關聯法則探勘於中文斷詞及建立相似度索引典之應用

Chinese Word Segmentation and Similarity Thesaurus Construction by Association Rule Mining

梁哲瑋
交通大學
資訊科學研究所
is86011@cis.nctu.edu.tw

柯皓仁
交通大學
圖書館
claven@lib.nctu.edu.tw

楊維邦
交通大學
資訊科學研究所
wpyang@cis.nctu.edu.tw

摘要

關聯法則探勘（Association Rule Mining）已經廣泛應用在許多領域，用以找出物件之間的關聯性。本研究主要目標為應用關連法則探勘，不使用任何額外的資源，例如事先定義的字典或是詞鍵關聯網路等等，直接從語料庫中學習中文斷詞規則以及建立中文相似度索引典。透過不同的前置處理，包括不同交易長度、交易項目、探勘流程，可以使用關聯性法則探勘來解決資訊擷取（ Information Retrieval）系統中相似度或是關聯性的計算問題，以增進資訊擷取系統正確率。

關鍵詞：中文斷詞（Chinese Word Segmentation），查詢擴展（Query Expansion），關聯法則探勘（Association Rule Mining），資訊擷取（Information Retrieval），相似度索引典（Similarity Thesaurus）

一、緒論

中文語法上有意義的單元是詞，可能是單字詞、雙字詞或是更長的詞。例如想檢索「海關」相關的文件，若資訊擷取系統中使用單字「海」以及「關」作為索引，則可能檢索出「面海的房間」、 「關島」這類旅遊相關的文件。無法精確對語義建立索引。因此中文資訊擷取系統必須將詞視為語義的最小單位。

中文斷詞是已經研究多年的問題，最基本的方法可以使用字典比對，但是一般字典的辭彙有限，使用字典比對方式需要大量人力建立字典，並且會遇到未知詞、新詞、專業術語等等無法判別的問題，僅僅使用字典比對並無法檢索到精確的結果。另外，使用者下達查詢時所使用的詞鍵與文件中所使用的詞鍵往往不盡相同，而且詞鍵有同義詞以及近似詞，例如使用者下達查詢「SARS」，但是文件的用語是「嚴重急性呼吸道症候群」，或是「非典型肺炎」，也無法檢索到精確的結果。

二、相關研究

（一）中文斷詞相關研究

中文斷詞的研究大致有語料庫統計（Statistical Analysis）、字典為本的比對方法（Dictionary-based Ap-

proaches）、句型分析（Syntactic Analysis）、概念為本（Concept-based Approaches）等等方法。本研究採用關聯法則探勘，類似於語料庫統計的方法。簡單的統計方法使用交互資訊(Mutual Information)[5] 來偵測詞鍵邊界，單字t_1 、t_2交互資訊定義為

$$MI(t_1,t_2) = \log_2 \frac{freq(t_1t_2)}{freq(t_1) * freq(t_2)}$$

當兩個相鄰單字(Bigram)之交互資訊強度大於其他相鄰單字的交互資訊，則這兩個單字可以組成一個詞鍵。這個方法考慮了t_1和t_2在所有文件中共同出現的頻率以及個別出現的頻率。Dai [4] 使用語境資訊（Contextual Information）的方法，計算方式類似交互資訊，但是權重的計算多考慮了單字的位置、單字附近的語境資訊等等。

（二）建立相似度索引典與關研究

相似度索引典（Similarity Thesaurus）[1]是紀錄詞鍵和詞鍵關係的矩陣，和共同出現矩陣(Co-occurrence Matrix)不同。共同出現矩陣用來表示任何兩個詞鍵在所有文件中共同出現的頻率，相似索引典則表示詞鍵和詞鍵在概念空間上的距離。 Qiu [1]用文件來作為概念空間的索引，類似TF-IDF方法，TF-IDF利用詞鍵出現頻率(Term Frequency, TF)和逆向文件頻率(Inverse Document Frequency, IDF)來表示文件向量。而[1]則是使用詞鍵出現頻率(TF)和逆向詞鍵頻率(Inverse Term Frequency, ITF)來表示概念空間上的詞鍵向量，計算概念空間中每個詞鍵的距離可以建構出相似度索引典。

三.研究方法

（一）關聯法則探勘

關聯法則探勘 [3]主要是針對交易(Transaction)型態的資料庫加以分析，以便找出隱藏其中的資訊。例如一個超級市場中，顧客在每筆交易所購買商品種類而形成的資料庫。關聯法則探勘就是要從該資料庫中找出顧客的購買行為模式，像是顧客買了麵包之後, 其中30%會同時去買牛奶的行為，就是一個關聯法則。

