基於新聞評論的熱點話題發現系統研究
Research on the System of Hot Topic Detection Based on News Reports

程軍軍 Jun-Jun Cheng,[1] 劉雲 Yun Liu[2]
[1,2]北京交通大學通信與資訊系統北京市重點實驗室
04271002@bjtu.edu.cn,[1] liuyun@bjtu.edu.cn[2]

摘要

在當前資訊爆炸的時代，如何從特定媒體中獲取有價值的資訊，是人們需要考慮的問題。本文針對新聞報導的特點，提出了一套熱點話題（Hot Topic）發現的演算法，旨在找出當前環境中討論較熱的若干話題，並將該系統實現。本文重點介紹演算法的三大功能模組：預處理（Pre-processing）、聚類（Clustering）和聚類後處理（熱度打分），並且針對資料特性擇優選取相似度公式，對提出的熱度打分公式進行測試，此外還將本熱度打分演算法與其他方法進行比較，最後通過實驗的方法驗證了本演算法是有效的、合理的。

關鍵字：資料採擷（Data Mining）、文本聚類、熱點發現、熱度排序。

Abstract

In the current era of information explosion, it is to be considered that how to dig out valuable information from a specific media. In this paper, we analyze the characteristics of news reports, and propose a set of algorithms about hot topic detection, whose aim is to identify a number of hot topics in the current environment, and finally we achieve this system. In addition, reference is made respectively to some skills used in the modules of the system which contains message pre-processing module, text clustering module and post-clustering module (hot topic detection module). We select a better similarity formula and test the hot topic detection formula which is compared with another method. At the end of the issue, we test the data saved in the database, and finally we prove that algorithms in the system are rational and essential.

Keywords: Data mining, Text Clustering, Hot Topic Detection, Temperature Sorting.

1 緒論

隨著電腦網路的不斷發展，網路資訊已經成為日常生活中的重要組成部分，越來越多的人通過網路來獲取資訊。面對海量資訊，人們很難快速地瞭解當前網路中討論較熱的話題。如何有效準確地發掘網路資料來源，已經成為了眾多學者關注的問題。

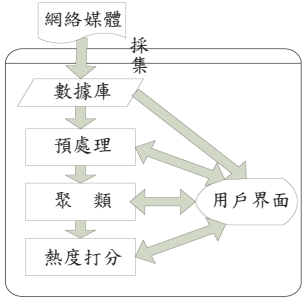
國外比較著名的相關工作主要是話題識別與跟蹤（Topic Detection and Tracking，簡寫為TDT[1]）。TDT識別與跟蹤的物件從特定時間和地點發生的事件擴展為具備更多相關性外延的話題，相應的理論與應用研究也同時從傳統對於事件的識別跨越到包含突發事件及其後續相關報導的話題識別與跟蹤。此外，文獻[9]以熱點詞為基礎，提取由熱點詞構造而成的熱點句，進而對其進行聚類並最終發現熱點話題。

在國內，也有很多學者研究該問題。文獻[3]以綜述的形式介紹了TDT的研究內容以及研究方向；文獻[5][7]介紹了資料採擷方面的相關技術以及遇到的一些問題。在熱點話題方面，文獻[2]給出了適用於BBS環境的話題發現系統，文獻[4]對網路中熱點詞語與熱點話題的關係進行了分析，提出了流量內容中熱點詞語相關度演算法，在此基礎上，提出了熱點話題發現演算法；文獻[6]提出了基於同構的資訊溫度定義，並構造了文本資料庫熱點挖掘等系統，文獻[8]詳細介紹了WEB與情監控和上報系統的架構和功能，並對幾種主要的聚類、分類算法進行了比較。

但是，目前關於熱點發現的演算法都是基於熱點詞與話題的附屬關係，基本原則為出現熱點詞頻率較高的話題即為熱點話題。這樣有可能會出現局部較熱，但是由於話題較分散，因此在類的層次上並不是熱度較高的現象。為了避免該現象的發生，本文從宏觀的角度先對相似話題聚類，然後在類的基礎上進行熱度計算。

本文主要研究網路中新聞報導的熱點話題發現。新聞報導的特點為：篇幅較長，主題思想表述清晰，內容逼近真相；費時、費力，其報導的是當前讀者最關心的問題；同時具有科學性，其威力與魅力建立在紮實細緻的調查研究上。

此處先給出本系統的整體功能圖（圖一），後面的2、3、4部分分別對本系統中的三大功能模組：預處理模組、聚類別模組、熱度打分模組，進行詳細的論述，最後第5部分通過實驗說明本系統的可靠性和準確性。第6部分主要說明下一步的工作。



