程式人 | 王建興

從搜尋引擎到文字探勘（中）

文件當中包含了許多非結構化的資料，可以採向量的方式來計算並排列出文件及詞相似度

文/ 王建興 | 2014-09-19 發表

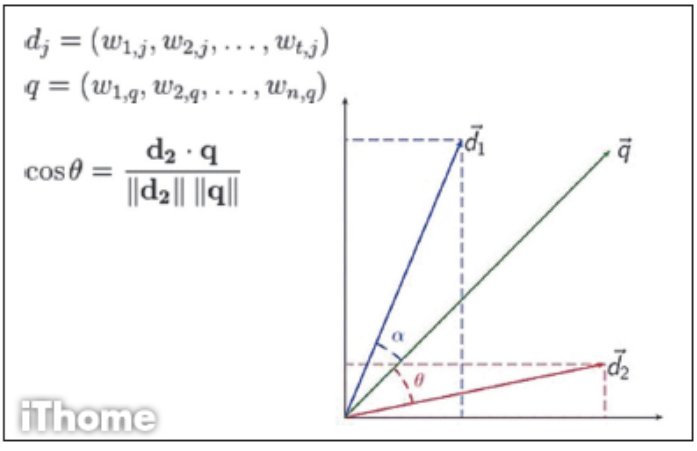
使用過搜尋引擎的人都知道，其搜尋結果不一定是完全符合所有查詢之關鍵詞，也可能是部份相符，搜尋引擎會將符合程度愈高的結果排在愈前面。當然，如 Google 之類的搜尋引擎，還會搭配其他的因素，來決定搜尋結果順位。不過，基本上，還是以相符程度為主要的基準。

搜尋引擎是怎麼決定每一份索引過的文件和查詢字串「相符程度」的呢？還記得在前文中我們提到，搜尋引擎的索引結構是將每個文件及該文件中所含有的「詞」，表示成為一個向量。這麼一來，搜尋引擎便將沒有結構的文字資料表示成為有結構的數值型資料了，而且，所有搜尋引擎所索引的文件，都成了同一個向量空間裡的向量了。

以向量來評估出文件之間的相似度

相信許多人都學過線性代數，在同一個向量空間裡的兩向量，其夾角可以透過兩向量的內積來計算得到。而夾角則關係到兩個向量的差異，當夾角愈小，代表兩向量的差異愈小，夾角愈大，代表兩向量的差異愈大。這告訴我們，若是我們可以用向量表示每個索引的文件時，就可以用文件之間的向量表示方式，求得其夾角，而評估出兩向量的差異。當夾角為零度時，代表兩向量相同，若夾角為九十度、即垂直時，則代表兩向量完全不相似。

如果我們將查詢文字表示為向量 q ，而將每份文件表示成為向量dj，那麼，我們可用利用向量的內積求出兩向量夾角 θ 的餘弦cosθ。透過便可以評估向量 q 與向量dj的相似程度。如下圖所示：



向量 d1 與向量 q 的夾角 a 小於向量 d2 與向量 q 的夾角 θ，這意謂著 d1 相較於 d2 ，與 q 更相似。只要找出所有索引的文件中與 q 足夠相似的文件，便可把它們列在搜尋結果中，而且可以依據該文件的向量表示 di 與向量 q 的cosθ值來排列，即可依據相似程度來排序。而這種評估相似性的方式，也被稱為 Consine Similarity。

綜合頻率和重要性指標，來表示一詞的權重

還記得我們是將所有的文件中總共有 n 個可能詞，分別用 w1 到wn來表示，藉以利用此 n 維的向量來表示一個文件。在這種表示方式下，我們將一個文件表示為向量 V，其中的第i個元素為 v(i)，而 v(i) 的值為詞wi的權重。

那權重應該是什麼呢？就其意義來說，權重應該是代表該詞在文件中的重要性，所以有一個很直覺的表示方式，便是計算該詞在文件中出現的頻率。這個評估方式不是沒有道理的，當某個詞在單一文件中出現的愈多次，權重自然應該要愈高。

