中文
成大研發快訊　第六卷 第十期 - 2008年十二月十二日
[ http://research.ncku.edu.tw/re/articles/c/20081212/1.html ]

# 應用超空間模擬語言模型與演化式推論演算法於網路精神科文件語意樣式萃取之研究

禹良治[1]、吳宗憲[1,*]、葉瑞峰[2]、張鳳麟[3]

[1]國立成功大學電機資訊學院資訊工程學系
[2]國立嘉義大學資訊工程學系
[3]奇美醫院精神科
chwu@csie.ncku.edu.tw

IEEE Trans. Evolutionary Computation, Vol. 12, No. 2, pp. 160-170, April 2008.

人們在日常生活中，可能會遭遇某些負面生活事件(Negative Life Event)，例如：感情問題、課業壓力、家庭問題等，並因此引發憂鬱情緒、焦慮、自殺意念等憂鬱症狀(Depressive Symptom)。當人們遭遇憂鬱問題時，往往因為擔心前往精神科看診會被貼上精神疾病的標籤，而轉向網際網路尋求協助。目前已有許多心理衛生網站，如：心靈園地(http://www.psychpark.org)、董氏基金會(http://www.jtf.org.tw)等，提供網路留言版、討論區與電子郵件諮詢等服務，提供網友上網撰文抒發情緒的困擾，並等待專家回覆建議事項。然而這些網路精神科文章數量龐大，以致於往往無法即時回覆，亦造成人工處理的負擔；如果電腦能夠理解文章中的負面生活事件，便可事先根據事件種類進行文章分類，加速線上諮詢流程，甚至可進一步提供文章檢索與自動諮詢等服務。因此，本研究之目的即在探討如何自動從網路精神科文件中找出負面生活事件的重要特徵。

負面生活事件在文字表現上的主要特徵為不同長度的字詞組合，例如：”兩年前，我失去了雙親”代表一個家庭類的負面生活事件，其中字詞組合<失去，雙親>即為重要特徵；同理，”夫妻經常為錢的事情吵架”便可以<夫妻，吵架，錢>來表示。在本篇文章中我們將不同長度的字詞組合稱為語意樣式(Semantic Pattern)，而這些語意樣式正是判斷負面生活事件的重要特徵。因此，本研究提出演化式文本探勘(Evolutionary Text-Mining)架構，目的在從未標記的網路精神科文件中自動擷取可變長度之語意樣式；此架構主要可分成兩部分：超空間模擬語言(Hyperspace Analog to Language, HAL)模型及演化式推論演算法(Evolutionary Inference Algorithm, EIA)，如圖1所示。