圖一　熱點話題發現系統功能圖
Fig.1 Function of Hot Topic Detection System

2 預處理

本系統使用網路爬蟲（Parser）等現有的軟體將指定網站的內容存儲到本地硬碟，同時去除HTML檔中特殊的標記，將我們感興趣的文本提取出來，並按照一定的格式存儲到資料庫中。需要提取的內容包括標題、文章內容等。

然後分別讀取每一篇文檔，對其進行分詞處理，去掉停用詞（Stop Words）（自己定義），統計每一個詞的詞頻，應用經典TFIDF方法計算出每個詞的權值，最後將這些資訊分別保存到資料庫。

3 文本聚類介紹

首先，在文檔向量表示部分，我們對標題、正文、文章長度進行表示，在計算詞語權重的時候，使用下面的公式：

$$w_i = tf_i \times \log(N / df_i)$$

該方法為經典的TFIDF方法。其中詞頻用TF（Term Frequency）來表示，IDF為逆文本頻率指數（Inverse Document Frequency）。

在相似度計算方面，使用下面的計算公式：

$$Sim(D,T) = \frac{\sum_{i \in H} t_i d_i}{\sqrt{(\sum_{i \in H} t_i^2) + (\sum_{i \in H} d_i^2) - \sum_{i \in H} t_i d_i}}$$

該公式為廣義Jaccard相似度公式，用於計算每篇文章與其後面文章的相似度。在本文的試驗中，我們選取的相似度閾值為0.405（該值是在數次試驗中聚類效果最佳的值）。

聚類的方法有很多，其適用範圍以及研究的物件都不相同。考慮到新聞報導的特殊性，和動態聚類在空間和時間的大量消耗，本文決定參照TDT[1]專案的single pass演算法，並做一定的修改，以適合現有的資料物件。具體演算法流程如下：

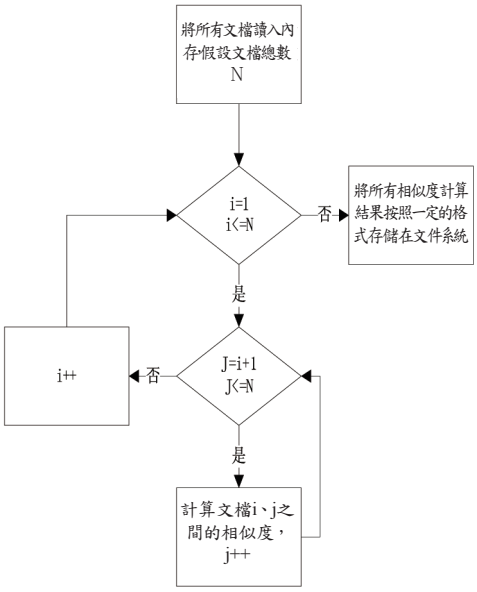
(1)將所有文章（總數為n）讀入系統，可看作一個集合S，在全域範圍內計算每一篇文章與其後所有文章的相似度$Sim(D_i, D_j)$，$i<j\leq n$；

(2)按照預先指定好的順序讀取S中的某一篇文章T_i，如果其不屬於任何已存在的類，則新建一個類C，將S中與T_i相似度大於某一閾值的文章都歸入類C，同時從S中刪除包含於類C內的所有文章；

(3)當S中的所有文章都不屬於任何已存在的類時，本演算法結束，否則轉到第二步。

本演算法與single pass[1]的主要不同處[2]在於single pass是增量聚類（Incremental Clustering），文章在第一次讀入的時候就可以確定所屬的類別，其對文章的先後順序有很大的依賴性，而且該演算法是在局部範圍進行聚類，精度不會太好，同時其採取的相似度公式為傳統的餘弦公式；而本文所採取的演算法將所有文章讀入系統，然後在全域範圍內尋求最優解，同時還可以為後面的功能模組準備中間資料，採用的相似度公式為廣義Jaccard。兩種方法時間複雜度是相同的，但本文的演算法需要有足夠的空間來存儲一些演算法實現過程中必須用到的資料。

該演算法的流程圖如圖二。



圖二　聚類流程圖

Fig.2 Flow Chart of Clustering

4 熱度打分策略

談論一個事件的新聞報導數目越多，就認為該事件越熱；談論一個事件的若干個新聞報導主題越集中，說明該事件越熱；談論一個事件的新聞報導的平均長度越長，在很大程度上，其所談論的事件也就越熱。