Agrawal[2] 等人在1994 年提出Apriori 演算法，用以解決關聯法則探勘的問題。Apriori方法如表1 所示，這個方法的主要概念是重複讀取資料庫，並且在每次讀取資料庫後產生長度相同的強項目集合（Large Itemsets）。另外只針對候選項目集合（Candidate Itemsets），而非所有可能的項目集合（Itemsets）來作支持度(Support)的計算，以減少計算時間。

$L_1 = \{\text{large 1-itemsets}\}$
for ($K = 2; L_{K-1} \neq 0; K++$) do begin
  $C_k = apriori - gen(L_{k-1})$
  forall transactions $t \in D$ do begin
    $C_t = subset(C_K, t)$
    forall candidates $c \in C_1$ *do*
      c.count++;
  end
  $L_K = \{c \in C_K \mid c.count \geq \min sup\}$
end
$Answer = \cup_k L_k$

表 1 Agrawal[2] 提出的 Apriori 演算法

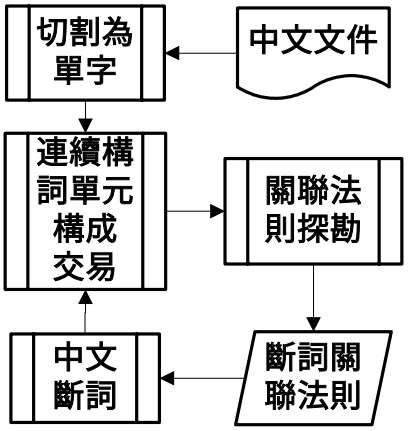
資料探勘系統（Data Mining）和關聯式資料庫（Relational Database）

可以使用不同的耦合方式，本論文考量實作方便，採用緊密耦合（Tightly Coupling）模型。採取 Sarawagi [6]等人所提出的方式，使用 SQL-92 語法實作 Apriori 演算法。

（二）中文斷詞

組成中文詞的每個單字之間會有關聯，根據這個觀察，本研究假設相鄰而且關聯性強的字可以組成中文詞，因此中文斷詞問題可以視為關聯法則探勘問題。

將文件中固定長度的連續中文字視為交易，每個中文單字則為交易項目，過程中判斷出來的中文詞也視為一交易項目。流程如圖一所示



圖一 中文斷詞關聯法則探勘

1. 將語料庫中的文件分割為中文單字。
2. 固定長度[1]的連續中文單字或是詞組成一筆交易，使用雙連字串模型〈Bigram Model〉右移一個單字產生下一筆交易。例如當交易長度固定為3時，則 “設立專責醫院” 可以如表2 所示分割成四筆交易。本論文考慮的詞鍵組合方式包括相鄰的中文單字，例如「醫」和「院」；相鄰的英文字母與數字，例如本論文採用的資料集常出現「N95口罩」，其中「N」與「9」與「5」便可組成詞鍵。由於相鄰的中文字、英文字母與數字雖然可以構成詞，但是本論文使用3.3節所述的相似索引典來處理這類關係，所以斷詞系統不考慮相鄰的中英文單字。本例中，「N95」和「口罩」構成兩個詞鍵，而非一個詞鍵，但是相似索引典中紀錄「N95」和「口罩」有極強的關聯性。
3. 使用SQL 92實作Apriori 演算法，找出單字間的關聯法則。
4. 產生斷詞規則庫，關聯法則探勘會找出所有強項目集合，因為長度為2的強項集最為完整，因此我們使用長度為2的強目項集合所產生的關聯法則作為斷詞規則。
5. 使用斷詞規則庫將原始文件的單一中文字組成中文詞。
6. 回到步驟2，直到產生所需長度的斷詞關聯規則為止。

	一	二	三
1	設	立	專
2	立	專	責
3	專	責	醫
4	責	醫	院

表 2 「設立專責醫院」 在交易長度為 3 的表示法，每列為一筆交易，欄為交易項目。

使用關聯法則探勘可以找出單字的關聯。強項目集合的長度及個數和

[1] 此處所指的長度係指交易項目的個數，每個中文單字或探勘過程中產生的中文詞均視為一個交易項目。

交易長度有關，由於長度為2的強項目集合較為完整，所以使用長度為2的強項目集合產生關聯法則，作為斷詞規則。第一次執行這個程序可以產生二字詞的斷詞規則，經過二字詞斷詞，第二次執行關聯法則探勘時長度為2的強項目集合可能包含單字以及二字詞，所以可以由單字和單字組成二字詞、單字和二字詞組成三字詞，以及二字詞和二字詞組成四字詞。執行的次數依照所需產生的詞鍵長度而定。

