從新聞和部落格尋找熱門事件

陳建成 盧文祥
國立成功大學資訊工程研究所
p7694427@nckualumni.org.tw, whlu@mail.ncku.edu.tw

摘要

新聞是生活中不可或缺的知識與資訊來源，電子化和網路化使得新聞資訊量的暴增，在如此大量的資訊中，使用者無法迅速搜尋想要獲取的新聞。另外，我們認為傳統的新聞檢索系統僅利用時間性與相關性的技術不能完全滿足使用者需求，因此我們想要提供使用者更有效的搜尋機制，以幫助使用者更方便找到所想要的資訊。我們觀察到部落格資訊可以彌補新聞資訊的不足。於是，我們提出一個新的 HOT Event Extraction 機制，利用新聞與部落格這兩大資訊來源，讓使用者可以在 Web 上有效地檢索熱門事件。

1. 導論

在Web上有非常大量新聞與部落格的資訊，如此大量的資訊常會讓使用者無法迅速找到他所想要尋找的資訊，另外我們觀察到新聞與部落格的一些差異性，有些在部落格很熱門的資訊，新聞上卻只有一點點資訊存在，像是轟動一時的”巴士大叔”就是從部落格先興盛起來的，某些Query在新聞找不到資訊，而部落格卻有和Query相關的訊息，像是”彎彎”等，因此在本論文中我們提出方法利用部落格的資訊來彌補新聞資訊的不足。另外部落格與新聞的觀點也常常不一樣，像是單單一個”林義傑橫越撒哈拉沙漠”這個事件主題，新聞通常會夾雜一些政治人物的看法，例如像是”蘇貞昌說…”，而部落格中通常會是發布者的觀點，因此一些常用詞的分析就會有差異，我們可以利用部落格和新聞在相同事件的不同觀點，來找到一些可能是熱門的事件。因此我們想藉由新聞與部落格這兩個網路資訊來源，並利用其資訊內容不同的觀點彼此互補，來找到熱門事件，讓使用者可以很快速從新聞和部落格找到他們所想要的資訊。

本研究主要的目標在於找出熱門事件，而我們簡單地定義熱門事件為很多人想要知道的事件，我們利用有別於 TFIDF 的 Bursty 方法，利用資訊量與時間持續性的關係來找出 Terms(關鍵詞)的 Bursty，而且我們進一步利用部落格每篇文章的的瀏覽訊息，來幫助找到含有 Popularity 性質的 Terms，除此之外我們另外使用時間標記來得到 Terms 的 Novelty 程度，藉由這三項因素來計算 Terms 的 HOT Score，利用這些 Terms 的 HOT Score，我們可以藉此計算出每個 Event 的熱門程度，當這些 Events 包含越多這樣的 terms 或者這些 Terms 的 HOT Score 高，我們認為這 Event 很有可能是熱門的事件，而這些事件會是人們所想知曉的。

2. 相關研究

2.1 TDT (Topic Detection and Tracking)

TDT (Topic Detection and Tracking)就是一種針對整串資訊流的分析，可以偵測和追蹤Topic的流動。Yang et al. (2002) 提出達成First Story Detection (FSD)任務的方法，也就是找出新發生的事件。(1)利用Supervised Learning Algorithm將文件先做Topic的分類，將文件分成他們所屬於的類別。(2)找出文件的Name Entities，他們使用BBN's Hidden Markov Model Software這個工具擷取出文件中7種形式的Named Entities，其中包含了Person, Organization, Location, Date, Time, Money 和 Percent這七種。(3)最後再利用文件的Term和Name Entities以TDIDF的權重重新表示文章當成Feature Vector，並用此偵測出新發生的事件。我們考慮他們提出的Name Entity的方法，因為當有大量文件，且文件內容雜亂，Name Entity當做重要的Term，會是一大幫助，不過本研究所尋找的HOT Event與他們所要的FSD有所不同，本研究偏重在更大範圍的資訊結果，找出和某一Name Entity相關的熱門新聞。

2.2 News System

Gabrilovich et al. (2004) 建立Newsjunkie 系統，這個系統主要是提供使用者個人化的新聞並且找出Novelty的事件。Newsjunkie這個系統利用Topic Words和Name Entities來表示新聞，且利用這兩個性質來找出新的新聞。這是一個以新聞為主的搜尋系統，與本研究的系統很相似，他們也採用Name Entity的方式幫助使用者找到最新資訊，除了都使用Novelty外，另外我們提出兩個不同的熱門指標Bursty和Popularity。

