

Concept Cluster Based News Document Summarization

Student: Cheng-Chang Liu

Advisor: Dr. Hao-Ren Ke, Dr. Wei-Pang Yan.

Institute of Computer Information and Science

Nation Chaio Tung University

ABSTRACT

A multi-document summarization system can reduce the time for a user to read a large number of documents. A summarization system, in general, selects salient features from one (or many) document(s) to compose a summarization, in the hope that the generated summarization can help a user understand the meaning of the document(s).

This thesis proposes a method to analyze the semantics of news documents. The method is divided into two phases. The first phase attempts to discover the subtle topics called concepts hidden in documents. Due to the phenomenon that similar nouns, verbs, and adjectives usually co-occur with the same representative term, we describe a concept by those terms around it, and use a semantic network to assist the description of a concept more accurately.

The second phase distinguishes the concepts discovered in the first phase by their word senses. The K-means clustering algorithm is exploited to gather concepts with the same sense into the same cluster. Clustering can diminish the problem about word sense ambiguity and reduce concepts with similar sense.

After the two above phase, we choose five features to weight sentences and order sentences according to their weights. The five features are lengths of clusters,

location of a sentence, $tf*idf$, distance between a sentence and the center of the cluster to which the sentence belongs, and the similarity between a sentence and the cluster to which the sentence belongs.

We use the news documents of Document Understanding Conferences 2003 (DUC2003) and its evaluation tool to evaluate the performance of our method.

keywords : English news summary, feature selection, concept clustering



以概念分群為基礎之新聞文件自動摘要系統

研究生：劉政璋

指導教授：柯皓仁博士，楊維邦博士

國立交通大學資訊科學研究所

摘要

多文件摘要系統可以有效節省在閱讀大量文件時所耗費的時間。一般摘要系統會從文件中挑選出具有較多語意資訊的句子構成摘要，以期從摘要中即可瞭解文件所涵蓋的概念，幫助讀者更快速瞭解文件內容。而多文件摘要更需要注重刪除重複的資訊。

本研究透過分析文件內容語意，藉以幫助挑選文件中具有較多語意資訊的句子。分析過程可分成兩大步驟：1) 找出隱藏在文件中的重要概念：多文件中會包含許多較小的主題，稱之為概念；我們使用描述鍵結去敘述一個概念，由於一個有代表性的詞彙，常伴隨著類似的名詞、形容詞、動詞等一起出現，因此我們利用這些出現在代表性詞彙周圍的詞來描述該詞彙，並加入語意網路，來加強描述的準確性。2) 分析內容語意要先分辨出哪些概念是相同或是不同的，亦即語意歧異解析，並將相同的概念分在同一群之中；我們使用 K-Means 分群法，將前一步驟找出的概念加以分群，以解決語意歧異的問題，並去掉重複的概念。確定文件內容語意之後，根據概念的分群結果、句子的資訊含量、句子在文章中的位置等不同的特徵選擇出最能代表文件集的句子。

在實驗中，使用 DUC2003 提供的新聞文件，以及評估軟體，評估軟體是比較系統自動產生的摘要與專家寫作的摘要之間的相似度，系統將會利用該評估程式來提供一個客觀的數據。

關鍵字：英文新聞文件摘要、特徵選擇、概念分群

致謝

首先感謝指導教授柯皓仁老師及楊維邦老師的悉心指導，讓我學習到完成一篇論文或一項研究所需經歷的整個過程與自我挑戰的階段，也讓我了解到作為一個研究生所須具備的實事求是與追根究底的精神。

感謝實驗室的伙伴們，由於你們對我的關懷與照顧，讓我的研究生活變成一種樂趣，適時溫馨的幫助與實驗室和樂的氣氛是使我完成論文的一大動力，在那種環境下，具有讓人放鬆心情及專心思考的魔力。

感謝親愛的家人永遠不變的支持與鼓勵，也感謝朋友的關懷。在研究所的這兩年當中，有摸索、有茫然、有討論、有歡笑、有瓶頸、也有解決方法，這段過程是獨一無二且滿懷回憶的。謝謝你們。



June, 2005!

目錄

英文摘要	I
中文摘要	III
致謝	IV
圖目錄	VI
表目錄	VII
方程式目錄	VIII
第一章 緒論	1
第一節 自動化摘要系統	1
第二節 研究動機	2
第三節 研究目的	3
第四節 論文架構	3
第二章 相關研究工作	5
第一節 文件摘要相關研究	5
第二節 以摘錄方式為基礎之文件摘要	6
第三節 以句子特徵選擇為基礎的文件摘要	9
第三章 改良型概念描述法	13
第一節 前置處理	13
第二節 前後文描述概念	20
第三節 加入語意網路描述概念	24
第四節 概念以及句子分群	29
第四章 語句語意權重摘要	39
第一節 句子的特徵	39
第二節 叢集的特徵	42
第三節 句子挑選	43
第五章 實驗結果分析與評估	45
第一節 實驗步驟	46
第二節 實驗結果	47
第六章 結論與未來研究方向	56
第一節 結論	56
第二節 未來研究方向	57
參考文獻	58

圖目錄

圖 1 相關研究工作.....	6
圖 2 W. LAM 提出的財經新聞樣版[9].....	8
圖 3 DUC2003 原文範例.....	14
圖 4 經過 NLP PROCESSOR 處理完的 POS 標籤.....	15
圖 5 片語化之結果範例.....	18
圖 6 轉小寫之結果範例.....	19
圖 7 SEMANTIC NETWORK 共現矩陣範例[5].....	25
圖 8 INFOMAP 的語意網路範例[5].....	26
圖 9 K-MEANS 演算法[11].....	31
圖 10 句子對應到叢集.....	35
圖 11 句子對分群結果的對應情形 1.....	36
圖 12 句子對分群結果的對應情形 2.....	37
圖 13 叢集特徵圖例.....	43
圖 14 權重比例調整 1.....	47
圖 15 權重比例調整 2.....	48
圖 16 調整向量長度變數.....	49
圖 17 調整 TF*IDF 門檻變數.....	50
圖 18 調整分群數量變數.....	51
圖 19 調整加入語意網路變數.....	51
圖 20 兩種加入語意網路方法的比較.....	52
圖 21 兩種句子對應法的比較.....	53

表目錄

表 1 詞性標記.....	17
表 2 字彙經過 STEMMING 之後的變化.....	17
表 3 停用字的部分列表.....	20
表 4 TF*IDF 例子.....	21
表 5 前後文描述概念範例.....	24
表 6 加入語意網路方法一的範例.....	28
表 7 加入語意網路方法 2 的範例.....	29
表 8 概念分群範例.....	33
表 9 Tf-IDF 範例.....	40
表 10 預計的實驗變數說明.....	46
表 11 調整權重比例 ROUGE-1 數值，從第二變數開始調整.....	48
表 12 調整權重比例 ROUGE-1 數值，從第四個變數開始調整.....	49
表 13 ROUGE 分數比較(部分數值取自 DUC[1]).....	53
表 14 DUC2003 全部類別評估數值.....	55



方程式目錄

方程式 1 貝氏定理應用在特徵挑選上[12].....	10
方程式 2 TF*IDF 公式	21
方程式 3 平方誤差準則	31
方程式 4 句子對應到叢集的方式	34
方程式 5 相似度特徵計算	41
方程式 6 相似度特徵計算 2.....	41



