News Recommendation based on Entropy Computation

[1]Yu-Min Zhang, [1,*]Ying-Chih Lin and [2]Xun Zhou
[1]Department of Applied Mathematics,
[2]Industrial Ph.D. Program of Internet of Things,
Feng Chia University, Taichung, Taiwan, R.O.C.
[*]yichlin@fcu.edu.tw

Abstract

在自由開放的多元化社會裡，各家網路媒體從不同角度切入的報導突顯出特定事件的眾多觀點，這對於激發各種思考與討論有正面的裨益。然而，對讀者來說，想獲取客觀且完整的資訊勢必得透過閱讀各家網媒的論述，這樣耗時且缺乏效率的方式，大大提升讀者通盤了解事件的門檻。因此，本研究以熵(Entropy)為基礎，計算文章的資訊含量，以此建構一套即時的新聞推薦機制，應用於新聞關聯性的計算並推薦予讀者，以協助讀者廣泛地獲取資訊。本研究的實驗是針對兩岸、社會、運動以及國際類的新聞進行推薦效果的測試，結果顯示對於兩岸及國際類的新聞有不錯的推薦效果，而對於社會、生活及運動類的新聞，推薦的成效則有待改善。

1. Introduction

語言是人類重要的溝通媒介，而描述一個事件更是一個個字詞所建構出來的文章，在進行大多數語言相關的處理及應用前，都必須能夠分辨出該文章所包含的字詞，因此分辨字詞是一道不可或缺的環節。不同於英文單詞間以空白隔開，中文是接連數個字詞所緊密串接而成的字串，因此中文文章得預先經過適當切割，才能分辨文章內含有那些字詞，故中文斷詞(Chinese Segmentation)是處理中文文章不可或缺的技術之一，也是許多文章分析的基礎技術。再者，雖然斷字斷詞有基於字串匹配、人工智慧、統計學習等眾多方法，但以基於字串匹配的方式，也就是詞庫斷詞法最為直覺，也最簡單。這個方法先利用標點符號將整篇文章斷開成句子的集合，再針對每一個句子由左到右依序掃描，遇到詞庫裡有的詞就標示出來，遇到不認識的字串就分割成單字詞；若遇到複合詞（如土地公）就以長詞優先法(maximum matching)為基礎[1]。這個作法的運算原則是「長詞優於短詞」，也就是由句子的一端開始，試著比對出在詞庫中最長的詞作為斷詞的結果。例如詞庫裡有「土地公」、「土地」、「公」這三個字詞，當在文章中遇到「土地公」時會優先把這三個字斷在一起，而非斷成「土地」與「公」兩個詞；接著將最大匹配後的詞排除，剩餘部分再重複利用長詞優先法進行斷詞，直到處理完整篇文章為止。一般來說，如果使用的詞庫夠大，文獻指出長詞優先法斷詞可達到90%以上的斷詞準確率[2]，在複雜性與成本考量的情況下，這個作法有還不錯的效果。

台灣社會在民主化浪潮的席捲下，再加上網際網路的普及，立場不再被箝制的網路媒體（簡稱「網媒」）越來越多，且各家網媒對事件的描述往往自行其是，更甚者有斷章取義的情形發生。從一個多元化社會的角度來看，這些從不同角度切入的文章報導正好突顯出各種不同的觀點，但是網媒針對特定事件而不同觀點立場的描述，往往模糊了焦點，表達出來的意見也常常莫衷一是。描述同一事件時，各家網媒發佈的新聞內文除了既有的政治立場、觀點外，尚有文章資訊量多寡的差異性，且各個領域的新聞都存在著類似問題。

多元化社會裡，許多網媒形成百家齊鳴的情形，對於激發各種思考當然有著正面的影響，然而，鑒於同事不同工的現象，讀者欲獲取客觀且完整的資訊勢必得透過閱讀各家網媒的論述，這樣耗時且缺乏效率的方式，大大降低讀者通盤了解事件的可能性。因此，降低讀者廣泛閱讀新聞的門檻，才能讓讀者能輕易以更多更全面的方式得到事件的各種面貌。基於上述理由，本研究以計算文章的資訊量為基礎，建構一套即時新聞推薦機制，應用於新聞關聯性的計算並推薦予讀者，以協助讀者廣泛的獲取資訊。