Del Corso et al. (2005) 利用兩種關係屬性建立出News Graph，(1)News Source和News Article間的關係，也就是News Source發佈News Article。(2)News Article和News Article間的關係，也就是報導相同新聞事件的關係。為了利用這兩種關係建立一個Graph，他們提出Time-Aware Ranking Algorithms，類似HITS Algorithms (Kleinberg 1998)，另外還考慮時間的因素，將News Article排序，給予使用者最重要的News Article。這篇論文提出有別於傳統找出最新新聞事件的新方法，他們主要是想找出最重要的新聞，方法是利用新聞來源與新聞文章的關係，與本研究最大的相關性在於他們是由新聞來源與新聞文章所推薦出來的重要新聞，而本研究則是由部落格與新聞文章利用HOT Event Extraction所選出的結果。

2.3 Weblog System

Mei et al. (2006) 提出一個可以同時找到Subtopic Themes和Spatiotemporal Theme Patterns的機率模型，並分析這些Patterns在數個領域的應用。他們採取三個步驟找到那些與時間相關的Theme Patterns (1)從Weblog擷取Common Themes；(2) 對每一位置產生Theme Life Cycles；(3)

對每一時段產生 Theme Snap-Shots。他們並且說明他們所找出來的Theme Patterns可以運用於四個方面(1) Search Result Summarization (2) Public Opinion Monitoring (3) Web Analysis (4) Business Intelligence。這篇論文主要是針對部落格的研究，與本篇研究最相關的部份在於他們從部落格中所擷取出來的Themes，另外他們提出了四項應用，是否我們的HOT Event Extraction方法也能在此四項應用上有所幫助。

3. 方法

3.1 系統架構

我們想要建立一個尋找HOT Event的新聞搜尋系統，利用部落格與新聞這兩大資訊來源，幫助使用者更容易更方便查詢到與 Query 相關的熱門資訊。

圖 1 為我們的系統架構圖，主要分成三大部分。第一部分為 Finding High Frequency Relevant Terms，這個部分的工作主要從新聞和部落格文件中尋找出與Query相關的Terms。第二部分 Finding Term's Novelty，Bursty，and Popularity，這部分主要在尋找出與 Query 相關的 Terms 的 Novelty， Bursty ，和Popularity。第三部分為Extracting HOT Events，主要是利用 Terms 的 Novelty，Bursty，和 Popularity 來找出 HOT Events。

3.2 HOT Event

我們定義HOT Event由Novelty、Bursty和Popularity這三個要素所構成。

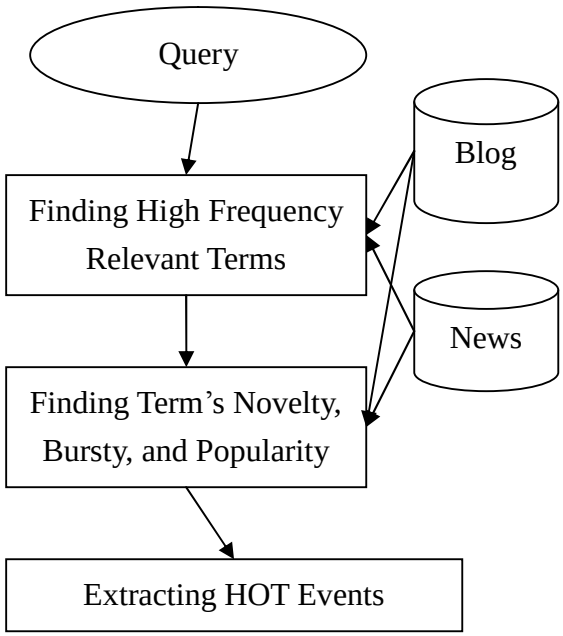

圖 1 系統架構圖

一般來講 Novelty 代表著較新穎的事物，人們總是對新的東西感到好奇，對新發生的事想要去一探究竟，於是人們會開始去追尋對這些新的事物的認知與了解，所以我們認為Novelty是HOT Event中不可或缺的重要特性。