第一章 緒論

第一節 自動化摘要系統

電腦科技的迅速發展是大家有目共睹的，加上近幾年來網路的應用越來越廣泛，使得許多資訊陸續進行數位化，以利在網路上傳播。不過隨著數位化的發展，也使得資訊大量增加，使用者在獲取資訊上不再像之前困難獲取管道稀少，反而是輕易的獲取大量資料。在這種現象之下，困難的反而是如何過濾掉不需要的或者是重複的資訊，使得使用者可以找到真正所需要的資訊。

對於資訊量過多的問題，有學者提出了幾種方法來幫助使用者找到所需的資訊。其中一種方法是先把相似的東西群聚起來，再將群聚完的資料進行諸如文字過濾(Text Filtering)等之處理，把使用者真正需要的資料篩選出來[9]。

文件自動摘要是眾多文字探勘技術(Text Mining)的重要項目之一，其目的在於從指定的文件集中取出滿足使用者需要的摘要。摘要的目的在於減少閱讀原始文件的時間，但是卻能夠讓使用者瞭解到原始文件的主要意涵[10]。讀完摘要的使用者應該能夠回答文件中與主題相關的問題，或者是能夠進行一些有關該文件集的工作。

文件摘要依照原始文件集的數量多寡，可分為單文件摘要與多文件摘要。單文件摘要把單篇文件的內容精簡化與重點化，注重的是能否有效地刪除沒有必要的資訊，並留下真正能代表文件內涵的資料；多文件摘要則是把多篇相同主題的文件融合在一起，除了刪除無用的資料外，尚需有效率地過濾重複在多篇文章中所出現的資訊[20]。

本論文中提出了以特徵選擇法(Feature Selection Approach)選擇出含有最適當語意的句子當作摘要。Chen[12]提出以挑選句子為摘要的多個理由：

(1) 可以反應文件的主題

在效果評估上看得出來的確可以選出具有豐富語意的關鍵字，尤其對於內容寫法一般性的新聞文件，更為明顯。

(2) 快速


不需要做語意訓練，也不需要太多該領域的專業知識。直接用特徵選取的速度是十分快的。

(3) 方便控制摘要長度

藉由挑選句子的多寡來決定整篇文件的壓縮度，可以從零到選取整篇文章。

(4) 可設定重要字彙

使用者可以事前設定一些關鍵字，例如使用者比較有興趣的幾個主題或是概念，設定完之後可以在挑選的時候，加重這些字彙的權重，優先挑出有包含這些關鍵字的句子，更能符合使用者的需求。



本論文以新聞文件為所要摘要的原始文件，新聞文件有下列幾項特點。第一，用詞比較一般化，專有名詞較少。第二，由於讀者為一般大眾，因此內容不需專業背景知識即可吸收。第三，容易找到相關主題的報導，同一事件會有多家媒體報導。針對這三個特點使用統計方式的摘要方法，將可以快速產生具有足夠精確度的多文件摘要。

第二節 研究動機

目前摘要系統的作法大致上可以分成兩類：第一類需要使用專業領域知識 (Domain Knowledge)，因此必須倚賴領域專家建立才能使用；第二類無須依靠領域知識，使用一些統計的技巧直接從原文抽取。

使用領域知識來達成摘要系統可以有效的抽取出文件內的主題，但是需要事前大量人工的建立領域知識，包括語言知識、文件主題背景知識、文件寫法

等。建立領域知識的方法目前還很難突破，且都需要該領域的專家建立，距離自動化建立尚有一段距離。

本論文的研究動機是希望用修改原文抽取的方式，並加入自動產生的語意網路，提出一套能夠描述概念且分辨語意歧異的系統，利用不同的概念涵蓋整個文件集，產生出一個能夠符合原文主題的摘要。

第三節 研究目的

摘要是要產生一篇能夠涵蓋原本語意的短文，因此有兩個目的；第一，找出文章中的主題。第二，從多文件中精簡原文。

為了達到第一個找出主題的目的，使用傳統的資訊擷取方法 TF*IDF 先找出新聞文件中可能出現的重要字彙，再以附近出現的字彙去描述該重要字彙，最後使用分群演算法來過濾去掉不相干的雜訊。

要達到第二個目的必須找出含有豐沛語意的句子，依照摘要的內容多寡選取重要度高的句子以達到精簡化的目的，所以我們以特徵選擇出多文件中每個句子的權重，依照權重的高低來當作摘要的句子。另外針對的是多文件的摘要，所以第一個目的中的分群也是為了找出多文件中有重複的主題，進行重複主題的過濾。最後針對上面兩個目的，設計適當的實驗來驗證作法是否合理。並分析描述概念的方法與分群是否能夠有效的過濾出資訊。本論文提出的方法可以適用在單文件或是多文件摘要，但是會把焦點放在多文件摘要，並以預先分類完的新聞文件當作原始文件集。

第四節 論文架構

本論文分成六章。第二章介紹與文件自動化相關的研究；第三章介紹我們如何使用改良型的主題偵測技術(Modified Topic Detection Method)找出以及描述句子中的概念；第四章針對之前選出的概念找出對應的句子(Sentence Mapping

Method)，並以探索式的法則(Heuristic Approach)選出五個特徵，進行句子的權重分析；第五章說明系統實作與實驗結果的分析討論，以驗證本論文所提方法的可行性；最後一章是結論與未來可繼續發展的方向。



第二章 相關研究工作

本論文提出新聞文件自動摘要方法，著重於概念的抽取，以及解決語意歧異的問題，使得在摘要裡面的句子能夠涵蓋最大語意，並去除掉描述重複事件的句子，達到摘要精簡化、去重複化、主題抽取化等目的。

第一節 文件摘要相關研究

自動摘要的作法，大抵可分為「摘錄」(Extraction)與「摘取」(Abstraction)兩種[14]。「摘錄」的結果為文件中重要文句的重組，其作法比較不依賴額外的知識或資源，主要是根據使用者的需求，從文件本身中選取重要文句，編輯組成使用者預期的長度即可。相反地，「摘取」的結果則不限於文件中的文句，其作法需要較多人工準備的資源，如辭典、同義詞庫、詞性標記、語法樹等[15]，經自然語言處理後，自動生成涵蓋原文重點的簡潔文句。由於「摘取」所需資源較多，目前以「摘錄」為主的研究占較多數。

「摘取」的概念擷取經常會透過文法壓縮的方式取得。雖然需要的資源多且處理程序複雜，但是經過摘取處理得到的概念(Concept)會比較接近摘要所需。

「摘取」的步驟通常有主題融合(Topic Fusion)、文字生成(Text Generation)，但不需要文字擷取(Text Extraction)[14]。不過近年來大部分有關摘要的研究重心都在探討如何從原文摘錄，「摘錄」的方式可以比較接近原文件，因為是抽取原文件的句子，基本上所產生的摘要內容是被限制在原文件的。

對於自動產生摘要可能的方法，我們參考了三大類的文獻。如圖 1所示，分成三類，第一類使用自然語言處理(Nature Language Processing, NLP)處理，由語言的文法特性抽取出文章的主題，重組句子形成摘要。第二類使用統計的方式，計算字詞的頻率，依照字詞的重要程度決定文件的概念形成摘要。第三類介紹IR(Information Retrieve)相關技術，包括前置處理、主題擷取等。