2. Related Works

推薦系統(Recommend System)是根據使用者的興趣偏好、購買行為等，向使用者推薦可能有興趣的資訊或商品。為了解決資訊超載帶來的困擾，推薦系統應運而生，透過分析使用者的行為，發現使用者的個人化需求、偏好等，然後再推薦給使用者感興趣的資訊或商品。以 Google、百度為代表的搜尋引擎能透過輸入關鍵字準確地找到自己想要的相關資訊，但如果使用者無法好好描述自己心中所想的，那麼搜尋引擎也派不上用場。和搜尋引擎不同，推薦系統不需要使用者提供明確的需求訊息，相對來說，推薦系統屬於主動出擊的類型。推薦系統已經廣泛應用在許多領域，而最典型且

已經成形的該是電子商務領域。推薦系統的做法主要有協同過濾推薦、以內容為基礎的推薦與混合推薦等三種[3]，說明如下：

- 根據使用者的歷史資訊來做推薦，這種方法稱為協同過濾推薦(Collaborative Filtering Recommendation)。協作過濾原理是先找到與使用者有類似興趣的其他使用者，然後將他人有興趣的內容推薦給目前的使用者。其基本想法相當直覺，就是透過他人的推薦來作選擇。

- 利用商品的內容描述，計算使用者的興趣與商品之間的相似度來作為推薦的依據，則稱為以內容為基礎的推薦(Content-based Recommendation)。這種做法不需要其他使用者的幫助，不用倚賴其他使用者對商品的評價意見，沒有新上架或冷門商品的問題，甚至可以進一步用在商品設計與開發的評估。

- 由於推薦方法各有所長，所以在實際應用中常常使用混合推薦(Hybrid Recommendation)的做法，例如分別用以內容推薦方法和協同過濾法產生一個推薦預測結果，然後再組合起來，以強化各推薦技術的優點並彌補其缺陷。

另一方面，熵(Entropy)是常見不確定性的數學度量之一，其概念最早起源於熱力學，用於度量一個系統的無序程度，Shannon 於 1948 年將熵的概念引入資訊理論中[4]。當不確定性越高則熵亦越高，意味著具有更多的資訊量；反之，不確定性越低則熵也越低，代表包含較少的資訊。所以當某個事件各種可能發生情形的機率為平均分布時，那麼該事件的平均訊息量會最高，此平均訊息量就是所謂的熵。假設有一個事件 X，其值域為$\{x_1, x_2, ..., x_n\}$，機率分布為$\{p(x_1), p(x_2), ..., p(x_n)\}$，則熵的計算方式為$-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$。而以最大熵原則為基礎的最大熵模型也已經廣泛出現在各個領域，在自然語言中的詞性標注、中文斷詞、機器翻譯中也都經常出現[5]。

此外，一個字詞的重要性應該隨著它在文件中出現的次數成正比增加，但同時也應該隨著它在文件集合中出現的頻率成反比下降。假設一篇文章 A 包含有 N 個詞 $w_1, w_2, \ldots, w_N$，這些詞在另一篇文章 B 對應的詞頻(Term Frequency, TF)分別為 $TF_1, TF_2, \ldots, TF_N$，那麼文章 A 和 B 的相關性可以表示為 $\Sigma(TF_i)$；另外，假定一個詞 w 在 D_w 篇文章出現過，那麼使用逆文本頻率指數(Inverse Document Frequency, IDF)來評估，公式為 $\log_2(D/D_w)$，其中 D 是文章的總篇數，所以若一個詞出現在大量文章中，則 D/D_w 的比值會很接近 1，IDF 也就很接近 0。綜合起來，TF-IDF 的計算方式為 $\Sigma(TF_i \times IDF_i)$，以此作為判斷兩篇文章相似度以及文章相關性排序的依據[6]。

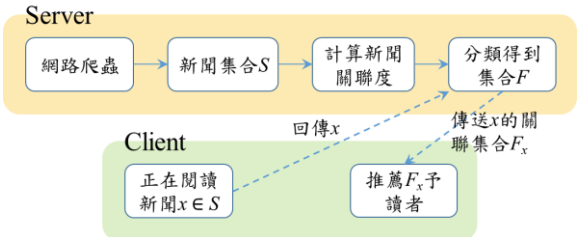

Figure 1. 新聞推薦機制的架構

3. Material and Method

本研究在於建構一套即時的新聞推薦機制，應用於新聞關聯性的計算並推薦予讀者，以降低讀者廣泛閱讀新聞的門檻，讓讀者能輕易以更全面的獲得事件的各種面貌及看法。Figure 1 是這個機制的架構圖，在一開始系統會透過網路爬蟲程式抓取各家網媒的每日新聞，並予以中文斷詞後交由 TF-IDF 演算法、餘弦相似性公式、Entropy 計算任兩篇新聞文章間的關聯性評比，最後加以整理儲存在 Server 端以供 Client 傳送及回傳資訊。底下說明每個程序的步驟：