產生斷詞規則

由於關聯法則並非對稱，也就是 ”醫→院” 和 ”院→醫” 是不同的法則，有不同的支持度（Support）和信心度（Confidence）。因為”醫院”和”院醫”兩個組合方式構成詞的可能性不相同，所以權重的定義是非對稱性。相鄰的單字或詞W_1,W_2構詞的權重如(1)，定義為該關聯法則的支持度 $S(R_{w1\to w2})$ 和信心度 $C(R_{w1\to w2})$ 的權重和。

$$W_{w1,w2} = \alpha \bullet S(R_{w1\to w2}) + (1-\alpha) \bullet C(R_{w1\to w2}) \quad (1)$$

例如「醫→院」規則中，支持度定義為「醫」這個字在所有交易中出現的比例，信心度定義為所有出現「醫」字的交易中，「醫」、「院」兩字一起出現的比例。

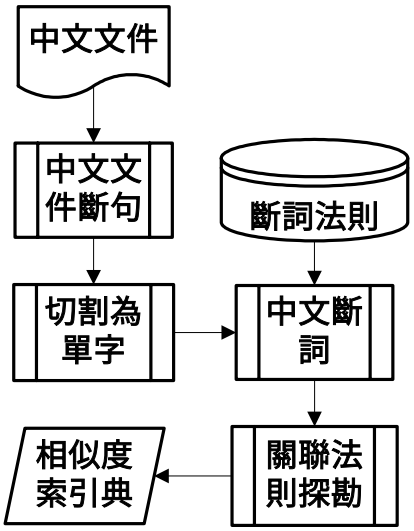
若規則庫中無W_1、W_2組成的規則，則權重為零。若大於零，則W_1、W_2可以構成新詞。由於Apriori 演算法會刪除小於最小支持度的項目，所以最小支持度可以視為門檻值（Threshold），在此不必額外設定門檻值。

規則	支持度	信心度
醫→院	0.084	0.459
感→染	0.031	0.473
疫→情	0.028	0.582
口→罩	0.008	0.593
海→關	0.001	0.667

表3 長度為 2的強項目集合產生的關聯法則實例

（三）相似度索引典

相似度索引典用來表示詞鍵之間的相似程度如圖三所示，並且計算權重。我們使用關聯法則探勘來自動建立相似度索引典。由於英文詞與中文詞也會有關聯，例如”SARS”和”嚴重急性呼吸道症候群”相關，所以相似度索引典的建立必須考慮英文詞。出現在同一句子的詞鍵通常是相關的，根據這個觀察，我們將一個句子視為一筆交易，構成該句子的單字和詞鍵就是交易的項目。如表4中「SARS」和「專責醫院」共同出現在一個句子中，也就是這兩個項目是在同一筆交易。找出交易項目間的關聯法則，就可以找出「SARS」和「專責醫院」之間的關係，再定義關係的權重，就可建立相似度索引典。流程如圖二所示。



圖二 建立相似度索引典

1. 將語料庫中的文件分割為中文單字以及英文單字。
2. 使用2.2節產生的斷詞規則庫將符合規則的中文單字合併為有意義的詞。
3. 依照標點符號，將中英文單字和詞組成句子，一個句子視為一筆交易，交易的長度不固定。由於句子的邊界定義明確，不會有3.2節中的斷詞邊界問題，所以不需要使用N-相連字串模型〈N-gram Model〉。表4為交易的實例。
4. 使用Apriori 演算法作關聯法則探勘，找出詞鍵之間的關聯
5. 計算詞鍵關聯權重，建立相似度索引典。

SARS 專責醫院 全國 設立 十 家

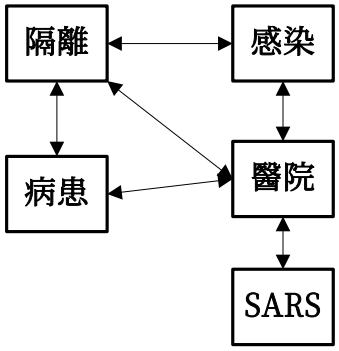
表4 「SARS專責醫院全國設立十家」的交易組成方式

相似度索引典權重

相似索引典中，詞鍵 t1與詞鍵 t2 的關聯權重$W_{t1,t2}$定義為關聯規則$R_{t1\to t2}$以及關聯規則$R_{t2\to t1}$的支持度S和信心度C加權總和：

$$W_{t1,t2} = \alpha \bullet S(R_{t1\to t2}) + (1-\alpha) \bullet C(R_{t1\to t2}) + \beta \bullet S(R_{t2\to t1}) + (1-\beta) \bullet C(R_{t2\to t1}) \quad (2)$$