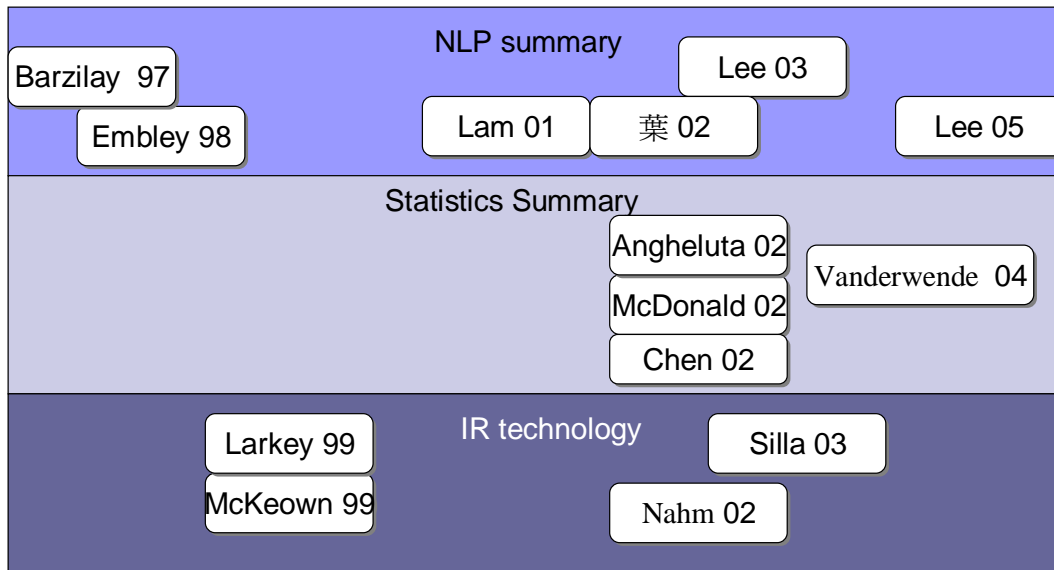


圖 1 相關研究工作

McDonald [14]提到摘要可以分成三大類依照摘要目的(Intent)、摘要的焦點(Focus)、摘要的範圍(Coverage)等三個條件來區分。摘要目的可能包含指出原文中的主題，或提供原文的片段，或提出的摘要可以供其他人來評估原文的主題。摘要的焦點指的是摘要是一般性(Generic)的還是詢問相關(Query-Relevant)的，一般性的摘要是指從原文擷取出來的，詢問相關的是指藉由使用者指定主題，由指定的主題找出與其相關的摘要內容。摘要的範圍則可分成多文件與單文件摘要。

近幾年也有些研討會，例如Message Understanding Conferences(MUC)[3]，Document Understanding Conferences(DUC) [1]，日本NTCIR的TSC[15]每年會有針對摘要系統的互相比較，包括單篇與多篇文件的摘要評比，提供評估的方法。

第二節 以摘取方式為基礎之文件摘要

目前的實作摘要方式主要有兩種，利用資訊擷取的結果來形成摘要。第一種方式使用自然語言處理以瞭解文件中的意思，才能夠把相關的句子聚集起來，產生摘要。使用這類的方式比較複雜，每種語言不同，也會影響處理的方法。第二種方式使用統計找出哪些句子或詞較具有代表性，再由這些句子或詞

來做擴展形成摘要。本節將介紹第一種方式的相關實作。

使用自然語言處理的摘要系統，通常會有幾種方法，機器學習(Machine Learning)、決策樹(Decision Tree)、分類器(Statistical Classifier)、類神經網路(Neural Network)等[9][19]。在 FIDS(Financial Information Digest System)[9]系統中使用了三個方式：分類法、資訊擷取(Information Extraction)、資訊詢問(Information Enquiry)。FIDS 是針對金融新聞文件做出摘要的系統，文章進入系統會先分類，共分成五類：公司表現、經濟結構、合併或併購、服務或產品、債券。資訊擷取是在以自然語言書寫的文件中找出指定的項目[19][3]，因此每一個類別之中都會有各自預先定義好的樣版(Template)，再從分類過的文章取出資訊填滿這些樣版。填滿資訊的樣版，可以被用來當作摘要，或者是存在資料庫中提供給使用者詢問時的答案。圖 2 舉出 FIDS 所定義出的財經新聞樣版。



1. company name
- 2: date of period
3. company performance (good, fair, poor)
4. balance sheet data
 - 4.1 revenue
 - 4.2 net income (net income, income after tax)
 - 4.3 asset
 - 4.4 turnover (turnover, sales)
 - 4.5 earning per share (earning per share, value per share)
 - 4.6 sales (sales, sales revenue, amount of sales)
 - 4.7 loss (loss in sales, decrease in sales)
 - 4.8 delinquency (mistake decision, wrong decision, poor investment)
 - 4.9 income (sales before tax, total sales, gross revenue)
 - 4.10 liability/loan (liability, loan, debt, borrowing)
 - 4.11 expectation gain
5. new product/service (good sales, optimistic sales value)
 - 5.1 selling (good)
 - 5.2 growth (good)
 - 5.3 improvement in market position (gain the market share, leading)
 - 5.4 lower cost (reduce the cost, decrease in cost)
 - 5.5 company restructuring (new branch, expand size of company)
6. performance in the last period (compare to)
 - 6.1 overall performance (excellent, good, fair, poor, very poor)
 - 6.2 reason for change in performance (wrong investment, lack of planning, poor planning, inexperience, market change, change in key people)
7. financial issue:
 - 7.1 no. of share issue:
 - 7.2 amount get from public
8. major business activities:
 - 8.1 amount earning
 - 8.2 other income/loss (other investments, business)

圖 2 W. Lam 提出的財經新聞樣版[9]

類似使用樣版的方式還有 DiscoTEX 系統[19]、Fuzzy Ontology 系統[21]。DiscoTEX 使用機器學習在訓練文件集中找出適合的規則，再利用規則推導 (Rule Induction) 出測試文件中哪些字彙要填入樣版中。使用樣版來擷取資訊的方法雖然對於填入到樣版的資訊是有高的正確率，但是有兩個需要克服的地方。第一，對於特定領域文件的內容以及結構必須十分熟悉，訂出的樣版需要專家的專業知識，經過分析該領域的文件才能決定適合的項目，如此的前置步驟需要花費許多人力，且在轉換文件領域時，必須再重新執行類似的步驟，才能得到適合該領域的樣版，因此目前要做到全自動化仍然有相當大的進步空間。第二，對於該特定領域文件中沒有被樣版對應到的資訊，將會遺失，尤其人類的書寫習慣常常會用不同的描述方式來描述同一件事情，雖然描述的字詞

可能相似，但是若與樣版不符的話，是無法抽取出來的。

第三節 以摘錄方式為基礎之文件摘要

在上一節中，提到了以文法或是知識庫的方式擷取概念，並進而產生摘要。接下來介紹可以適用於一般性文件的摘要方式，這些方式利用一般性的知識庫例如字典檔或是計算詞頻等，去找出重要的字彙或句子，從原文中擷取出來當作摘要。

Angheluta[7]提出以三個步驟來交互產生階層式的主題。第一步是對文章中的名詞建立語彙鍊結[8]，使用字典檔 WordNet 來找出文件中的同義字，把同義字都連結在一起。第二步是使用兩種經驗式(Heruiistics)的方式找出句子中的主題與副主題，第一種是以句子中出現名詞片語的位置來判斷，第二種是看主題字彙是否持續的出現在該句子中。第三步是判斷哪些為主題哪些為副主題，通常主題的話會貫穿全文，副主題的話只會聚集在某一段落之中。經過以上三個步驟之後，會形成一個階層式的樹狀結構，包含了所找出的主題、副主題。

