The Application of Web Crawler in City Image Research

網路爬蟲在城市形象研究中之應用

- 準確把握城市形像一直是城市管理者面臨的挑戰。
- 利用網路爬蟲技術抓取特定城市的網頁，對抓取的數據進行處理和分析，得到代表城市的關鍵詞。
- 我們提出了一種分佈式爬蟲架構，並基於該架構收集與城市相關的數據。
- 隨後，成功爬取了幾個示例特定城市(三個 杭州 濟南 青島)。 通過數據預處理、數據分析等方法對爬取的數據進行分析。

一、介紹

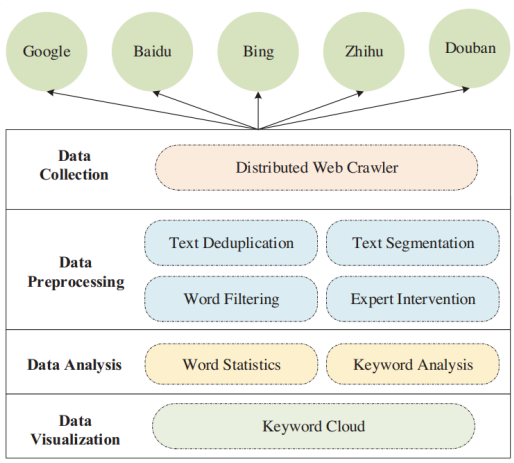
- 隨著互聯網的發展和新媒體的興起，傳統紙媒和電視的影響力正在下降。 然而，隨著網路媒體平台上演講和視頻數量的急劇增加，網路輿論對城市形象的影響越來越大。
- 通過分析與特定城市相關的網路大數據，我們可以獲得代表該城市的關鍵詞，從而幫助城市管理者更好地把握城市形象。
- 網路爬蟲是根據特定的網路搜索策略自動從互聯網上爬取網頁信息的程序。它廣泛用於需要大數據抓取的互聯網搜索引擎和應用程序。
- 目前主流的搜索引擎有Google, Bing, Yahoo, 百度 知乎 豆瓣等。使用搜索引擎，我們可以很容易地收集和檢索給定關鍵字的網頁信息。
- 通過搜索引擎收集特定城市的相關網頁和討論信息是研究城市形象的一種簡單有效的方法。

二、相關工作

- 通過收集公共信息，使用文本分割、詞性標註、句法分析等一系列步驟，可以準確評估目標城市的網路圖像[8]。黃等人。使用文本挖掘技術來調查 2005 年至 2013 年間關於澳門的 TripAdvisor 在線評論內容的演變 [9]。
- 阿里等人。提出了基於模糊本體的情感分析和語義網路規則語言基於規則的決策來監控交通活動，並為旅行者製作城市特徵極性圖[10]。黃等人。提出城市時尚形象的概念，並基於網路新聞大數據對37個城市的時尚形象進行對比分析[11]。
- 在網路爬蟲研究領域，Shrivastava 對網路爬蟲進行了系統的研究，包括網路爬蟲的特性、網路爬蟲使用的策略、網路爬蟲的一般架構等[4]。
- 帕蒂爾等人設計了一個有效的深度網路採集系統，提出了一個兩階段增強的網路爬蟲框架來有效地採集網路界面[12]。
- Fen等人 提出了分佈式爬蟲系統。結果表明，該方法比單機網路爬蟲系統[15]更高效、更穩定。
- 巴赫拉米等人 介紹了一種可擴展的網路爬蟲，該爬蟲採用了在 Windows Azure 雲平台 [14] 中實現的雲計算技術。
- 綜上所述，城市形象的主要研究工作集中在城市圖像檢測、城市圖像傳播、城市圖像構建等方面。一些研究人員已經開始使用數據挖掘、自然語言處理等技術來提取與形象相關的信息用於網路形象檢測的城市。使用網路爬蟲進行城市形象研究的成果很少。