由於相似度索引典是對稱性矩陣，也就是詞鍵t1和詞鍵t2的關聯度等於詞鍵t2和詞鍵t1的關聯度。所以將權重定義為關聯法則t1→t2 以及關聯法則t2→t1加權支持度和信心度的和。



圖三 相似索引典實例

查詢自動擴展

相似度索引典是資訊擷取系統中寶貴的資源，在資訊擷取系統的許多問題都可以使用，例如文件分類（Text Categorization），自動查詢擴展（Automatic Query Expansion）等等。本論文將相似度索引典應用於自動查詢擴展，使用以概念為本的自動查詢擴展[1]，將相似度索引典應用在資訊擷取系統上。

如 2.2 節所述，每個詞鍵可以視為是在概念空間中的概念，每個詞鍵向量 k_v 和查詢向量 q 在概念空間上的相似程度可以定義為：

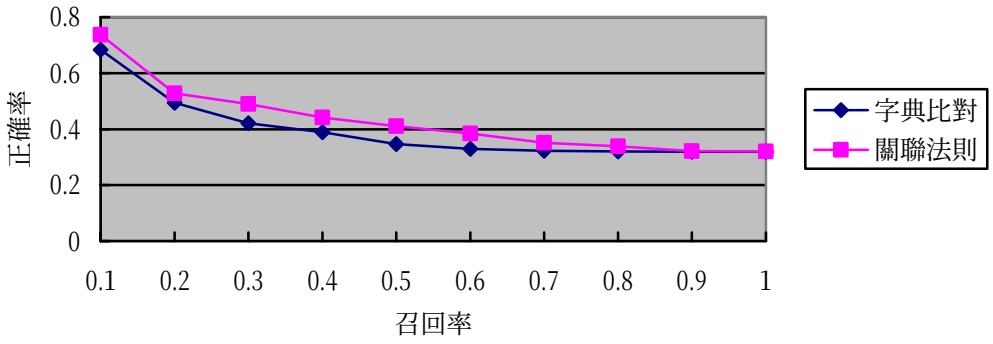


圖 4 在資訊擷取系統中使用關聯法則和一般性字典的效能比較

$$sim(q,k_v) = \vec{q} \bullet \vec{k_v} = \sum_{k_u \in Q} w_{u,q} \times c_{u,v} \qquad (3)$$

其中 $c_{u,v}$ 為相似度索引典中詞鍵 u,v 的關聯權重，若相似度索引典中沒有定義 u,v 的權重，則設為 0。$W_{u,q}$ 為查詢 q 中包含的某個詞鍵 u 在查詢中的權重。所以使用概念空間中的某個詞鍵 v 和查詢中的每個詞鍵，以及相似度索引典中 u，v 的關聯程度可以計算出某個詞鍵 v 和查詢 q 的相似程度。

自動查詢擴展是找出和查詢 q 最相似的 n 個詞鍵作擴展，重新調整該 n 個詞鍵在查詢中的權重 (Re-weighting) 。調整後的查詢 q'中詞鍵 v 的權重定義成：

$$w_{v,q'} = \frac{sim(q,k_v)}{\sum_{k_u \in q} w_{u,q}} \qquad (4)$$

四、實驗與討論

本論文使用中時電子報”SARS 戰疫前線” (http://forums.chinatimes.com/report /SARS/) 專題報導的部分文章作為資料集，總共收錄 274 篇。其中 25%的文章拿來作為查詢（Query），其他 75%作為關聯法則探勘的交易資料庫，也就是訓練集（Training Set）。實驗方式為人工針對每個查詢，勾選出相關的文件，作為評估標準。將這些查詢輸入本系統，檢索結果和人工訂定的評估標準作比對，算出正確率(Precision)和召回率(Recall)。

（一）斷詞規則庫評估

使用常用詞頻統計作為比對基準，常用詞頻統計是從一般文章中統計出經常共同出現的單字作為中文詞鍵，所包含的詞鍵屬於一般用詞。斷詞規則庫使用關聯法則探勘，從資料集中學習斷詞規則，可以學習出特定領域的用語。圖 4 的實驗結果可以看出關聯法則產生的斷詞規則庫效能比一般字典佳。