Silla Jr. [10]中提出以分群演算法把語意相關的句子分在一起，句子之間的相似度是以出現在句子中名詞的距離來計算。至於名詞之間的距離則是以名詞出現在 WordNet 中的位置決定。WordNet 中定義了單字的上下位詞、同義詞，基本上如果上下位詞的層級離的越遠，則代表兩個字越不相關。藉由 WordNet 能夠判斷字與字是否相似，進而進行分群，再從每一群之中找出主題。

接著介紹以統計頻率為基礎的技術，利用字詞出現的多寡，或是位置來辨別出不同字詞的重要性。我們也將利用以下的方法做為本論文自動摘要系統的基礎。計算詞頻(Term Frequency)來找出屬於摘要的句子最早是 Luhn 在 1958 年提出的[14]。之後的相關研究加入了其他規則以選擇當作摘要的句子。

Chen[12]提出計算機率的方法，使用了六個以字為基礎的特徵(Features)以及三個以句子為基礎的特徵來計算每一個字或者是句子的權重，以找出最有可能為摘要的句子。

Chen[12]使用貝氏定理(Bayes' Rule)，並應用所定義出來的特徵來決定句子成為摘要的機率，數學模組如方程式 1：

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{P(F_1, F_2, \dots, F_k | s \in S)P(s \in S)}{P(F_1, F_2, \dots, F_k)}$$

方程式 1 貝氏定理應用在特徵挑選上[12]

$P(s \in S)$ 是一個常數代表摘要長度佔文件原文長度的比例。在事前會先給一個訓練文件集，由訓練文件集跟專家所做出來的人工摘要作比對，藉此估計出 $P(F_j | s \in S)$ 和 $P(F_j)$ 所應該佔的機率值。

以下為綜合曾經被提出的特徵值，分成兩類，第一類是針對字彙，第二類是針對句子

字彙的選取特徵：

(1) 字彙使用 TF-IDF 計算[12] [14]：

$$P_{wcon} = tf \times idf$$

(2) 標題字[12] [14]：

假如在本文中的字彙有出現在標題上的，則加重該字分數。

(3) 提示片語(Cue-phrase) [12] [14]：

提示片語，例如：“In conclusion”，“this letter”，“results”，“summary”，“argue”，“propose”，“develop”，“attempt”等。有這些片語出現的句子通常具有涵蓋整篇或是整段的內容，在這種情況下加重後續字彙的分數。

另外也定義了反向的提示片語，例如：“hardly”，“impossible”等，出現這種字彙的話減少其權重。

(4) 偏見字(Biased Word) [12]：

在文章中對於某個字彙的語意上想要更深層，或是更能顯示該字彙所要表達的意思就可以使用偏見字。舉個例子來說，「All men (people) are created equal.」，為了強調「人人平等」因此在「men」之後加註了「people」。出現偏見字時加重其權重。

(5) 跟主題相關的字(Topic Sensitive Word Feature) [12]：

之前特徵都是專注在單一字彙，相對的這個特徵比較著重在跟主題字彙共同出現的字。由於主題通常是名詞跟動詞，而跟隨在附近的通常是修飾這些主題的，修飾字可以是形容詞、副詞。

(6) 大寫字[12] [14]：

出現大寫字加重其權重，只適用於英文。McDonald[14]則是認為在句子中的專有名詞也是相當重要的，但是目前專有名詞抽取的技術未成熟，暫時以大寫字代替。

句子的選取特徵：

(1) 句子出現的位置[12] [14]：

出現在段落的首句或最後一句，增加句子的權重。

(2) 句子所在段落的位置[12]：

該段落是整篇文章的首段或是最後一段。

(3) 句子的長度[12] [14]：

摘要內的句子，其長度不宜過長或過短，大約在 5~15 字之間比較恰當。

在統計方法計算之後，有些研究會以分群法來辨別類似的主題，以避免在摘要中重複著同一個主題，也可以利用分群法來計算主題的重要性。

Angheluta[7]提出以主題分段演算法(Topic Segmentation Algorithm)先判斷文件的主題與副主題，再依照摘要的長短選取適當的字彙數量。若是在 50 字以上的多文件摘要，會進行字彙向量(由句子中的名詞、形容詞、動詞組成)的分群，選取每一群最靠近中心點的句子為摘要。

長度較長的摘要中，使用兩種分群法[7]：Covering 以及 K-medoid。覆蓋類別分群法(Covering Clustering Algorithm)先是找到向量裡面比較密集的幾個向量當作中心，再依照所需摘要的長度來決定中心點所涵蓋的範圍，涵蓋越多則摘要長度越長。K-medoid 則是初期先隨機選取 K 個中心點，每個向量依照與中心點的相似度進行分群，分完之後再重新在每群之中挑選一個中心，每個向量再依照相似度進行一次分群，直到中心點穩定為止。以這兩種方式選取適合句子當作摘要，可以挑選出最具有意義的句子出來。



第三章 改良型概念描述法

多文件摘要系統在於能夠以少數的句子，盡可能的把文件集之內所提到的事件都納入摘要。在本章中，針對新聞的文件集，先做前置處理(Preprocessing)，之後進行對候選概念(Candidate Concept)的描述，方法是採取文章中出現在候選概念前後之相關字來描述，並加入語意網路(Semantic Network)來確定字與字之間的關係，之後再對這些候選概念進行過濾(Filter)跟分群(Cluster)。

第一節 前置處理

DUC2003[1]所使用的新聞文件已經先經過分類，DUC2003 收集 2003 年的新聞，選定幾個主題，每個主題收集 10 篇新聞，每篇新聞給定一個識別號，新聞來源為 AP newswire、New York Times newswire、Xinhua News Agency (英文版)。由於我們採用統計方式找出候選概念，為了避免在之後的運算出現干擾的雜訊，降低統計數據的精確度，因此針對這些新聞文件先進行前置處理，包含片語化(Chunk)、詞幹轉換(Word Stemming)、轉小寫(Lowercase)、刪除停用字(Stop Word)...等。

3.1.1 斷詞切字(Tokenization)

在前置處理中，斷詞切字為第一步驟。英文的斷詞切字相對於中文容易許多，因為中文詞彙之間並沒有特別的標示，而英文字與字之間有空格區隔。本論文使用空格以及英文中常用的標點符號進行斷詞切字，同時處理句子、段落的判斷。

3.1.2 詞性判斷(Part of Speech, POS)

詞性的判斷使用的是愛丁堡大學在 1990 年發展的 NLP Processor[2]，NLP Processor 使用改良式 Penn Treebank tag-set[16]來做訓練的方式，使用的訓練語

料集(Train Set)為網路上所收集到的英文文件大約一百萬字。由於我們的摘要系統對象為新聞文件，內容偏向一般性文章，跟 NLP Processor 的訓練文件集十分類似，另外一方面，NLP Processor 的輸出格式為 XML 格式，可以利用 XML Parser 方便地存取檔案中的內容，因此利用此一工具。

A car bomb exploded Friday in a Jerusalem market crowded with Israelis shopping for the Sabbath.