3.1. 中文斷詞與詞庫

一篇新聞乃至一篇文章，其實就是各種字詞的堆砌，不同詞語間的組合形成語意各自迥異的文章，因此可將字詞視作一篇文章或一篇新聞的主體。而為了做更進一步的相似度計算，一個要先進行的程序是中文斷詞處理，以期將新聞文章轉變成詞語集合。此外，在各種中文斷詞法中，使用詞庫來斷詞不僅比較簡單，也可以達到不錯的成效，因此本研究以開源的中文斷詞程式 Jieba [7]使用的簡體詞庫為基礎，先將該詞庫（共十萬餘詞）轉成繁體，再添加一些常用詞彙整合成我們的詞庫，例如教育部成語典、大學校院名稱、地名、台灣名人等，總計十八萬餘詞。我們沒有直接使用 Jieba 來幫忙斷詞，因為 Jieba 的斷詞方式是先產生句子裡中文字所有可能成詞的情況，再使用動態規劃法(Dynamic Programming)找出基於詞頻的最大機率路徑，以此作為斷詞結果。然而，在我們的詞庫中缺少正確詞頻的紀錄，即便使用現有的詞頻，在測試時也發現斷詞結果不甚理想；同時，礙於經費的成本考量，我們也沒有合適的平衡語料庫得以統計詞頻，因此就以最簡單易做的「長詞優於短詞」為原則，自行開發中文斷詞程式，演算步驟如下：

Step 1: 將詞庫內的詞由小到大按順序排列，先比較詞的第一個字，再比較第二個字，依序比較下去。

Step 2: 讀取新聞文章 A，利用標點符號將其斷成句子的集合{A_i}。

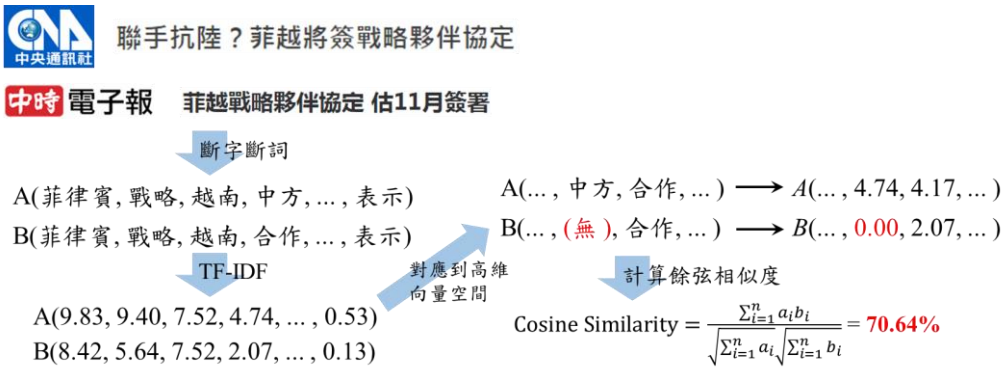

Figure 2. 兩篇文章的相似度計算過程

Step 3: 對每一個句子 A_i，由左到右依序掃描每一個字，並利用二元搜尋法比對在詞庫出現的位置。

Step 4: 繼續讀取 A_i 的下一個字並比對詞庫，直到找到最大匹配的詞。

Step 5: 重複 Step 2，直到處理完所有的句子為止。

3.2. 文章關聯性計算

TF-IDF 的想法為「如果某個詞或短語在一篇文章中出現的頻率高，並且在其他文章中很少出現，則認為此詞或者短語具有很好的類別區分能力」。簡而言之:「散佈性越大的詞語越不重要」。而一個詞 w 於一篇新聞內的權重為 TF × IDF，其中

- 詞頻(TF) ＝ 某一個詞在一篇新聞文章中出現的次數。
- 逆文檔頻率(IDF) ＝ $\log_2\left(\frac{\text{新聞文本總數}D}{\text{含}w\text{的新聞文本數量}D_w}\right)$