因此，在聚類後需要對各個類進行熱度打分，必須考慮上述三個約束條件，參考文獻[2]我們定義幾個參數，類平均長度、類平均相似度和類中文本的數量。

要求出類平均相似度，必須知道類中各個文檔的個體平均相似度。個體平均相似度定義為類中某一文檔與其餘文檔的相似度，取平均值。然後類中所有文檔的個體平均相似度再取一次平均值，就得到了類平均相似度。引入這個概念的目的是減少內部比較雜亂的類的熱度打分，可以說，其散度直接決定了其熱度。

類平均長度為對類中的文檔整體求一次平均值，主要目的就是消除有的文檔過長或過短從而對熱度造成的損傷。

最後，我們給出熱度打分公式：

$$h_i = f_i(TextCount) \times \log(f_i(avg(len)) \times f_i(avg(sim)))$$

其中，$f_i(InforCount)$表示第i類所包含的文本數，$f_i(avg(len))$表示第i類中所有文檔的平均長度，$f_i(avg(sim))$表示第i類中所有文檔的類平均相似度。

最後，我們需要為每個類找一個類標題，具體策略如下：選取類中每篇文檔詞頻大於預先給定值的若干個詞，然後在該類的範圍內計算詞的權值，最後選取權值較大的若干詞作為該類的類名。

5 實驗

系統設計時所使用的作業系統為Windows XP sp2，系統內部存儲為768M，外部硬碟為80G。本系統的開發環境為.net Framework 2.0，開發語言為c#，使用的資料倉庫為Microsoft Access 2007。

本文的資料庫包含395篇關於體育的報導（來自新浪體育，經過一定處理），長度在247到2440之間，內容涉及足球、籃球、網球、羽毛球、圍棋、田徑等多種體育專案，每個報導都存儲為一個文字檔案，然後按照一定的存儲格式將各個文本分別讀入資料庫。

通過簡單的統計可知，國內外大型門戶網站的體育板塊每天的新聞報導數目與本文選取的資料庫數目處於同一量級，因此本文的實驗是具有實際意義的。

5.1 相似度公式的選取

傳統的相似度公式還有餘弦公式，在向量化模型中，該相似度公式應用很廣，於是我們對餘弦公式和廣義Jaccard進行對比，我們選取文檔集中ID為7的文章，與其他文章進行相似度計算，結果如下（取若干結果）：

表1 相似度公式比較結果

Tab.1 Comparison Results between two Algorithms

ID	餘弦公式	廣義Jaccard
8	0.814265004053568	0.681454193239259
9	0	0.813438069264051
10	0.823122164710816	0.699372250542248
11	0	0.716001514296361
12	0.859844625275553	0.696822291670579
13	0	0.350968083466217
14	0	0.813065887294921
15	0.969457516456758	0.935940534556726
16	0.91664399444603	0.83732414791245
17	0.881751648295466	0.775544234551186
18	0	0.458321330541657
19	0	0.201357969847845

通過人工觀察的方法，可以知道文檔7與這些文檔都具有相似性，即敘說的是同一事件，而使用餘弦公式，卻出現了6個0值，相比之下，廣義Jaccard計算結果更小，更具彈性。因此，本系統實現過程中，採用的是廣義Jaccard相似度公式。

5.2 熱點話題發現實驗結果

根據前面介紹的各個功能模組，對395篇文檔進行熱點發現的實現，實驗結果如表2：

表2 熱點發現實驗結果

Tab.2 Experimental Results of Hot Topic Detection

編號	文章數	熱度	主題
0	43	194.856468859818	布萊恩特-巴特-主將-後衛-中國隊
1	22	144.353497021915	炮兵-紅魔-曼聯-小賴特-同維-阿貝
2	45	135.172702435054	馬刺隊-活塞隊-籃球賽-哈基姆-火箭隊
3	47	128.90583132251	進球-紐卡斯爾-英格蘭-皇馬-迪烏夫
4	35	125.045312470973	湖人隊-灌籃-布萊恩特-火箭隊-休士頓
5	20	69.5171090803414	西布羅-錦標賽-南安普敦-切爾西-暴力行為
6	24	64.1002991231312	智利隊-分別在-琳賽-比賽中-琴科-特勒
7	7	60.2166202862239	紐卡斯爾-上籃-芬蘭隊-喬科爾-比賽中-
8	8	51.0765886009867	迪烏夫-西布朗-前鋒-博爾頓-水晶宮-
9	14	36.2007302288331	蝶泳-悉尼奧運會-莫里

從實驗結果可以看出，討論最熱的前幾項都是關於籃球和足球的，但是各自的側重點不同，有談論中國男女籃的，還有談論NBA的，有討論英超曼聯的還有討論西甲皇馬的，等等。可以看出，本系統能夠直觀地將熱點話題展現出來，並提供熱度的值作為參考。