不過，只計算詞的頻率會有個缺點，就是我們分不出不同詞的差異。舉例來說，當某個詞 A 廣泛的出現在許多文件，而另一個詞 B 卻只出現在特定的幾份文件，而它們都同樣出現在某一文件、而且次數也一樣多時，兩個詞扮演的影響力是一樣的。

但是，其實詞 B 更具備代表性，因為許多文件都可能含有詞 A ，卻只有少數文件含有詞 B。為此，有種 TF-IDF （Term Frequency - Inverse Document Frequency）的表示式發展出來。TF-IDF 的主要精神，便是綜合兩種指標來表示一詞在文件的權重。

TF（Term Frequency）代表的是詞在文件中出現的頻率，我們可以用這樣的式子來表示：

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中ni,j代表詞ti在文件dj中出現的次數，因此式子中的分母部份即代表文件dj中所出現的總詞數。為什麼不只利用ni,j是因為每篇文件的長短不盡相同，因此，必須利用分母來予以正規化，才能讓每篇文件的計算結果有相同的意義。

而 IDF（Inverse Document Frequency）則是用來表示一個詞的重要性。如果單看 TF，那麼不論該詞有多普遍，影響力都相同，但事實上，不同的詞會依其「獨特程度」而應該要有不同影響力。就如同「健康」和「氣爆」二詞若在新聞裡相比，「氣爆」肯定出現次數低的多了，相較而言，「氣爆」的獨特性也就高了許多，更具代表性了。而 IDF 正是試著要表現出這樣的特點。IDF 是怎麼評估的呢？

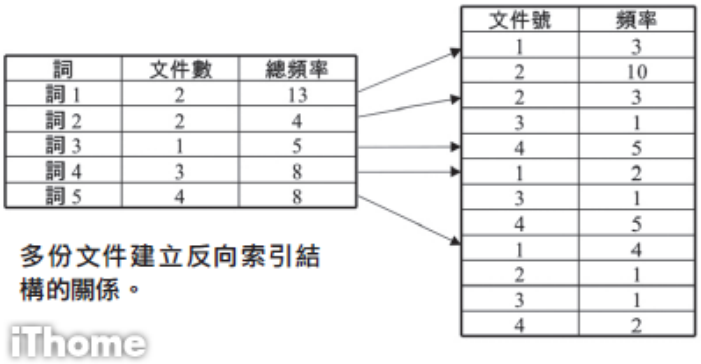
$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

基本上，詞ti的 IDF 值idfi是總文件數除以出現過詞ti的文件數後所得的商，再取對數。基本上，總文件數除以出現過詞ti的文件數後所得的商，是該詞出現於文件中之頻率的倒數，因此，才稱這個表示方式為「逆文件頻率（Inverse Document Frequency）。有了 IDF 之後，我們便可以評估某一個詞在所有文件中所扮演的重要性，而讓獨特、重要的詞能在計算時發揮更多影響力。

綜合 TF 及 IDF ，便可以同時將詞在特定單一文件中的影響力，以及某個詞在所有文件中的影響力一起納入評估，於是有了 TF-IDF。

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$$

將詞ti在文件dj中的 TF 及詞ti的 IDF相乘，即得詞ti在文件dj中的 TF-IDF 值。此值目前被普遍的用於文件向量表示中的權重表示。



詞	文件數	總頻率
詞 1	2	13
詞 2	2	4
詞 3	1	5
詞 4	3	8
詞 5	4	8

文件號	頻率
1	3
2	10
2	3
3	1
4	5
1	2
3	1
4	5
1	4
2	1
3	1
4	2

如果我們回憶搜尋引擎建立的反向索引結構：

你就會發現，這個反向索引結構基本上就記錄了每個詞的 IDF（左方表格），以及每個詞在每個文件中的 TF（右方表格），也就是說，這個結構不僅幫助我們搜尋到含有特定詞的所有文件，也可以同時計算出 TF-IDF 值，來做為文件向量中的權重。

因此，除了搜尋出結果之外，同時也可以得到排序時所需的重要資訊，因為排序時會利用計算查詢文字與文件向量的餘弦值來評估其相似性，而 TF-IDF 則是向量中表示權重的方式。