當兩篇新聞內的所有詞語都經過 TF-IDF 計算後，可得兩個 n 維向量 A = (a_1, a_2, …, a_n), B = (b_1, b_2, …, b_n)，接著只要利用餘弦相似性(cosine similarity)公式就能求得兩篇文章的相似程度。舉一個簡單的例子，假設有兩篇新聞文章，新聞 A 取自中央通訊社，標題為「聯手抗陸？菲越將簽戰略夥伴協定」，新聞 B 則由中時電子報下載，標題為「菲越戰略夥伴協定 估 11 月簽署」，則兩篇文章的相似度可由 Figure 2 計算而得。

然而，此方法是將新聞文章內每一個詞都拿來比對，可理解為將 A、B 兩篇文章內所有信息量（所有事件）拿來比較，這時有可能會錯失掉一些資訊。以 Figure 3 為例，在新聞 A 重要的事件一，在 B 卻無足輕重，但不能因此就說兩新聞文章毫無干係，多數情形反而是 B 補述了更多 A 所沒有描述且關於事件一的資訊。因此對於剛剛閱讀完新聞 A 的讀者而言，新聞 B 即具備了高可讀性，但是以上述的計算方法，此一情形的餘弦相似性

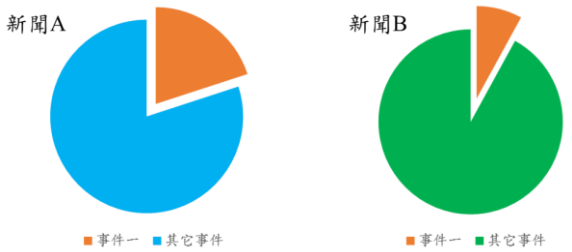

Figure 3. 新聞推薦機制的架構

會由 A、B 的其他事件（包含 B 對事件一所做的補述）沖淡而造成相似度偏低。因此本研究提出另一個關聯性算法，其主要想法如下：

"*如果在新聞 A 中最重要的詞語集合能在新聞 B 內找到相應集合，且兩者的 TF-IDF 分佈相近，則新聞 B 具有基於新聞 A 的關聯性*"

基於上述想法，假設 A = (a_1, a_2, …, $a_{n(A)}$), B = (b_1, b_2, …, $b_{n(B)}$)為兩篇文章 A、B 經過斷詞後的結果，且斷出來的字詞皆按照該詞的 TD-IDF 值由大到小排列。也就是說假如 $i < j$，則 TF-IDF(a_i, A) ≥ TF-IDF(a_j, A)、TF-IDF(b_i, B) ≥ TF-IDF(b_j, B)；再者，對於 a_i 來說，若文章 B 的斷詞結果裡沒有 a_i 這個詞，則 TF-IDF(a_i, B) = 0，反之若文章 A 的斷詞結果裡沒有 b_j 這個詞，則 TF-IDF(b_j, A) = 0。據此，取 TF-IDF 值為前 10%的詞作為關鍵詞彙，上述想法可以轉換成如下的計算公式，用以計算新聞 X 基於新聞 Y 的關聯性。換句話說，從讀者正在閱讀新聞的最重要詞語（最重要的事件）作為立基，向外計算與其他新聞的關聯性。需要注意的是這個關聯性的計算方式不具備交換律，也就是說，B 基於 A 的關聯性不等於 A 基於 B 的關聯性。

- X 基於 Y 的關聯性 ＝ $\frac{\sum_{i=1}^{n(Y)/10} TF\text{-}IDF(b_i,X) \times TF\text{-}IDF(b_i,Y)}{\sqrt{\sum_{i=1}^{n(Y)/10}(TF\text{-}IDF(b_i,X))^2} \times \sqrt{\sum_{i=1}^{n(Y)/10}(TF\text{-}IDF(b_i,Y))^2}}$

Figure 4 是利用改良過計算方法來計算 Figure 2 內兩篇文章的關聯性，結果關聯性由原本的 70.64%增加為 97.28%。此一改良後的關聯性計算

法，是將新聞 A 中最重要的詞語於新聞 A、B 所構成的兩向量作餘弦相似性計算。經過這個計算，便可於已完成斷詞的新聞集合 S 內計算兩兩文章間的關聯性，並加以分類為 F_x 集合，這是所有新聞基於新聞 x 的關聯新聞集合。例如，基於新聞 A 計算 B, C, D 的關聯性分別為 0.3, 0.6, 0.9，在門檻值設定為 0.7 的情況下，得到基於 A 文章, C 和 D 文章與之有關聯，則 F_x = {C, D }。