圖 3 DUC2003 原文範例

圖 3 的原文輸入 NLP processor 之後，經由訓練語料集的比對找出每個單字詞可能的詞性，並以 XML 標籤標注，結果如圖 4 所示。詞性的標記方式如表 1，分成兩大部分：第一部份為開放類別分類，這些類別用來表示一般性沒有固定字來限制的詞彙，像是常用的名詞(Nouns)、形容詞(Adjectives)、動詞(Verbs)以及副詞(Adverbs)等；第二部分為封閉類別，這個類別的詞彙可以用少數有限的字來規定，或是藉由固定的組成方式來判斷。例如介系詞(Prepositions)、冠詞(Articles)以及 WH 開頭的字彙，這些詞性的字彙，都可以藉由規定好且個數有限的字來進行判斷[4]。

```

<NG>
  <W C="NNP">A</W>
  <W C="NN">car</W>
<W C="NN">bomb</W>
  </NG>
<VG>
  <W C="VBD">exploded</W>
  </VG>
<NG>
  <W C="NNP">Friday</W>
  </NG>
  <W C="IN">in</W>
  <NG>
  <W C="DT">a</W>
  <W C="NNP">Jerusalem</W>
  <W C="NN">market</W>
  </NG>
<VG>
  <W C="VBD">crowded</W>
  </VG>
  <W C="IN">with</W>
  <NG>
  <W C="NNP">Israelis</W>
  </NG>
  <W C="VBG">shopping</W>
  <W C="IN">for</W>
  <NG>
  <W C="DT">the</W>
  <W C="NNP">Sabbath</W>
  </NG>
  <W C="." T=".">.</W>
  </S>

```

圖 4 經過 NLP Processor 處理完的 POS 標籤

開放類別的種類(Open Class Categories)		
POS Tag	Description	Example
JJ	形容詞(adjective)	green
JJR	比較級形容詞(adjective comparative)	greener
JJS	最高級形容詞(adjective superlative)	greenest
RB	副詞(adverb)	however, usually, naturally, here, good
RBR	比較級副詞(adverb comparative)	better
RBS	最高級副詞(adverb superlative)	best
NN	一般名詞(common noun)	table
NNS	複數名詞(noun plural)	tables
NNP	專有名詞(proper noun)	John
NNPS	複數專有名詞(plural proper noun)	Vikings
VB	動詞(verb base form)	take
VBD	動詞過去式(verb past)	took
VBG	動名詞(gerund)	taking
VBN	過去分詞(past participle)	taken
VBP	非第三人稱動詞(verb, present, non-3d)	take
VBZ	第三人稱動詞(verb present, 3d person)	takes
FW	外國字(foreign word)	d'hoevre
封閉類別的種類(Close Class Categories)		
POS Tag	Description	Example
CD	數字(cardinal number)	1, third
CC	連接詞(coordinating conjunction)	and
DT	指定詞(determiner)	the
EX	there 存在詞(existential there)	<i>there is</i>
IN	介系詞(preposition)	in, of, like
LS	列表標題字(list marker)	1)
MD	語氣詞(modal)	could, will
PDT	前限定詞(predeterminer)	<i>both</i> the boys
POS	所有格結尾(possessive ending)	friend's
PRP	人稱代名詞(personal pronoun)	I, he, it
PRP\$	所有格代名詞(possessive pronoun)	my, his

RP	質詞(particle)	give up
TO	to (both "to go" and "to him")	to go, to him
UH	感嘆詞(interjection)	uhhuhhuhh
WDT	WH 開頭限定詞(wh-determiner)	which
WP	WH 開頭代名詞(wh-pronoun)	who, what
WP\$	WH 開頭所有格代名詞(possessive wh-pronoun)	whose
WRB	WH 開頭副詞(wh-adverb)	where, when

表 1 詞性標記

3.1.3 詞幹轉換(Word Stemming)

詞幹轉換，是去掉型態學(Morphology)上的詞類型態變化，使得經過變形的字尾能夠有個統一化的結尾，其目的在於做資訊擷取處理時能夠正確辨認字形不一樣但是同樣的字。在 Porter 的演算法中[6]，主要的方式是去判斷各種可能的字尾變化，依照英文文法來加以還原到最有可能辨認整個字，但是又不會殘留有型態上的變化。表 2 以例子來說明這個演算法的結果：

Term	Result
caresses	caress
ponies	poni
ties	ti
caress	caress
cats	cat
feed	feed
agreed	agree
disabled	disable
matting	mat
mating	mate
meeting	meet
milling	mill
messing	mess
meetings	meet

表 2 字彙經過 stemming 之後的變化


由於之後是採用統計的方法來計算概念的重要性，以及對概念作分群，需

要計算字彙出現的頻率、位置以及比較概念的相似度，所以詞幹轉換對於之後的處理效果有很大的提升。

3.1.4 片語化(Chunk)、轉小寫

英文在斷詞切字上面是比較容易的，但是以單一字彙為單位在語意的判斷上並不足夠，在一般的文章中會有很多名詞是使用多個字彙組成的，形成名詞片語，例如：next month、recent years、a critical point。這些片語可能單看一個字彙並沒辦法完全瞭解整個片語所代表的真正意義。所以前置處理如果沒有做片語化的步驟，在後面做統計步驟的時候，就會把相同但是可能代表不一樣意思的單字計算在一起，這樣的統計效果就沒辦法去區分出語意的歧異(Word Sense Ambiguity)。

片語化的方法是使用訓練文件集，訓練文件集中先以人工標出文章中的片語組合，再由程式去統計哪幾種詞性的組合可能是連在一起的片語，以機率的方法去判斷文件集中的字彙是否可形成片語。下面為新聞中的文件做片語化步驟之後的結果。



Federal prosecutors in Manhattan said Wednesday that one of the men accused of conspiring to bomb the U.S. Embassies in Kenya and Tanzania in August had met earlier with Osama bin Laden, the suspected mastermind of the attacks , and “asked him for a mission.” (Source : d30005t/ NYT19981007.0383.xml)

圖 5 片語化之結果範例

圖 5 的範例中，抓出了 Federal prosecutors(聯邦檢察官)、the U.S. Embassies(美國大使館)、Osama bin Laden(奧薩瑪賓拉登)、the suspected mastermind(嫌疑犯)等，都是屬於名詞片語，由這些片語可以大略推測整句話的意思。但是如果由單字詞來看的話，容易產生語意歧異的問題，無法推測正確的語意。所以抓取出來的片語可以當成是整句話的概念，之後的計算也是以這

些概念為主題，找出最能夠代表整篇文章的概念，形成摘要。

轉成小寫的步驟，是為了在計算頻率時防止有些單字以大寫出現在文章中。例如句子開頭、專有名詞之間等，在這些文法規則中雖然出現了大寫，造成與在其他地方出現的單字有了差異，但是單字的意思一樣的，為了消除此種誤差，先將全部單字轉成小寫，以減少後面的統計模組出現誤差。圖 6 為轉完小寫之後的範例。

federal prosecutors in manhattan said wednesday that one of the men accused of conspiring to bomb the u.s. embassies in kenya and tanzania in august had met earlier with osama bin laden, the suspected mastermind of the attacks , and “asked him for a mission.”
(Source : d30005t/NYT19981007.0383.xml)

圖 6 轉小寫之結果範例

3.1.5 停用字(Stop Words)刪除

停用字指的是在文章中沒有語意但是可以用來平順語意的詞，可能包括介系詞、指代詞、連詞、助詞等。這些停用字會常常在文件中出現，所以以頻率計算字彙重要程度的話，有些停用字因此突顯出來，但是這與語意的豐富與否並沒有關係，因此在前置處理中需要先過濾掉，以達到清理雜訊的目的。目前針對 DUC2003 的資料集總共選出 309 個字為停用字。表 3 為部分停用字的範例。