三、模型和架構

- Problem Model
- 系統架構
    - 系統框架由數據採集、數據預處理、數據分析和數據可視化四個模塊組成。



- Data Collection Module
  - 數據收集模塊負責收集網站上的數據。 我們的爬蟲從谷歌搜索引擎、百度搜索引擎、知乎問答社區等收集數據。每個搜索引擎都有自己的特點，網頁結構也各不相同。 因此，需要針對具體的網站設計相應的爬取策略。
  - 網路爬蟲必須解決幾個問題：
    1. 有大量網頁需要抓取；
    2. 每一個爬蟲執行都會分解很多任務；
    3. 目標網站始終具有反爬機制。
  - 因此，單一的網路爬蟲無法取得完全令人滿意的結果。 為了提高爬蟲的效率和準確性，我們提出了一種分佈式爬蟲架構，如圖2所示。

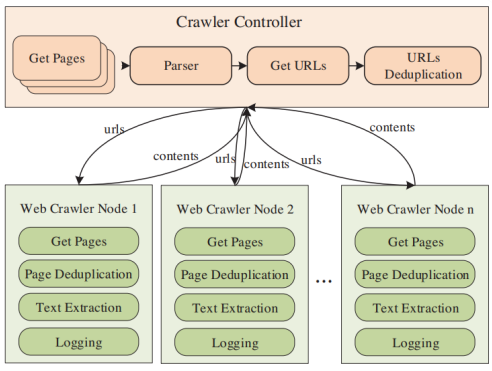


Figure 2. Distributed Web Crawler Architecture.

  - 爬蟲控制器的處理流程如下。
    1. 接收用戶輸入的關鍵字和指定的要查詢的頁數；

2. 啟動多線程，通過谷歌搜索引擎、百度搜索引擎、知乎等關鍵詞搜索相關內容，從Get Pages Module獲取結果；
3. 調用Parser Module進行頁面解析；
4. 通過get URLs模塊和URLs去重模塊獲取目標統一資源定位器（URL）隊列； 最後，控制器將 URL 分發給多台機器作為要執行的任務。

- 爬蟲控制器負責調度所有爬蟲節點，並按順序分配要爬取的鏈接。 同時，它從每台機器接收網頁內容並將這些內容存儲在數據庫中。 假設我們有一個包含 1000 個鏈接等待爬取的列表，並且有 10 個爬蟲節點。 爬蟲控制器的作用是按照某種機制將上千個任務分配給十個爬蟲節點，並接收來自各個爬蟲節點的信息。 每個爬蟲節點的網路爬蟲算法如下。
- 算法1：網路爬蟲算法。
  - 步驟 1. 接收 URL 並將它們存儲在本地 URL 隊列中。
  - 步驟 2. 確定 URL 隊列是否為空。 如果為空，則算法結束，否則轉步驟 3。
  - 步驟 3. 從 URL 隊列中獲取一個 URL，等待幾秒鐘，然後訪問該頁面。
  - step 4. 判斷頁面是否返回成功。 如果返回成功，則執行步驟5。否則，記錄錯誤日誌並執行步驟2。
  - 步驟 5. 使用 SimHash [17] 算法對頁面進行重複數據刪除。 如果網頁不重複，則執行步驟 6，否則執行步驟 2。
  - step 6. 使用線塊分發功能提取文本內容，並將文本內容、來源網址、標題等信息發送給爬蟲控制器，然後執行步驟2。
- Data Preprocessing Module
  - 數據預處理模塊負責對採集到的數據進行預處理。
  - 處理流程包括文本去重、文本切分、分詞、專家干預。 首先，比較存儲在數據庫中的記錄。 如果存在多條文本內容、

來源網站等信息完全相同的記錄，則確定存在重複文本。 我們將刪除多餘的重複記錄，只保留一條可查詢的記錄； 接下來，使用分詞工具將數據庫中的內容分成幾個詞，以附加的形式寫入指定文件； 此外，對分割的結果進行過濾。 在我們的研究中，我們使用 HIT 的 stop list 和 custom list 來去除停用詞和文本中一些無意義的詞，例如“2019”、“city”和“today”。 最後，使用人工干預再次去除無意義的詞，以保證結果的準確性。