而Bursty也就是忽然間媒體所爭相報導的事物，這種現象主要在於這些事情可能很重要，且大眾似乎也很想要知道這樣的訊息，於是媒體會開始大量的報導，這樣的 HOT Event 就是由媒體所帶動出來的一種熱門。

最後Popularity，在Web2.0的時代，User 參與度和資訊分享概念，帶動著Blog的大量興起，User可以自由自在編輯自己的Blog，在Blog發表屬於自己的文章，而這些文章可以表現出使用者的想法。於是我們利用 Blog 的討論程度所形成的 Popularity，替HOT Event加入新的要素，這個要素的

最大特色是考慮使用者的參與度。

因此，我們運用這三個要素來定義 HOT Event，也就是新穎性、流行性和重要性，預期這三個要素能有效的定義 HOT Event，而且我們提出的 HOT Event Extraction 機制能達到大部分的使用者需求。

3.3 尋找相關的 Terms

為了擷取新聞和部落格中的 Name Entity，我們使用 CKIP POS Tagger(中央研究院詞庫小組的中文斷詞詞類標記系統)，我們取詞類標記為 Na、Nb、Nc 這三類的 Term，其中 Na 為一般名詞、Nb 為專有名詞、Nc 為地方名詞，我們以這三類帶有語意型態的名詞，當成是和 Query 相關的 Terms。

首先我們找出 Query 相關的新聞和部落格文件，在此我們將新聞和部落格的文件分開來處理。將這些文件送入 CKIP POS Tagger，我們找出相關的一般名詞、專有名詞和地方名詞，例如 Query“林義傑”我們可以從相關新聞找到一般名詞“馬拉松”、專有名詞“查理、雷伊”和地方名詞“撒哈拉沙漠”。我們認為大部分的名詞能貼近使用者所想要的資訊，且 CKIP POS Tagger 所得到這三類的名詞的詞類標記可以給定簡單的人地物的語意類別，如此能提升 Term 的品質，不選擇動詞的 Term 在於文章中有太多的動詞屬於 Common Terms 會產生雜訊。

根據上述程序，我們得到新聞的相關 Terms 和部落格的相關 Terms，然後統計出三類名詞出現的頻率，我們取出較高頻的詞，這些高頻的詞我們定義為和 Query 相關的 Terms，因為高頻的 Terms 代表常和 Query 同時出現在同一篇文章，也就是表示這些 Terms 與 Query 的相依性高，因而我們認定這些 Terms 為相關。對每個名詞詞類的 Term 我們取其頻率前 20 名的 Terms 當作是有代表性的相關 Terms。於是我們可以得到與新聞相關的前 20 名的高頻 Terms 和與部落格相關的前 20 名的高頻 Terms。

3.4 擷取 HOT Events

首先我們先將每個時間區段與 Query 相關的 News 文件斷成一句一句，然後我們以三個句子合成一個 Passage，這些 Passages 就是我們之後運算的基本單位。三個句子所形成 Passage 的例子如“林義傑在去年十一月二日起與查理、雷伊並肩橫越撒哈拉沙漠 - 從最西緣的塞內加爾東行 - 經茅利塔尼亞、馬利、尼日、利比亞”。

再來便是計算這些 Passages 的 HOT Score，而這部分主要由 Novelty、Bursty 和 Popularity 這三個特性來計算出，我們計算出 Query 相關 Terms 的 Novelty 值、Bursty 值和 Popularity 值，再由這三個特性結合得到 HOT Score，然後將 Passages 中所有相關 Term 的 HOT Score 累加，以得到 Passages 的 HOT Score，利用這些 Passages 的 HOT Score 我們可以將時間區段內的 Passages 做排序，再然後找出 HOT Score 最大的 Passages 當成 HOT Events。