.	,	the	p	in	to	of
a	and	that	said	with	for	is
was	u	he	from	have	he	not
had	by	it	they	who	been	on
but	has	new	an	as	where	at
be	which	el	states	about	him	you're

⋮

whenever	wherever	whom	willing	written	yet	
----------	----------	------	---------	---------	-----	--

表 3 停用字的部分列表

第二節 前後文描述概念

經過前置處理之後，已經從文件中刪除掉許多雜訊，有利於之後所要進行的概念擷取。概念可能由單一或是多個字彙組成的集合，這個集合能夠作為一個概念性的描述，描述出主題的基本範圍。透過這個集合能讓系統瞭解到定義概念所代表的意思[21]。我們的想法是以文件中的名詞或者是名詞片語當作是概念的主軸。在 Silla Jr. 等人的研究中[10]也提出了以文章中的名詞、名詞片語為主題的作法，主要原因是名詞比其他詞性的單字含有較多的語意。但是若直接擷取文件中的名詞或者是名詞片語，仍然會有兩個問題：1. 資訊量仍然過多；2. 無法真正辨認出每個名詞所代表的語意。所以本論文中以 TF-IDF 統計方法來刪除過多的資訊量，以及使用前後文(Context)資訊來描述概念，並對概念加以分群，以確定其語意。

3.2.1 TF-IDF

經過前置處理之後的文件，裡面包含的名詞仍然很多，我們在分群之前再增加一個步驟來過濾多餘的字彙。這一步驟使用傳統資訊擷取過程中計算字詞權重的方式來過濾掉權重較低的字彙。權重公式 $\text{weighting} = \text{TF} * \text{IDF}(\text{Term Frequency} * \text{Inverse Document Frequency})$ ，TF 表示該字詞在某篇文件中出現的頻

率；IDF 表示該字詞出現過的文件數之反轉頻率。其公式如方程式 2：

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad freq_{i,j} \text{ 代表 } k_i \text{ 這個字在文件 } d_j \text{ 中出現的頻率}$$

$$idf_i = \log_2 \frac{N}{n_i} \quad N : \text{ 文件個數 ; } n_i : k_i \text{ 這個字出現過的文件數}$$

$$w_{i,j} = tf_{i,j} \times idf_i \quad w_{i,j} : k_i \text{ 在文件 } d_j \text{ 的權重}$$

方程式 2 TF*IDF 公式

表 4 顯示 TF*IDF 能夠凸顯哪種情形的字彙。

Word	Document				Total Freq	log(N/n)
	D1	D2	D3	D4		
W1	3	2	2	3	10	0
W2	0	5	1	1	7	0.1249
W3	2	0	5	0	7	0.301
W4	0	3	0	0	3	0.602

Word	TF				Weight=IDF*TF			
	D1	D2	D3	D4	D1	D2	D3	D4
W1	1	0.4	0.4	1	0	0	0	0
W2	0	1	0.2	0.333	0	0.1249	0.0249	0.041
W3	0.666	0	1	0	0.2	0	0.3	0
W4	0	0.6	0	0	0	0.3612	0	0

表 4 TF*IDF 例子

表 4 中假設在四篇文章中出現的四個字，先統計出 W1、W2、W3、W4 分別在四篇文章中出現的次數。「Total Freq」是只統計出現次數，若是只以次數出現越多的便越重要，則 W1 似乎是最重要的字，但是經過 TF-IDF 計算之後，W1 在每一篇文章中卻是最沒有語意資訊的。「log(N/n)」代表的是 IDF，計算的是字彙出現在文章數的多寡並取 log。「TF」是針對單一文章計算出現的頻率，並除以頻率最高的值作為正規化(Normalize)。灰色的方塊則是每篇文章中 TF-IDF 值最高的字彙。

TF-IDF 的算法可以用來凸顯在少數幾篇文章中，出現頻率高的字彙。以表 4 來看，W1 這個字彙在 IDF 的得分為零，由於在文件集中的每篇文章都有出現，表示這個字的意義可能是較為一般性的，或是一個停用字，因此比較不可能成為一個文章中在描述的概念。其他得分高的字彙，都是出現文章數量少，但是在單一篇文章中的頻率又很高。這種情形比較符合在文章的書寫形式，成為概念的機曾很大。所以在這步驟，我們先過濾掉 TF-IDF 權重輕的名詞，保留權重高的字彙，進行下一個步驟，如何描述一個概念，以進行概念分群。

3.2.2 前後文描述

在一般性的文章之中，如果是描述同一件事情的話，伴隨出現的字大部分是相似的[7]。依照這個想法，對之前所擷取出來的概念候選字(Candidate Concept)加以出現在該概念的前後文來描述，以增強對概念真正意義的表現。Chen 等人[12]提到了除辨識單一字彙的重要性外，也不能夠忽略出現在重要詞彙附近字的影響力，例如 Condemn(譴責)和 Intensively(強烈) 經常出現在一起，因此注意重要字彙附近的字往往能夠讓該概念的語意更明顯。

對於概念候選字的詞性再作進一步篩選，挑選了類名詞的詞彙當作概念，一般研究皆認為名詞或動詞比起其他的詞性具有較多語意[7][9][10][14][17]。因此從處理完的概念候選字再挑選出適合的詞性(Part of Speech)，包括一般名詞、複數名詞、專有名詞、複數專有名詞等，之後才進行概念描述的步驟。

我們依照一般文件書寫的習慣，找出這些概念出現時常會伴隨出現的文字，以這些文字來描述這個概念。先以出現在概念前面的 N 個字以及出現在概念後面 M 個字來當作描述概念的字彙。可以當作描述的字彙也是一樣要經過前置處理，但是前置處理僅包括片語化、轉小寫等，不處理過濾掉停用字以及 TF-IDF 過低的字彙。原因在於用來描述概念的字彙要確實是出現在概念周圍，如果刪除了這些字彙，有可能使得前後文描述時會跨過一些原本應該是鄰近的

字彙。

在實作時有關 N 、 M 兩個變數的設定，是採用概念前後 5 個字彙，加上概念本身的話，總共包含了 11 個字彙。在心理學的研究中顯示人類的短暫記憶通常是 7 ± 2 個字彙[12]。在本論文的實驗裡面，最後選定了 N 、 M 各等於 5，亦即為心理學研究結果的最小值，取最小值的目的是要使前後文的涵蓋範圍小一點，避免彼此相鄰的概念在描述的內容有過多的重複。因為之後需要經過分群演算法，相似度過高的內容拿去做分群的話，會影響分群的結果[13]。

從 DUC2003 的文件集之中，以一篇新聞文件示範如何對前述步驟所抓取出來的名詞、名詞片語來做描述。

BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.

