從搜尋引擎到文字探勘（下）

許多資料探勘演算法被釋出，開發者可以省下不少時間及心力，就能直接用這些現成的方法來進行分類，在Java的開放原始碼社群中，累積了質與量兼優的產出

文/ 王建興 | 2014-09-25 發表

在前文中，我介紹了 TF-IDF（Term Frequency - Inverse Document Frequency）的概念，有了 TF-IDF 之後，就有了一種將文件表示成向量的不錯方式，而且其實現代搜尋引擎的索引結構，基本上都能提供 TF-IDF 的資訊，因為搜尋引擎在排列搜尋結果時，也會依據文件本身與搜尋字串的相似程度來做排序，而這個相似程度便是基於向量的相似性來計算得出的。

將不具特定結構的文字表示成為結構性的向量資料之後，文字探勘就得以套用資料探勘的相關技術來解決一些問題。

例如，如果我們想做文件的自動分類，像是將新聞自動的分成政治、社會、娛樂、運動、科技、……等等，就可以在得到每份文件的向量表示方式之後，套用一些在資料探勘領域大家都已經熟悉的分類演算法來進行自動分類。

現成的資料探勘演算法

在過去，這些資料探勘的演算法，很多都仰賴開發者自行撰寫實作。但慢慢的，也有愈來愈多的開放原始碼的程式庫釋出，大大節省了開發者的時間及心力。

其中，包含了從搜尋引擎到資料探勘相關演算法的開放原始碼專案，而在 Java 的開放原始碼社群中，累積了質與量兼優的相當產出。這使得 Java 的開發者想要開發相關的應用，有了許多現成的好輪子可用，省去了重新打造輪子的功夫。

對 Java 而言，想要建立運用搜尋引擎的應用，可以憑藉著 Lucene 這個開放原始碼的專案，來輕易完成。Lucene 是 Apache Software Foundation 下的一個專案，威力強大、效率高，而且具有很好的自訂及擴充的彈性，這使得它在這些年來，早就成為 Java 應用中最廣為使用的搜尋引擎。

除此之外，還有各式圍繞在 Lucene周遭，以 Lucene 為基礎的各個開放原始碼專案，來滿足打造搜尋引擎時的各種不同需求。包括了提供更進階搜尋功能以及分散式機制的 Solr、自動收集網站資料的 Nutch、……等等。

運用 Lucene 來索引文件的好處不僅止於此，我們想要利用 Lucene 來做文字探勘的應用也很方便，因為一旦利用 Lucene 來對文件進行索引之後，便可以利用 Lucene 的 API 來取得、計算出索引結構中每份文件以 TF-IDF 表示的向量。而一旦有了每個文件的向量表示式之後，我們便可以套用資料探勘領域中所運用的相關演算法，來對文件做處理。

就像前面所舉的例子，倘若我們想做文件的分類，我們可以先將文件送進 Lucene 以建立索引的結構，然後取出表示每份文件的向量（在 Lucene 中稱為 Term Vector）。

在取得 Term Vector 後，就可以將各文件的 Term Vector 做為分類演算法的輸入。經過分類演算法計算之後，便可以得到分類的結果了。從這個例子中，就不難理解 Lucene 所提供的文件 Term Vector 資訊，對於我們要做文字探勘的處理有多麼大的幫助了。

機器學習演算法相關的開源軟體專案

Apache Software Foundation 為此能提供的幫助，還不僅只如此。在文字探勘的應用裡，機器學習的演算法會經常被運用到，像是自動分類、或是自動群集，目前都常被歸類在機器學習的領域中。

在之前，我們就介紹過 Apache Software Foundation 旗下的開放原始碼專案 Mahout，它就是一個專注在機器學習領域的程式庫專案。在 Mahout 中，實作了諸如群集（clustering）、分類（classification），以及協同過濾（collaborative filtering）等等的核心演算法，而這些演算法正好是我們在開發文字探勘的應用時時常會需要使用到的。