- Data Analysis Module
  - 數據分析模塊負責分析預處理後的數據。 可以分為詞統計和關鍵詞分析兩部分。 詞統計部分統計分詞結果文件中所有詞的出現頻率，按照詞頻從高到低將詞寫入新文件。 關鍵詞分析部分應用潛在狄利克雷分配[18]模型將所有關鍵詞按主題分類，並將分類結果寫入新文件。
- Data Visualization Module
  - 數據可視化模塊負責數據可視化。 在我們的研究中，我們使用了一個keyword cloud進行視覺展示，它是通過閱讀詞頻文件形成的。

四、實驗評估

- 實驗設置
  - 為了驗證上述方法的有效性，我們用python語言實現了上述爬蟲程序。 我們在服務器上部署爬蟲控制器，使用 5 台機器作為網路爬蟲節點。 開發環境為 Anaconda，操作系統為 Windows 10。每台機器都有 64GB RAM 和 Intel i5-8400 CPU。我們以杭州、濟南

和青島為例對所提出的方法進行實驗。 為了避開網站的反爬機制，不妨礙目標網站的正常訪問，將爬蟲的等待時間設置為2秒。

- 結果與評價
  - 以Google搜索結果為例，表一展示了三個城市原始統計結果中排名前十的詞和詞頻。 這個結果並沒有去除停用詞和無意義的詞。

TABLE I. ORIGINAL STATISTICAL RESULTS

City / Rank	Hangzhou		Jinan		Qingdao	
	keyword	*count*	*keyword*	*count*	*keyword*	*count*
Top 1	Hangzhou	1205	Jinan	2214	Qingdao	2534
Top 2	China	858	China	1459	China	2157
Top 3	hangzhou	587	Shandong	852	Was	967
Top 4	Hotel	495	2019	623	Tsingtao	846
Top 5	2019	350	jinan	506	2019	761
Top 6	Com	276	10	477	Chinese	728
Top 7	City	275	Was	450	qingdao	694
Top 8	Chinese	273	Hotel	398	10	618
Top 9	More	226	00	374	City	552
Top 10	10	222	city	349	Beer	533

  - 從表一可以看出，在沒有詞過濾的詞頻統計中，前10個詞中有很多無意義的詞，如“杭州”、“10”、“0”。 詞過濾和人工過濾後的結果如表二所示。

TABLE II. FINAL STATISTICAL RESULTS

City / Rank	Hangzhou		Jinan		Qingdao	
	keyword	*count*	*keyword*	*count*	*keyword*	*count*
Top 1	Hotel	495	hotel	398	Beer	533
Top 2	Lake	213	Travel	251	hotel	463
Top 3	School	186	Beijing	238	German	425
Top 4	Travel	157	Class	187	International	372
Top 5	Tea	150	Book	186	Japanese	358
Top 6	Shanghai	140	Shanghai	182	Beijing	222
Top 7	West	134	Scholar	178	August	184

  - 從表二和圖三的結果可以看出，三個城市的關鍵詞存在顯著差異。 與杭州相關的關鍵詞主要是“湖”、“旅遊”、“茶”等。公眾對杭州的印象集中在茶、景點、旅遊； 與濟南相關的關鍵詞主要是“酒店”、“旅遊”等詞。 濟南公眾形像以旅遊、教育、航空為主； 與青島相關的關鍵詞主要有“啤酒”、“酒店”、“國際化”等，公眾對青島的形像以啤酒、旅遊、國際化為主。 這些結果與我們的看法是一致的，濟南的結果也被濟南市經理用於濟南形象的評價中。

五、結論和未來工作

- 在我們的研究中，我們使用網路爬蟲技術來捕獲和分析在線媒體平台中的城市相關信息。 程序的處理流程包括數據採集、數據預處理、數據分析和數據可視化。 從實驗結果來看，我們的方法可以很好地捕捉到城市的相關信息，並通過分析這些信息得到城市關鍵詞云。 這一結果有助於城市管理者掌握城市形象。 我們的研究成果已被濟南採用。 在未來的工作中，我們將構建一個詞源圖來分析每個詞的來源，並考慮一種更直觀的數據可視化形式。