Source : d30005t\APW19981104.0772.xml

Concept	Index Description	Length
German police	German police, raided, several locations, Bonn, receiving, word, a terrorist threat	6
several locations	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy	8
bonn	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence	9
word	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	10
a terrorist threat	raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials	11
the U.S. Embassy	several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	11

no evidence	Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	10
a planned attack	receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	9
found	word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	8
officials	a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	7
Wednesday	the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	6

表 5 前後文描述概念範例


上述的範例把同一句中的 11 個概念利用出現在他們前後的字彙作描述，由於在句首跟句尾的概念，其周圍的字彙較少，因此描述的字數變少。這種描述概念的方式是希望能夠讓概念的語意更明顯，使得在分群的時候可以判斷概念彼此的相似度；且如果描述的字彙內容很相似的話，在分群時要會被分在同一群之中。儘管在句首或句尾的概念用比較少的字彙描述，但是如果這些概念跟鄰近的概念很接近的話，也會使得描述內容相近，進而分到同一群內。所以描述的字彙少的話，還是可以依照相近的概念來取得相近的內容而分到同一群中，不會因為描述的字彙少而影響到分群的結果。

第三節 加入語意網路描述概念

本節中，我們試著在上一節的描述裡加入了語意網路來增加描述時的語意概念。語意網路(Semantic Network)的作法有相當多種，我們使用史丹佛(Stanford)大學所開發出來的語意工具 Infomap[5]。

Infomap[5]把一個文件集當成一篇文章，在文件中的每一個字將會被編成一個向量，向量依照字彙在一個範圍(Window of Text)之內出現的次數來編成，

範圍是可以被設定更改的。利用該字的向量可以知道該字在整篇文件中每個範圍的分佈情形，因此在共現矩陣(Co-occurrence Matrices)中每一列(Row)代表一個字，每一行則是一個範圍，矩陣中的值即是字出現在該範圍的次數。Infomap提出的方法是希望能找出哪些字與字之間的語意是相關的。由於在同一篇文章中，通常作者的寫法對於同一件事物都會盡量用同一個詞去表示。例如：在體育新聞中寫到 umpire 或 referee 都是表示裁判的意思，但是很少有文章會同時提到這兩個字。因此 Infomap 採用「共現」(Cooccurrence)的方式來計算。一開始依照字出現的頻率選訂出語意基本字(Content Bearing Words)，再訂出一個可調式範圍，在這個範圍之內的每一個字伴隨語意基本字一起出現的頻率，就把這些頻率定在共現矩陣裡面。圖 7 為共現矩陣的範例。選定的語意基本字為「Music」、「Food」，基本字最好具有語意清晰以及不易混淆的特質，不過在挑選時還是以頻率高低為選擇基準。



<p>HOT-FROM-THE-OVEN MEALS: Keep hot food HOT; warm isn't good enough. Set the oven temperature at 140 degrees or hotter. Use a meat thermometer. And cover with foil to keep food moist. Eat within two hours.</p>	<p>“Change is always happening,” said the ebullient trumpeter, whose words tumble out almost as fast as notes from his trumpet. “That’s one of the wonderful things about jazz music.” For many jazz fans, Ferguson is one of the wonderful things about jazz music.</p>
---	--



	eat	hot	jazz	meat	trumpet
Music	0	0	3	0	1
Food	1	2	0	1	0

圖 7 Semantic Network 共現矩陣範例[5]

依照實驗的結果，字與字之間如果向量相似的話，在語意上的意義通常是比較相近的。Infomap 會利用共現矩陣，並減少字向量的維度，再利用餘弦(Cosine)來比較向量的相似度。

最後使用奇異值分解 (Singular Value Decomposition, SVD)[18]降低維度，最後利用餘弦算出矩陣內每個向量的相似度，得到一個語意網路。

圖 8 為一個例子來說明 Infomap 語意網路，以「attack」這個單字透過語意網路找尋出在文件中相關的字詞：

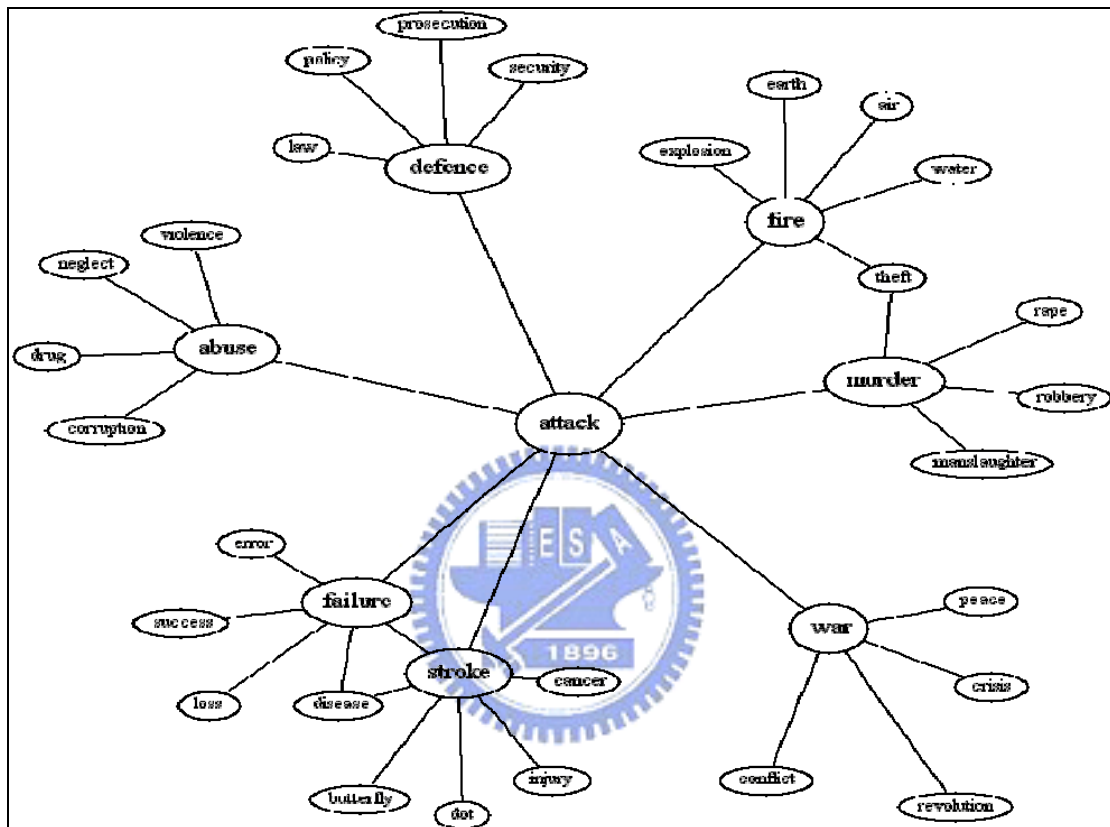


圖 8 Infomap 的語意網路範例[5]

Infomap 會在共現矩陣中找尋最常共同出現的字彙當作「attack」的第一層相似字彙，再依次找出第一層的字彙跟哪些字彙有共現關係。在描述概念時加入語意網路，可以更加反應描述時的精確度。

3.2.2 中使用了前後文的方式描述概念，在這本節中嘗試把上述所提的語意網路也加入描述概念的方法中，提出了兩個方法來實作。第一個方法是在前後文中只取跟概念有在語意網路上出現的字彙。第二個方法則是仍然取用前後文來當描述概念的方式，但是對於有出現在語意網路上的字彙則加重其權重。

表 6 為一個例子來表示第一種加入語意網路的方式：

<p>BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.</p> <p>Source : d30005t\APW19981104.0772.xml</p>
--

Concept	Index Descript	Length
	Add Semantic Network	Length
German police	German police , raided, several locations, Bonn, receiving, word	6
	German police , raided, several locations, Bonn, found, Wednesday	6
several locations	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy	8
	German police, raided, several location, Bonn, receiving, word, a terrorist threat, the U.S. Embassy	8
Bonn	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence	9
	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence	9
word	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	10
	German police, raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	10
a terrorist threat	raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials	11
	raided, several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials	11
the U.S. Embassy	several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	11
	several locations, receiving, a terrorist threat, the U.S. Embassy, no	6

	evidence, a planned attack	
no evidence	Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	10
	raided, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	6
a planned attack	receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	9
	receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials	8
officials	a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	7
	receiving, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, officials, Wednesday	7
Wednesday	the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday	6
	a terrorist threat, the U.S. Embassy, no evidence, a planned attack, founds, Wednesday	6