接下來，我們依序介紹 Novelty、Bursty 和 Popularity 的計算方式，以及一些其他部分的細節。

3.4.1 Novelty

Novelty 主要在模擬事物的新穎度，我們利用 Equiangular Spiral[1]的半徑長來模擬 Novelty 的值，

$$n_i = A\exp(\theta_i \times \cotB) \qquad (1)$$

式子(1)中，n_i 代表 i^{th} *Term* 未經過 Normalization 的 Novelty 值，A 是用來控制 n_i 的長度差，θ_i 是 i^{th} *Term* 與基準日相差的天數，cotB 是來控制每隔 360 天 n_i 相差的倍數。其中我們把 1 天當成 1 度來看待，每隔 360 天其半徑會相差一倍。我們有人工設定 A=1、B=$7\pi/16$，為了不讓每天的 Novelty 的差異度太大，如此每隔 360 天 n_i 會相差 3.4 倍。我們以 2007 年 1 月 1 日定義為基準日(Initial Time)，θ_i 就代表與基準日相差的天數。當越接近最近發生或剛發的事，其 θ_i 值就會較大，使得 n_i 也會比較大，而其 Novelty 值就會相對的比較高，細部如圖 2 所示。最後我們把所有的值做 Normalization，主要是方便與之後的 Bursty 的 Popularity 的整合計算。

式子(2) N_i 代表經過 Normalization 後，i^{th} *Term* 的 Novelty 值。

$$N_i = \frac{n_i}{\sum_{n_k \in \{n\}} n_k} \qquad (2)$$

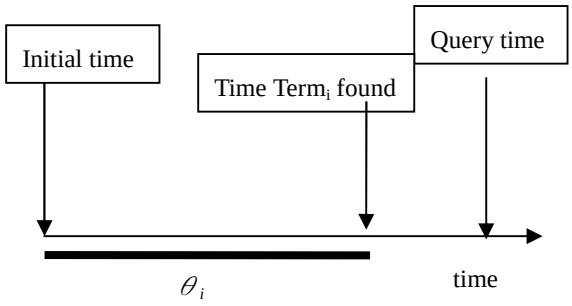

圖 2 i^{th} Term 的 Novelty 計算

3.4.2 Bursty

Bursty 主要在模擬媒體報導量的 Burst，我們找出每個 Burst 的持續時間和報導的最大量，時間的持續性代表著這個話題是否使用者需要持續性的了解，而最大量可以代表著很多使用者當時可能都很想知道這些話題，因而，我們可有這兩個變數得到相關 Terms 的 Bursty。最後我們把所有的 Bursty 值 Normalization，以方便之後的運算。

$$b_i = \sum_{Peak_j \in \{Peak\}} D_{ij} \times H_{ij} \qquad (3)$$

式子(3)中，b_i 為 i^{th} *Term* 未經過 Normalization 的 Bursty 值，D_{ij} 為 i^{th} *Term* 在 $Peak_j$ 的 Duration，也就是 *Peak* 的持續時間，H_{ij} 為 i^{th} *Term* 在 $Peak_j$ 的 Height，也就是 *Peak* 的最高點。

式子(4)為 Normalization 的式子，將 Bursty 做 Normalization，B_i 為 i^{th} 的 *Term* 的 Bursty 值。

$$B_i = \frac{b_i}{\sum_{b_k \in \{b\}} b_k} \qquad (4)$$

[1]http://www-history.mcs.st-andrews.ac.uk/history/Curves/Equiangular.html

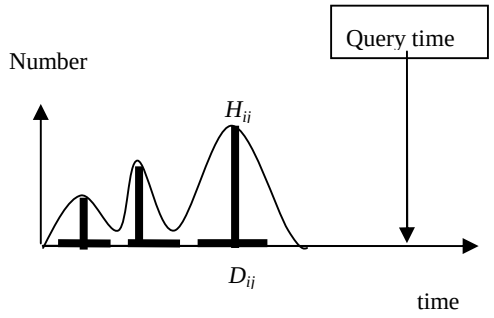

圖 3 i[th] Term 的 Bursty 計算

3.4.3 Popularity

Popularity 主要在模擬使用者對這些事件的關注程度，我們認為越多人瀏覽的東西可以代表著更多人想知道的事物，因此，我們利用使用者所發表的 Blog 文章並紀錄 Crawler 抓取這些文章時的瀏覽度來定義這個 Popularity 值，我們將一段時間內關於某個話題的瀏覽度全部加起來。

$$p_i = L_i \qquad (5)$$

式子(5)，p_i 代表 i^{th} *Term* 未經過 Normalization 的 Popularity，L_i 為 Query 與 i^{th} *Term* 在 Blog 中被瀏覽的次數。最後，我們利用式子(6)做 Normalization，其中 P_i 代表 i^{th} *Term* 的 Popularity 值。