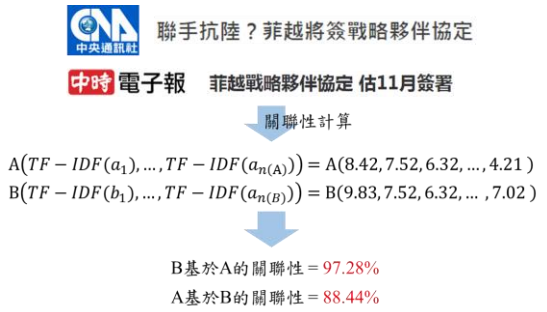

Figure 4. 改良過後的文章關聯性計算

3.3. 推薦文章的排序

文章分類的最終目的是當使用者在閱讀新聞 x 時，伺服器端能收到這個訊息，且經過搜尋新聞分類結果 F 後能回傳新聞集合 F_x 推薦予讀者。然而，F_x 裡仍然有許多與 x 相關的新聞，對於正在閱讀新聞 x 的讀者而言，F_x 裡相關的新聞訊息仍然太多太雜亂，因此這裡利用熵來衡量新聞的資訊量，再以 TF-IDF、餘弦相似性做進一步推薦指數(Recommendation Index, RI)的計算。熵的計算方式為 $-\sum_{i=1}^{n} p(w_i)\log_2 p(w_i)$，其中 $p(w_i)$代表詞 w_i 出現的機率，而在實作的過程中應用了 IDF 的統計資料，將原本新聞文本總數與含某詞的新聞文本數量比值(D/D_w)倒數成 D_w/D，以這個值衡量 w_i 在整個新聞集合中出現的機率。因此，文章 B 基於 A 的 $\mathrm{RI_A(B)}$可由下列公式計算而得：

$$\mathrm{RI_A(B)} = \frac{(1 - \mathrm{CosSim(A, B)}) \times \mathrm{Entropy(B)}}{\mathrm{Entropy}(A)}$$

這裡餘弦相似性的計算仍就使用所有詞彙的比對，這是因為兩篇新聞除了描述特定事件外，對於其他事件的資訊仍會有些許重疊，如 Figure 3 所示。推薦指數 $\mathrm{RI_A(B)}$的涵義是「當讀者讀完新聞 A 之後，再去讀新聞 B 能帶來多少對於主要事件的新資訊」，由此可知推薦指數應與新資訊量的多寡成正比。因此，比對兩篇文章的所有詞彙是為了去除掉所有交互重疊的資訊，從而推斷出此一新聞能帶給讀者多少新的資訊。最後再依照推薦指數重新排序 F_x 內的文章，再回傳予使用者，如此一來，當使用者閱讀完新聞 A 後，能接著推薦與 A 描述事件相關且資訊量最豐富的新聞，以便於讓讀者快速地掌握特定事件的全貌。

4. Experiments

實驗所使用的測試資料是透過爬蟲程式擷取中央社、自由時報、中國時報、蘋果日報、聯合報、以及法廣新聞等網站的部分新聞所組成。首先擷取這些新聞網站的歷史新聞共六千餘篇，以做為 TF-IDF 計算並分群的依據。接著從中挑出五篇做為主要的閱讀新聞，其資訊如 Table 1 所示。

文章代號	發佈時間	新聞網站	類別
	新聞標題		
A	2016/3/8	中央社	兩岸
	張志軍：不認同一中 就是改變台海現狀		
B	2016/3/8	自由時報	社會
	肇逃闖禍逃回家 車牌掉現場露餡		
C	2016/3/7	中國時報	運動
	全國高中甲級排球聯賽 豐商、東山奪冠亞軍		
D	2016/3/8	頻果日報	國際
	王毅談南海：航行自由不等於橫行自由		
E	2016/3/5	法廣	國際
	難民危機：默克爾向希臘施壓		

Table 1. 主要閱讀的新聞資訊

測試實驗進行的步驟如下：

(1). 測試者依序閱讀 Table 1 的五篇新聞文章
(2). 針對每一篇新聞 X，閱讀系統所挑出 RI 值最高的十篇新聞
(3). 測試者閱讀完十篇新聞後，對於每一篇文章給出兩個回應：
- 是否與新聞 X 相關？
- 對於了解與新聞 X 描述的事件，這十篇新聞的資訊量由大到小排序