Mahout 本身，亦基於 Apache Software Foundation 另一個名為 Hadoop 之專案的平臺。相信對雲端計算略有涉獵的讀者都知道，在Hadoop當中，提供了執行 map/reduce 演算法的計算環境，這使得於其上執行的程式，得以具備良好的規模可擴充性。因此，這邊的好消息是，因為 Mahout 中的機器學習演算法都有提供 map/reduce 的版本，能運行於 Hadoop 之上，這使得這些機器學習演算法的實作，有能力處理大量的資料規模，因為可以透過增加節點來提高計算能力。

當我們從 Lucene 的索引中取得文件的 Term Vector 之後，可以將它們餵給 Mahout 中的機器學習演算法，計算出我們想要的結果。如此一來，便可以將 Lucene 與 Mahout 整合起來，利用機器學習的演算法來處理大量的文件，從文件中挖掘出可能有用的資訊。

在Mahout裡面，除了提供 API 讓程式設計者可以直接控制之外，它也提供了現成的程式可以直接讀取 Lucene 的索引結構或是獨立的文字檔案，先將它們轉換成為 TF-IDF 的向量表示方式之後，再執行所指定的 Mahout 機器學習演算法，而演算法的執行可以是在多節點的 Hadoop 平臺之上。

善用已經發展出來的各種軟體技術，打造文字探勘應用已經比過去容易得多

至此不難發現，這幾個專案成果之間的整合非常的好。在過去，光是計算單一文件的Term Vector 就是一件花功夫的事，再加上所需的機器學習演算法，往往都可以耗去開發者相當多的心力和時間。如今，所有的輪子都具備了，而且還幫你上了雲端，等於可以不需要額外勞神於解決計算規模的問題之上。

大家都說現在是一個大數據（Big Data）的時代。因為，大數據的資料來源不僅僅只有具結構的資料，文字型、不具結構的資料在我們的生活中也到處都是。從大量的文字資料中，我們其實很有機會發展出各種有潛力的有趣應用，這正是文字探勘技術的目標。

就像每個人在 Facebook 上每天所發布的近況更新都有文字，而每個人在轉貼文章連結時也都有文字。當某個使用者更新或轉貼連結、甚至只是按讚時，他就和這些文字產生了關聯。這些文字某種程度上和他的個人偏好有所相關，我們可以從中分析出特定使用者和那些關鍵字詞有高度的相關性，從中探尋出該使用者的興趣或是偏好。

一旦有能力發現使用者的興趣或偏好之後，我們可以進一步發展很多應用。例如，我們可以利用文字分類技術將使用者自動分類、在社交網路服務上推薦興趣相近的使用者，甚至可以將這個資訊應用在廣告的投放上。

目前的網路廣告愈來愈講究投放的精準度，以便增加廣告的效果，其背後的中心思想無非希望所投放的廣告能夠更投使用者所好。因此，若是可以透過使用者在網路上的行為，使其和文字產生關聯，便可以進一步分析其偏好及興趣，也就有利於廣告精準度的提高了。

而在這 Big Data 的時代，面對的可能是更龐大的資料規模，開發者要解決的不只是怎麼從文字中找出有用資訊的議題，同時也要面對規模可擴充性的問題。

對於幸運的 Java 開發者而言，有龐大的 Apache Software Foundation 專案資源做靠山，可以很流暢地憑藉著本文中所介紹的幾個專案，一條龍式的解決開發文字探勘相關應用時的重要問題。這使得開發者可以更將心力放在應用本身，而不需要花費太多的心神在大家都用的到的輪子之上。

也期待有志於此的開發者，能夠在開放原始碼社群貢獻的強大支持下，開發出更有意思的文字探勘應用。

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（中）」

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（中）
- 程式人該關注的程式碼品質
- 物聯網概念下的程式設計
- 以雲為集散地，物聯網連接人與物
- 開發應用，API先行
- 犧牲的架構：為了砍掉重練的架構

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

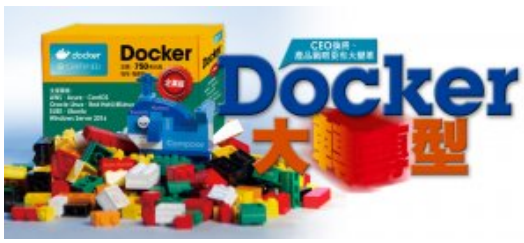
思科營收連六季負成長，準備再裁1100名員工

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