$$P_i = \frac{p_i}{\sum_{p_k \in \{p\}} p_k} \qquad (6)$$

3.4.4 利用 Novelty、 Bursty 和 Popularity 計算出 HOT Score

我們定義 HOT Score 由 Novelty、Bursty 和 Popularity 的結合，如式子(7)所表示

$$HOT_i = aN_i + bB_i + cP_i \qquad (7)$$

式子(7)中，我們利用參數 a=0.336、b=0.34、c=0.324 將式子(2)(4)(6)線性結合得道 HOT_i，而 HOT_i 代表 i^{th} *Term* 的 HOT Score。式子(8)中，P_e 就是 e^{th} *Passage* 中的 Terms，HS_{P_e} 代表著 e^{th} *Passage* 的 HOT Score，我將 e^{th} *Passage* 中所包含 Term 的 HOT Score，加權起來成為 e^{th} *Passage* 的 HOT Score。

$$HS_{P_e} = \sum_{Term_i \in P_e} HOT_i \qquad (8)$$

4. 實驗結果

4.1 資料收集

我們利用一個自行設計的 Crawler 收集五個月(2007/01/01 ~ 2007/05/31) 的新聞和部落格資料，每天大概有 1000 則的新聞文章和 5000 篇部落格文章，其中新聞來源是從 Yam 天空(http://www.yam.com)所取得，主要是抓取 Yam 天空每天所發佈整理的新聞，其中包含中廣新聞網、聯合新聞網、中央社 、路透社、中央商情網、中時電子報、法新社、鉅亨網、TVBS 與大台灣旅遊網這些媒體。部落格，主要來源從 Yam 天空部落、Yahoo 部落格與 UDN 網路城邦。

實驗的 Query 為人工挑選的，目前只用了 10 個 Query 做一些實驗測試，Query 為林義傑、黃海岱、黃富生、馬兆駿、陳綺貞、海珊、明華園、艾迪墨菲、葛萊美獎和比爾蓋茲。

4.2 實驗設計

我們將每一則找出來的 Event 評

估其熱門程度(HOT Level)，評估方式為人工標定，分成五個等級的評估，將熱門程度從 1 分到 5 分，5 分代表最熱門，1 分代表最冷門。

4.3 熱門事件因素分析

這部分的實驗主要在測試三種因素 Novelty(N) 、 Bursty(B) 與 Popularity(P)的各自影響力，我們利用這三個因素的所有組合，找出各個因素組合的 TOP10 HOT Events 的總合 HOT Score，其結果如同表 1 所示。

表 1 使用不同要素組合的 HOT Score

Feature	TOP10 HOT Events HOT Score
Novelty(N)	36.2
Bursty(B)	36.7
Popularity(P)	34.9
Novelty + Bursty (N+B)	36.1
Novelty + Popularity (N+P)	35.7
Bursty + Popularity (B+P)	36.6
Novelty + Bursty + Popularity (N+B+P)	37.3

由圖 4 看出，使用 N+B+P 三個要素的結果會是最好的，若是使用兩個要素 B+P 得到最佳的結果，若是只考慮一項單獨要素，以 Bursty 的結果較好。我們由此可知，實際上 Bursty 的效用最大，單用 Popularity 並沒有我們預期的效果出現，但是 Popularity 與 Bursty 的結果有加分的效果。而新聞的 Novelty 與 Bursty 再加上了部落格的 Popularity 後，其結果會比單單只用新聞 Novelty 與 Bursty 的結果好。

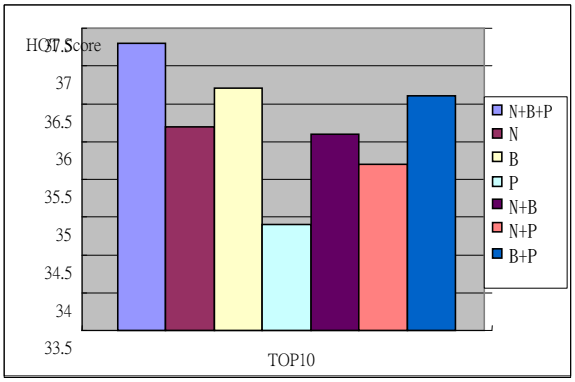

圖 4 Novelty(N) 、 Bursty(B) 、 Popularity(P)不同組合的 HOT Score 比較

4.4 與其他方法效能比較

這一部分的實驗，我們利用 TF、TFIDF 與 Ordering by Time (OBT) 的方法當作 Baseline 來與我們 HOT Event Extraction (HEE) 方法做比較，主要想檢視 HOT Event Extraction 的效能，表 2 為我們測試四個方法在不同個數的 HOT Event 的 HOT Score 的結果。