實驗方式為邀請四位測試者分別進行，測試結果分別列於 Table 2 與 Table 3。在相關性的正確率評估，計算的是測試者對於系統所挑出十篇相關新聞的認定結果。由 Table 2 的結果可看出對於兩岸、國際類的新聞，系統能挑出比較相關的文章；而對於社會類的新聞 B，系統挑出的文章幾乎都與 B 無相關。造成這個結果的原因在於新聞 B 描述的是在新北市板橋中山路發生的一起車禍事件，系統挑出的新聞則是描述發生在其它地點的車禍及酒駕事件，測試者大多認為這與新聞 B 描述的關聯性不大，導致這部分的正確率偏低。

	User1	User2	User3	User4	avg
A	80%	100%	100%	100%	95%
B	10%	10%	40%	0%	15%
C	89%	100%	44%	78%	78%
D	70%	100%	60%	70%	75%
E	89%	89%	78%	78%	84%

Table 2. 推薦新聞相關性的正確率

接著探討的是在閱讀完新聞 X 之後，閱讀那篇新聞能更了解新聞 X 描述事件的全貌，也就是能獲取更多的資訊量。由於每個人主觀認定的資訊量相當程度的差異，所以這裡採用一個命中率的評估方式。系統根據 RI 值挑出最高的前三篇新聞，只要能落在測試者認定資訊量最高的前三篇中，這樣就算命中。換言之，在 Table 3 中以 User1 對新聞 A 而言，系統挑出 RI 值最高的前三篇中有兩篇也在 User1 認定資訊量最高的前三篇新聞中，而只有一篇新聞落在 User4 認定的前三篇文章中。

	User1	**User2**	**User3**	**User4**	avg
A	67%	67%	67%	33%	59%
B	0%	0%	0%	0%	0%
C	67%	33%	33%	67%	50%
D	33%	67%	67%	67%	59%
E	100%	100%	100%	100%	100%

Table 3. 推薦新聞的命中率

從 Table 3 的測試結果，可以看到每個測試者對新聞 B 的命中率皆為 0，這是因為 Table 2 的測試結果顯示系統挑出的新聞與新聞 B 的關聯性不大，導致這裡得到皆無法獲取資訊量的結果。再者，這裡也得到與 Table 2 一致的結果，對於兩岸、國際類的新聞，根據 RI 值大小推薦的文章能有較高的資訊含量。甚至在新聞 E 的測試中，系統挑出的三篇文章被所有測試者認定是資訊含量最高的前三名。

5. Conclusion

在一個自由開放的社會，各家言論宛如萬言堂般流傳，同時在這個資料量爆炸的時代，網路媒體多不勝數，連最貼近我們生活的新聞亦是各家網媒眾說紛紜，對於閱讀網路新聞的民眾而言，想要在這個資訊洪流中得到一個事件完整且客觀的面貌，實屬不易。因此本文章以資訊量的計算為基礎，試圖建構一個即時新聞的推薦機制。當讀者在閱讀新聞 X 時，這個機制能迅速地找出與 X 相關的新聞，並將找到的新聞按照資訊量的多寡加以排序，最後推薦予讀者。此舉讓讀者能迅速掌握整體事件的各個面向，節省許多精神與時間。從實驗結果可以得知，本文章所設計的即時推薦機制對於兩岸及國際類的新聞有不錯的效果，而對於社會、生活及運動類的新聞，推薦的成效則有待改善。

Acknowledgements

本研究工作由科技部計畫補助，計畫編號為 MOST 104-2633-S-035-001。

References

[1] K. J. Chen and S. H. Liu (1992) “Word identification for mandarin Chinese sentences,” *Proceedings of the Fifteenth International Conference on Computational Linguistics*, Nantes, pp.101–107.

[2] C. -L. Goh, M. Asahara and Y. Matsumoto (2005) “Chinese word segmentation by classification of characters,” *Computational Linguistics and Chinese Language Processing*, 10(3), pp. 381–396.

[3] F. Ricci, L. Rokach, B. Shapira and P.B. Kantor, *Recommender Systems Handbook*, Springer, 2011.

[4] C. E. Shannon (1948) “A mathematical theory of communication”, *Bell System Technical Journal*, 27(3), pp. 379–423.

[5] M. Sun, M. Zhang. D. Lin and H, Wang, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, Springer, 2013.

[6] H.C. Wu, R.W.P. Luk, K.F. Wong and K.L. Kwok (2008) “Interpreting TF-IDF term weights as making relevance decisions,” *ACM Transactions on Information Systems*, 26(3).

[7] Jieba, https://github.com/fxsjy/jieba