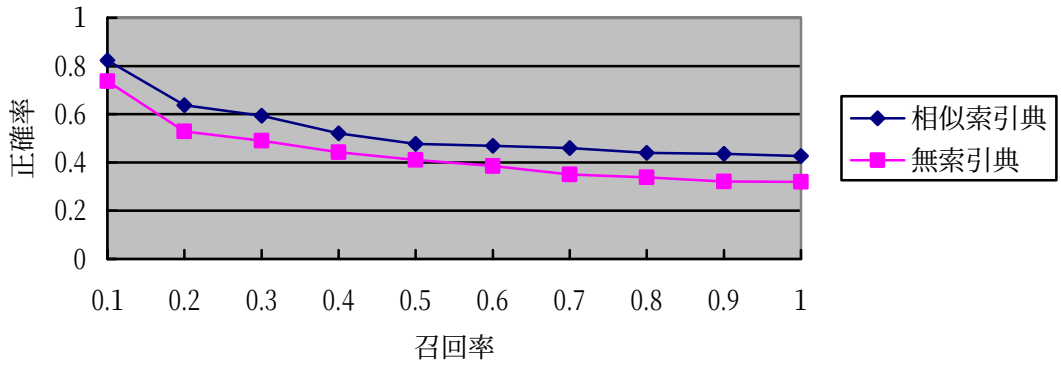


圖 5 加入相似索引典對資訊擷取系統的影響

（二）相似度索引典評估

相似度索引典紀錄詞鍵之間的相關程度，使用 3.3 節所述的查詢擴展方法作自動查詢擴展來評估相似索引典對資訊擷取系統的影響。資訊擷取系統中使用關聯規則方法作中文斷詞，圖 5 的實驗結果顯示加入相似索引典作自動查詢擴展以後，資訊擷取系統的效能會提升。表 5 是相似索引典的部分結果，

詞鍵 1	詞鍵 2	權重
SARS	醫院	1.032
SARS	專責醫院	0.976
N95	口罩	0.871
海關	口罩	0.462
衛生署	疾病管制局	0.455
醫院	耳溫	0.436

表 5 相似索引典部分結果，權重=0.6 支持度+0.4*信心度

實驗中可以發現信心度對權重的影響比支持度還大，尤其在出現次數較少，但是出現模式固定的詞鍵上。例如「疾病管制局→衛生署」這條法則的支持度為「疾病管制局」一詞在所有句子中出現的比例，只有 0.001。而信心度為「疾病管制局」出現的句子中，「衛生署」也一併出現的比例，高達 0.778。而相似度索引典的權重是對稱性，本論文將權重定義成「疾病管制局→衛生署」和「衛生署→疾病管制局」兩條法則權重的總合。其中「疾病管制局→衛生署」的權重比「衛生署→疾病管制局」大，因為前者的信心度高於後者。

五、結論

本研究提出了將資訊擷取的問題，包括中文斷詞問題以及詞鍵關聯問題轉化成資料探勘問題的方法。實驗得知，應用在資訊擷取系統中，關聯法則斷詞的結果較字典比對的方法準確度為高；關聯法則建立的相似度索引典，透過查詢自動擴展的機制，也可以大幅提升資訊擷取系統的準確度。

本研究的兩個問題根據不同的假設，使用不同的前置處理流程，和不同的參數設定，但是執行相同的關聯法則探勘。斷詞問題使用固定的交易長度，詞鍵關聯問題使用句子或是句子片段作為交易。本研究提出的方法也可以根據問題特性做假設，透過不同的前置處理和參數設定，推廣來解決其他資訊擷取系統中需要做關聯程度和相似度運算的問題。

六、參考文獻

[1] H. P. Frei and Y. Qiu, “Concept based query expansion” in *Porc. ACM SIGIR93*, pp.160- 169, 1993.

[2] R. Agrawal and R. Srikant, ”Fast Algorithms for Mining Association Rules.” In *Proc. 20th International Conference on Very Large Data Bases* , pp. 487 - 499,1994.

[3] A. Swami, R. Agrawal and T. Imielinski, “Mining association rules between sets of items in large databases” in *Proc. ACM SIGMOD93* pp.207-216, 1993

[4] C. S. G. Khoo , Y. Dai and T.E. Loh, “A new statistical formula for Chinese text segmentation incorporating contextual information” in *Proc. ACM SIGIR99*, pp. 82-89, 1999

[5] K.T. Lua, “Experiments on the use of bigram mutual information in Chinese natural language processing” , in *Proc. Interational Conference on Computer Processing of Oriental Languages*,1995

[6] R.Agrawal, S.Sarawagi and S.Thomas, “Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications.” in *Proc. ACM SIGMOD98,* pp.343-354 , 1998