我們由圖 5 可看出 HOT Event Extraction (HEE)與 TF 所表現的的結果相近，我們認為結果相近的原因在於 HOT Event Extraction 也有使用到 TF 的因素，而利用 HOT Event Extraction 會比 TF 的結果稍微好一些，TFIDF 效果其次，而利用 OBT 的方法其結果普遍比較差，因為比較熱門的事件不一定會是最新的消息，這部份也驗證了只使用 Novelty 的誤差會比較高的結果。

表 2 HOT Event Extraction(HEE)與 Baseline 方法的 HOT Score 比較

Method	TOP N HOT Events HOT Score				
	N=10	N=20	N=30	N=50	N=100
HEE	37.3	69.1	99.8	156	306.9
TF	35.9	67.4	96.9	154.7	314.5
TFIDF	34.1	65.1	94.9	151.2	307.3
OBT	29.6	58.9	87.8	142.4	287.9

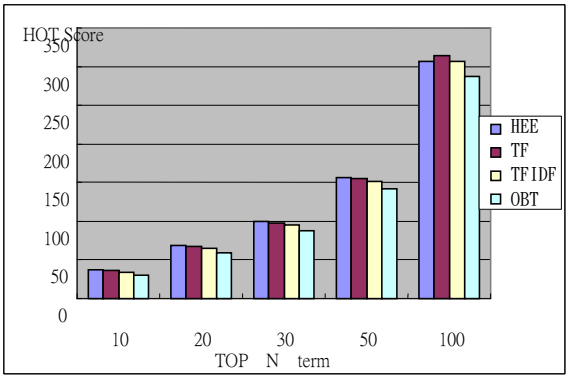

圖 5 HOT Event Extraction (HEE) 方法與其他 Baseline 方法比較

5. 結論

我們建立一個 HOT Event Extraction 的搜尋系統，利用這個系統可以幫助始用者快速獲取資訊。系統結合了新聞與部落格這兩大資訊來源，利用這來兩大資訊來源以及 HOT Event Extraction 的方法，提供使用者檢索的功能，讓使用者可以更有效率地找到其所想要的相關資訊。

我們利用實驗驗證了，使用時間區段可以產生正面的效果，會使事件比較集中，因此我們的系統能夠將熱門時間區段的熱門事件傳送給使用者。另外，我們驗證了 HOT Event Extraction 對於整個搜尋結果會有少量的改善，這部份會比單純只用時間來排序或是 TFIDF 來的好。但是我們相信如果能再利用一些其他的因素和方法來改善 HOT Event Extraction 的效能，我們相信 HOT Event Extraction 會有更好的貢獻。

從實驗結果我們也發現利用三種特性 Novelty、Bursty 和 Popularity 的 HOT Event Extraction 會比 TF 的效果好一點，當 Term Frequency 一樣的時候，我們 HEE 方法可以用 Novelty 或 Popularity 找出較有效的 HOT Terms。另外，如何讓 HOT Event Extraction 更顯出它的效果來，將會是我們所需要改進的主要方向。

我們認為 HOT Event Extraction 可以應用在其他方面，像是 Search Result 的重新排序，熱門事件的推薦系統，或是一些不同事件的評比，和一些部落格的搜尋等，HOT Event Extraction 可幫助找到一些熱門的事件，尤其是在那麼龐大與雜亂的部落格資訊中，可藉此過濾掉一些不必要的資訊。

最後我們討論到使用者對 HOT Event 認知上的差異，這部份的差異並非我們用一個簡單的機制所能去模擬的。但認知上的差異，仍是一個值得我去探討的問題，也是一個非常有趣的工作方向。

6. 未來工作

目前我們系統所使用的部落格與新聞資訊來源並不夠完整，基於版權問題與對這些來源網站的 Request 次數限制，我們只能抓取到其中的一部分資訊，尤其是在部落格這個部分，由於部落格的資訊量實在是太大了，