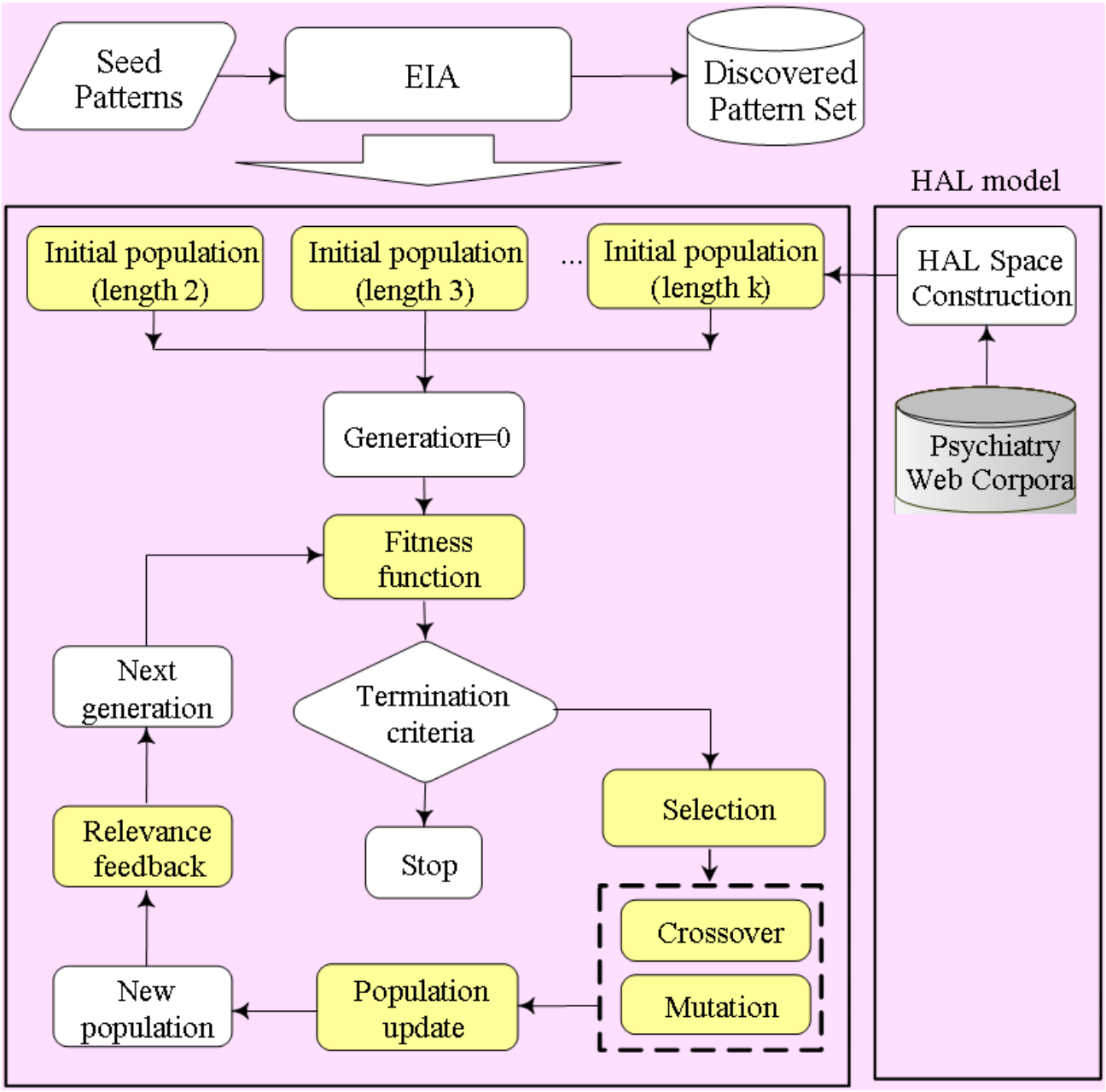


圖1、 語意樣式擷取流程圖

超空間模擬語言模型是一個用來表示文字語意的模型，其係以高維的空間記錄文字的前後文脈資訊(Context Information)，並以此模擬人類認知語言的能力，這是因為當人們遇到不認識的字時，往往會由其附近的文字來推論此字的意義。因此，超空間模擬語言模型在文字意義的表示上是利用一個高維度的向量空間來表示文字與其前後文脈之關係(如圖2所示)，其中每一個向量代表一個字詞，而向量維度代表其前後文脈的字詞分布。因此，當兩個字詞有相似的前後文脈時，其意義較為相近，而此文脈資訊亦可從大量文章資料統計得到。

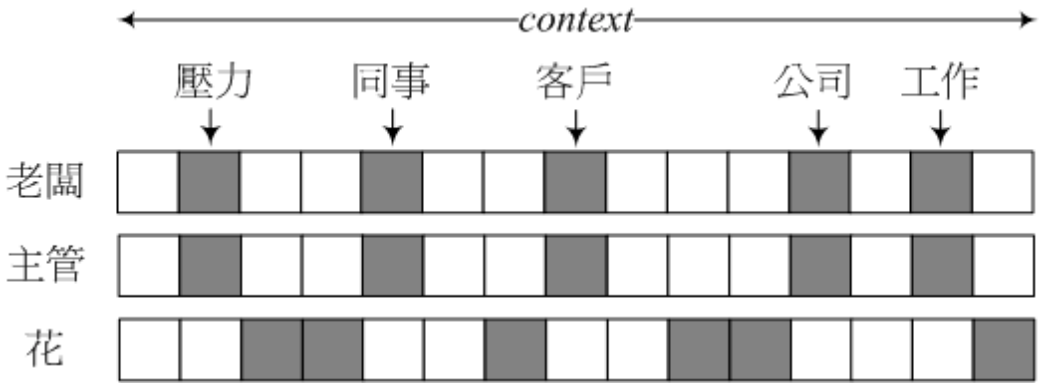


圖2、 超空間模擬語言模型示意圖

圖2顯示”老闆”、”主管”與”花”三個字的前後文脈分布情形，可以想像的是”壓力”、”同事”、”客戶”等字較常出現在”老闆”及”主管”的前後文中，與”花”的前後文大不相同，因此可判斷”老闆”及”主管”在語意上較為相近。本研究中，我們使用超空間模擬語言模型來表示語意樣式，並藉由比對語意樣式的前後文脈分布來計算其相似度。