有用之處在於，它不僅利用特定詞出現的頻率，來評估該詞在特定文件中的重要性，同時還將該詞在所有文件中的獨特性也考慮進來。因此，愈獨特、在單一文件出現次數又多的詞，就可以浮現。

有些應用，像是想找出文件中的關鍵詞，就可以利用計算特定文件中每個詞的 TF-IDF 值，接著找出該值最高的若干個詞，來做為該文件的關鍵詞。像是處理新聞或專利文件時，我們想要快速的知道某段文字的「特徵」，我們可以用少數幾個關鍵詞來表示。這時，TF-IDF 就可以派上用場，幫助我們以自動的方式，計算出每一份文件的代表關鍵詞。

TF-IDF 除了關鍵詞的應用之外，也可以用於文件摘要的應用。文件摘要的常見作法是摘出文件中的重要句子。有那些句子重要呢？其中的一個方式就是評估句子中所含的詞的重要性，當一個句子中含有愈重要的詞、愈多重要的詞時，這個句子通常愈重要。這時，也可套用 TF-IDF 值來做加總計算。

由上述的例子不難明白，TF-IDF 的評估方式，對於想要以數值的方式處理文字，是扮演相當重要的角色，而它也讓我們得以連結到文字探勘的領域去。

相關報導請參考「從搜尋引擎到文字探勘（上）」「從搜尋引擎到文字探勘（下）」

作者簡介

王建興

目前在一家網路應用軟體公司擔任技術長的工作，專長是物件導向設計以及Internet應用系統的開發。他過去的研究興趣包括：點對點網路、分散式網路管理、行動式代理人、感知網路。從企業應用軟體系統，到個人行動裝置上的應用，他都有一些開發的經驗。並且對於網路創業及網路應用的發展趨勢，持續保持高度的關心。

更多 iThome相關內容

- 從搜尋引擎到文字探勘（下）
- 從搜尋引擎到文字探勘（上）
- 以雲為集散地，物聯網連接人與物
- 結合刀鋒與橫向擴充性，Pure Storage推高密度快閃陣列
- 無人汽車應用新進展 1：專業級賽車也開始挑戰要無人化
- VR應用再進化實例 2：人體VR也將取代醫學院人體解剖實習

熱門新聞

【MIS必看】WannaCry勒索病毒猖獗！上班第一天如何處理病毒未爆彈？TWCERT/CC教你6步驟自保

2017-05-14

如何躲過WannaCry勒索蠕蟲風暴？週一上班先不要開電腦，照著這些方法做

2017-05-14

XP也有救！WannaCry 2.0勒索軟體肆虐，微軟破例再釋Windows XP修補

2017-05-14

別哭！被WannaCry加密的檔案有機會救回

2017-05-16

WannaCry企業自救術

2017-05-19

微軟譴責美國政府監控政策引發WannaCry災情!

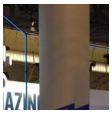
2017-05-15

史上最大WannaCry勒索蠕蟲，到底有沒有WannaCry 2.0變種？

2017-05-14

WannaCry的幕後攻擊者是誰？新證據指向北韓

2017-05-16



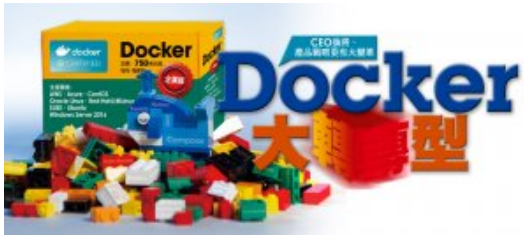
思科營收連六季負成長，準備再裁1100名員工

2017-05-18

XP用戶中了WannaCry先別慌，資安業者有解

2017-05-19

專題報導



Docker大轉型

SD-WAN崛起，企業重新定義連外網路

2017超融合應用伺服器採購大特輯【一線大廠篇】

企業Chatbot新機會

強化資安意識，鞏固惡意郵件進逼的最後防線

更多專題報導