, "*Auto crawler: An Integrated System for Automatic Topical Crawler*," Computer and Information Science, 2005, pp. 462-467.

[4] Jamali M., Sayyadi H., Hariri B. B., et al., "*A Method for Focused Crawling Using Combination of Link Structure and Content Similarity*," Web Intelligence, 2006, pp. 753-756.

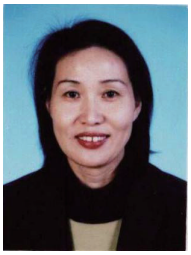
[5] 汪濤、樊孝忠，*鏈結分析對主題爬蟲的改進*，電腦應用，Vol. 24（B12），2004，pp.174-176。

[6] Sen Qin, Guan-Zhong Dai, and Yan-Ling Li, "*Design and Implementation of Web Hot Topic Talk Mining Based on Scale-free Network*," Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, 2006, pp.1184-1189.

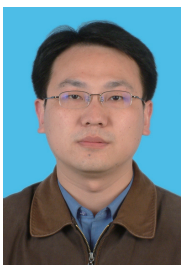
[7] Dorogovtsev S. N., Mendes J. F. F., "*Evolution of Network: From Biological Nets to the Internet and WWW*," London: Oxford University Press, 2003.

作者簡歷

熊菲（**Fei Xiong**），中國北京交通大學電子資訊工程學院在讀博士生，主要研究領域是資訊網路技術、觀點動力學、複雜網路。在北京交通大學獲得通信工程學士學位。



劉雲（**Yun Liu**），中國北京交通大學電子資訊工程學院教授。研究方向為觀點動力學、資訊網路安全、電腦通信、智慧交通系統。在北京交通大學獲得通信與資訊系統博士學位。



李勇（**Yong Li**），中國北京交通大學電子資訊工程學院講師。研究方向為密碼學和資訊安全。在中國科學研究生院資訊安全國家重點實驗室獲得博士學位。