演化式推論演算法係將超空間模擬語言模型整合至傳統的演化式計算中，使其能夠計算語意樣式之相似度，並逐步從網路精神科文件中推論出更多相似的語意樣式，其做法說明如下。首先，給定一組種子樣式(Seed Pattern)，並且每個種子樣式皆以超空間模擬語言模型來表示；為了推論出不同長度的語意樣式，當種子樣式輸入後，演化式推論演算法將隨機產生長度為2到k的初始族群(Initial Population)做為候選者，同樣地，族群中的樣式也以超空間模擬語言模型來表示，如此，適應函數(Fitness Function)便可根據種子樣式與語意樣式的文脈分布計算適應值，適應值較高代表兩者較為相似；接著根據適應值選取(Selection)部分語意樣式做為親代(Parents)進行交配(Crossover)與突變(Mutation)，經此過程後將產生一組新的語意樣式，即所謂的子代(Offspring)。此時，子代也將經過適應性評估，並且只有當子代的適應值高於親代時才會取代親代，經過族群更新後(Population Update)，新的族群也同時產生。此時，相關回饋(Relevance Feedback)挑選出相關的語意樣式，並根據相關樣式的文脈資訊調整種子樣式的文脈分佈，使其更相似於相關的語意樣式，以便在下一世代(Generation)擷取出更多相關的語意樣式。表一即為部份部分種子樣式及推論所得之語意樣式

表一 部分種子樣式及推論所得之語意樣式

| Types | Seed Pattern | Pattern Induction from Web |
|---|---|---|
| 家庭(Family) | <小孩,受傷>, <先生,吵架> | <先生,爭吵><br><先生,吼><br><老婆,爭論><br><夫妻,衝突><br><先生,爭吵,錢><br><太太,爭論,錢> |
| 感情(Love) | <婚姻,破裂>, <老婆,外遇> | |
| 學校(School) | <老師,責罵>, <考試,失敗> | |
| 工作(Work) | <薪水,減少>, <工作,停止> | |
| 社交(Social) | <朋友,過世> | |

在實驗評估上，本研究以心靈園地及董氏基金會心理衛生特區等網頁資料共5,000份文件做為實驗語料；比較對象為資料探勘領域常用的Apriori演算法，接著分別執行演化式推論與Apriori演算法找出語意樣式，並針對兩種方法所得的語意樣式進行評估；首先，我們邀請15位參與者寫出生活中不愉快的經驗，並依此

收集69句包含貝面生活事件的句子做為測試資料，接者以未知樣式率(Out-of-Pattern rate, OOP rate)評估兩種方法的推論能力。所謂未知樣式率是指測試句所含的語意樣式沒有被演算法找到的比例。表二即為演化式推論演算法(EIA)及Apriori的未知樣式率比較，其中*_*_Multiple與*_*_Worst分別代表執行演化式推論演算法30次與其中最差一次的結果，*_RF_* 代表有使用相關性回饋的結果，而#代表使用符號檢定(Sign Test)評估兩未知樣式率之差異為顯著。

表二 未知樣式率比較

| | OOP Rate | Reduction (%) | | | |
|---|---|---|---|---|---|
| | | over Apriori | over EIA_Worst | over EIA_Multiple | over EIA_RF_Worst |
| Apriori | 60.9% (42/69) | — | — | — | — |
| EIA_Worst | 97.1% (67/69) | +59.4 | — | — | — |
| EIA_Multiple | 42.0% (29/69) | -31.0# | -56.7# | — | — |
| EIA_RF_Worst | 40.6% (28/69) | -33.3# | -58.2# | -3.3 | — |
| EIA_RF_Multiple | 27.5% (19/69) | -54.8# | -71.7# | -34.5# | -32.3# |

# performance difference is statistically significant (p<0.05)

實驗結果顯示演化式推論演算法比起Apriori有較佳的推論能力(EIA_Worst除外)，主要原因包括：演化式推論演算法結合超空間模擬語言模型與相關性回饋，並且可重複執行多次；超空間模擬語言提供有用的前後文脈資訊使演化式推論演算法能藉由給定的種子樣式自動從網路文件中推論出更多相似的語意樣式，相關性回饋則可確保演化方向正確以增進其推論能力，此外，藉由重複執行演化式推論演算法可同時增加語意樣式的數量與多元性。更重要的是，演化式推論演算法僅需少量的標記資料(即種子樣式)便可直接利用網路上龐大且未標記的資料進行推論，比起傳統必須仰賴大量標記資料的方法，更適合在網路環境下使用。