我們必須有更強大的 Crawler 來抓取這些資訊。另外就是部落格文章的品質過濾，這部分再將來也會是一個有趣的方向，因為資訊量大，相對的雜訊也會變大，如何從那麼大量的的文章中，擷取出有效的資訊，會是一個仍待解決的重要問題。

Name Entity Identification 這部份會影響整個 HOT Event Extraction 結果，有效的 Name Entity Identification 工具會幫助我們有效地找出和 Query 相關的 HOT Terms，對於 HOT Event Extraction 有正面的影響，或許我們還可以利用這些 Name Entity 所建立的社群關係，幫助資訊的找尋。

未來我們也將考慮加入使用者的 Click Stream，因為 Click Stream 能模擬出使用者的意圖，這部分也許可以帶給使用者更完善的搜尋結果與熱門推薦，也可運用在個人化的資訊整理。在未來等系統上線後，我們將可得到這些使用者的 Click Stream 資訊，之後我們將把這些資訊引用到我們 HOT Event Extraction 的方法中，預期這將會能使整個系統的效能有所提升。

最後，我們相信新聞與部落格是無國界的，中文資訊只是 Web 上的一小角，使用者所想要的國際資訊有時候只能在國外的新聞或部落格中找到，而且國內外的部落格資訊也會有互相流通影響的情況。此時跨語言的新聞檢索系統將是無法避免的，我們必須要能分析到其他語言的資訊，而這也會是未來發展的趨勢與方向，這時候整合國內外的新聞與部落格資源將會是未來發展的重要議題。

7. 參考文獻

M. Atallah and R. Gwadera. Detection of significant sets of episodes in event sequences. In Proceedings of the International Data Mining Conference, pages 3－10, 2004.

S. Boykin and A. Merlino. Machine learning of event segmentation for news on demand. Commun. ACM, 43(2):35-41, 2000.

C. H. Brooks and N. Montanez. Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering. In Proceedings of the 15th International WWW Conference, pages 625－631, 2006.

S. Chung and D. McLeod. Dynamic topic mining from news stream data. In Proceedings of International Conference on Ontologies, Databases and Applications of Semantics, pages 653－670, 2003.

K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In Proceedings of the 11th Text Retrieval Conference. National Institute of Standards and Technology, 2002.

A. Das, M. Datar, A. Garg. Google News Personalization: Scalable Online Collaborative Filtering. In Proceedings of the 16th International WWW Conference, pages 271－280, 2007.

G. M. Del Corso, A. Gullf and F. Romani. Ranking a Stream of News. In Proceedings of the 14th International WWW Conference,

pages 97 – 106, 2005.

E. Gabrilovich, S. Dumais, and E. Horvitz. Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In Proceedings of the 13th International WWW Conference, pages 482 – 490, 2004.

N. Glance, M. Hurst, and T. Tornkiyo. Blogpulse: Automated trend discovery for weblogs. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.

M. Henzinger, B. Chang, B. Milch, and S. Brin. Query-free news search. In Proceedings of the 12th International WWW Conference, pages 1 – 10, 2003.

J. Kleinberg. Authoritative Sources in a Hyperlinked Environment, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 46(5), 604-632, 1998.

J. Kleinberg. Bursty and hierarchical structure in streams. In Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, 2002.

R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In Proceedings of the 12th International Conference on World Wide Web, pages 568{576, 2003.

G. Kumaran and J. Allan. Text classification and named entities for new event detection. In Proc. of the SIGIR Conference on Research and Development in Information Retrieval, 2004.

Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In Proceedings of SIGIR'05, pages 106-113, 2005.

Q. Mei, C. Liu and H. Su. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In Proceedings of the 15th International WWW Conference, pages 533 – 542, 2006.

Q. Mei, X. Ling, M. Wondra, H. Su and C.X. Zhai. Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In Proceedings of the 16th International WWW Conference, pages 171 – 180, 2007.

Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of KDD '05, pages 198{207, 2005.

J. Perkio, W. Buntine, and S. Perttu. Exploring independent trends in a topic-based search engine. In Proceedings of WI'04, pages 664{668, 2004.

R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In Proceedings of the ACM SIGKDD 2000 Workshop on Text Mining, pages 73 – 80, 2000.

Y. Yang, J. Zhang, J. Carbonell, and C. Jin. Topic-conditioned novelty detection. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, pages 688 – 693, 2002.