由於我們在進行熱度打分的時候綜合考慮很多因素，因此避免了文章數多就一定「熱」的假像。例如，實驗中有一個文章數達到60的類，其熱度僅僅11左右。這種現象的出現主要是因為該類內部文章的內容過於散亂，討論內容過於廣泛。從上面的表中，也可以得到相應的結論，例如編號為2和3的兩類，2的熱度高於3，而3包含的文章數比2中的文章數多。

本文的主要任務為探索熱點話題發現系統，因此，主要精力集中於得到當前網路中的熱點話題，以及熱度排序，其他實驗資料，本文未給出。

5.3 熱度打分公式比較

參考文獻[6]中對資訊溫度的定義，即$H(t)=(\sum \alpha_i x_i^2)^{1/2}$，其中$\alpha_i$為權重，$x_i$為熱度分析中需要考慮的因數，i取3，即我們同時考慮類中包含的文本數、類平均長度以及類平均相似度。

該演算法的缺點是顯而易見的，由於類平均相似度是小於1的數，而平均長度和文本數都較大，因此在加權平方和的情況下，類平均相似度的影響可以忽略不計，這與實際不符。如果要克服這個缺點，類平均相似度對應的權值必須調得很大，這就增加了可操作的難度。權值變化，熱點發現結果也在變化。當類平均相似度的權值為10^4量級，其他兩個權值都為1，該演算法的結果與本文提出的演算法相似。但是在表2中編號為5的類在此演算法的情況下會具有最高的熱度，而分析資料庫中的文本可知，這是不準確的。

因此，本文所提出的熱度打分較為合理，準確度也較高，得到了較好的效果。

6 結論

本文基於新聞報導的熱點發現系統提出了一套切實可行的熱點發現演算法，它可以用作資訊挖掘領域對比較重要的或者受關注程度較高的熱點資訊進行挖掘。同時，本文對新聞報導的熱點事件發現只是本領域內的一個初探，為將來下一步的工作做好了準備。

但文中的演算法仍有不足：類名都是以關鍵字的形式給出；聚類演算法中閾值是事先給定的；聚類演算法較為單調，針對不同類型的資料，效果可能不同。下一步的工作將重點研究動態自檢驗閾值的選取以及聚類後文本摘要自動生成等技術，以及找出一個兼可靠性與準確性的評測方法用來監測熱點話題發現的準確性和效率。

基金專案

教育部高等學校科技創新工程重大專案培育基金專案（707006），通信與資訊系統北京市重點實驗室資助專案，北京市教育委員會共建專案專項資助。

參考文獻

[1] James Allen, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, "*Topic Detection and Tracking Pilot Study Final Report[J]*," 1998.

[2] 邱立坤、程葳、龍志禕等，*面向BBS的話題挖掘初探[J]*，全國第八屆電腦語言聯合學術會議論文集，2005年，8月，pp.401-407。

[3] 李保利、俞士汶，*話題識別與跟蹤研究[J]*，電腦工程與應用，2003，Vol. 39，No. 17，pp.7-10。

[4] 周亞東、孫欽東、管曉宏等，*流量內容詞語相關度的網路熱點話題提取[J]*，西安交通大學學報，2007，Vol. 41，No. 10，pp.1142-1150。

[5] 王斌，*淺析資料採擷的主要方法和研究方向[J]*，電腦模擬， Vol. 22，No. 10，2005，pp.1-3。

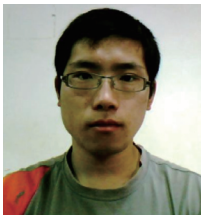
[6] 周啟海、黃濤、張元新等，*同構化資訊溫度與熱點發應用初探[J]*，WEB技術與應用，pp.3-9。

[7] 彭曙蓉、王耀南，*針對小文本的Web資料採擷技術以及應用[J]*，PCL技術應用200例，Vol. 7，No. 3，2006，pp.203-205。

[8] 鄭軍，*網路輿情監控的熱點發現演算法研究[D]*，哈爾濱：哈爾濱工程大學，2006。

[9] Kuan-Yu Chen, Luesak Luesukprasert and Seng-cho T. Chou, "*Hot Topic Extraction Based on Timeline Analysis and Multidimensinal Sentence Modeling[J]*," IEEE Transaction on Knowledge and Data Engineering, Vol. 19, No. 8, 2007, pp.1016-1025.

作者簡歷



程軍軍（Jun-Jun Cheng），中國北京交通大學電子資訊工程學院在讀博士生。研究方向為網路輿論、複雜網路、資料採擷。在北京交通大學獲得理學學士學位。

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。