表 6 加入語意網路方法一的範例

在表 6 的範例舉出原本的描述方式以及使用 Infomap 的描述方式，觀察出對於所描述的概念的確更準確。例如，美國大使館這個概念(the U.S. Embassy)，在原本的描述多了 Bonn、word、found、officials、Wednesday 等五個字彙，由於在語意網路裡面「the U.S. Embassy」跟上述五個字並沒有相關的關係，因此在加入語意網路之後可以消除這種情形，使得描述概念的字彙更清楚。

第二種加入語意網路的方法，是採用與所描述的概念有關的字彙，加重其權重。採用這種方式是希望能夠凸顯與概念有關的字彙，並且不會影響到原本描述字彙的組成。在未加入語意網路時，描述的字彙是在代表該字彙的向量維度中加入 TF*IDF 值，加入語意網路之後可以知道哪些字與所描述的概念比較有關係，因此增加其權重就是在增加該字彙的語意值，使其增大。表 7 為一個例子來說明如何透過語意網路加重權重：

BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.

Source : d30005t\APW19981104.0772.xml

Concept	Index Descript			
	Weight			
German police	German police	raided	several locations	Bonn
	4.47734+X	4.47734+X	5.17048+X	5.17048+X
	receiving	word		
	2.97326	5.17048		
the U.S. Embassy	several locations	Bonn	receiving	word
	5.17048+X	5.17048	2.97326+X	5.17048
	a terrorist threat	the U.S. Embassy	no evidence	a planned attack
	5.17048+X	2.97326+X	5.17048+X	5.17048+X
	found	officials	Wednesday	
	3.56105	4.47734	2.97326	

表 7 加入語意網路方法 2 的範例

上面的範例可以比照表 6，描述概念的字彙與在沒有加入語意網路的描述是相同，但是有出現在語意網路的字彙增加其權重，加重的值會在系統評估的時候來說明會增加多少才可以有效地凸顯字彙與概念的相關語意。

觀察表 6、表 7 以及原本的描述方式，可以發現在方法一中雖然描述比較能夠貼近概念的語意，但是在描述的字彙分佈比較散，遇到不在語意網路內的字彙會跳過，也會因為句子長度的關係使得描述的字彙無法像原本的描述字彙這麼多，所以雖然描述精準但是會有描述字彙不足的情形，連帶會影響到之後在分群以及後來計算特徵的權重。

第四節 概念以及句子分群

前一節之中，已經針對概念做出了字彙向量(Term Vector)，Angheluta 等人 [7]也針對由結構化的句子中抽出的字彙向量去做分群計算(Covering 和 K-Medoid)，由此分群的結果找出每個句子對應的中心點，因而可以找出哪些句子是相似的。類似分群的方法 McDonald [14] 也提出過，把整篇文章依照類似的字義去分段(Document Segmentation)，由分段之中更容易地判斷出隱藏在句子中的主題。因此在做句子的特徵判斷前，如能將類似語意的句子分割出來，將能提升後續特徵選擇的精確度。

本論文中分群的對象，是經過上述處理完以前後文去描述的概念向量。分群的方式大概可以分成兩種：一、劃分方法(Partitioning Method)、二、階層方法(Hierarchical Method)。劃分方法在中小規模的資料庫中發現球狀叢集的資料時較為適用，而階層式的方法雖然計算成本較小，但是每個階層完成之後，這個階層的規則將無法更改，一旦在上層出現錯誤，將會影響其後的分群結果 [11]。DUC2003 的資料已經把相似主題的新聞文件整理在一起了，且在我們前述的處理之後，每個新聞主題約有 2000 個概念，因此選擇使用劃分方法來分群，使用的分群方法是 K-means。

K-means 以 K 為參數，把 n 個物件分為 K 個叢集，以使叢集內具有較高的相似度，而叢集間的相似度較低，相似度的計算是根據一個叢集中物件的平均值來進行[11]。

K-means 演算法的處理流程如下。首先，隨機地選擇 K 個物件，每個物件代表一個叢集的平均值或中心。對剩餘的每個物件，根據其與各個叢集中心的距離，將它指定給最近的叢集。然後再重新計算每個叢集的平均值。這個過程不斷重複，一直到判斷準則函數收斂。通常，判斷準則函數會採平方誤差準則(Squared Error Criterion)，定義如方程式 3：

$$E = \sum_{i=1}^k \sum_{p \in C} |P - m_i|^2$$

方程式 3 平方誤差準則

E 是資料庫中所有物件平方誤差的總和，P 表示資料物件在空間中的點， m_i 是叢集 C 的平均值(P 和 m_i 都是多維的)。這個準則試圖使生成的結果叢集盡可能地緊湊和獨立。演算法總結如圖 9：

輸入：叢集的數目 K，以及包含 n 個物件的資料庫。

輸出：K 叢集。且使平方誤差準則最小

方法：

- (1) 任意選擇 K 個物件作為初始的叢集中心；
- (2) repeat
- (3) 根據叢集中物件的平均值，將每個物件(重新)指定給最類似的叢集
- (4) 更新叢集的平均值，也就是計算每個叢集中物件的平均值
- (5) until 平方誤差小於門檻

圖 9 K-means 演算法[11]

K-means 演算法嘗試找出平方誤差合數值最小的 K 個劃分。當結果叢集越密集，且叢集與叢集之間區隔特別明顯時，效果會非常好。對處理大資料集，該演算法是相對可以延展的和高效率的，因為 K-means 的複雜度是 $O(nkt)$ ，其中， n 是所有物件的數目， K 是叢集的數目， t 是疊代的次數。正常而言， $k \ll n$ ，且 $t \ll n$ 。K-means 演算法經常得到的是一個局部最佳值(Local Optimum)。

將 DUC2003 的文件集來做分群，由於文件已先經由專家分類過，每一個類別都有 10 篇文章，這 10 篇文章都是在講同一個新聞事件，所以裡面的內容是比較類似的(跟事前沒有經過分類的新聞文件相比)。但同一個新聞事件也是可以再細分為地點、對象、影響結果...等。符合 K-means 演算法叢集越密集、且叢集與叢集之間區特別明顯的特點，因此選用了該演算法，希望能到最好的效果。表 8 為從 DUC2003 選出一篇新聞文件(d30005t \ APW19981104.0772.xml)

的分群的结果：

Concept	Index
Concept from which Sentence	
第一群	
no evidence	found, no evidence, a terrorist threat, a planned attack ,the U.S. Embassy
BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.	
a planned attack	the U.S. Embassy, found, officials, no evidence, a planned attack
BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.	
officials	officials, found, officials, Wednesday, a planned attack
BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.	
wednesday	found, officials, Wednesday
BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.	
explosives	the sites, stockpiled, found, arms, explosives
The agency said it had received ``credible information" that Middle East terrorists had stockpiled arms and explosives at the sites, but none were found.	
the sites	the sites, found, arms, explosives
The agency said it had received ``credible information" that Middle East terrorists had stockpiled arms and explosives at the sites, but none were found.	
第二群	
a suspected top aide	Mamdouh Mahmud Salim, A suspected top aide, bin Laden
A suspected top aide of bin Laden, Mamdouh Mahmud Salim, is jailed in Germany pending a U.S. extradition request, raising concern about reprisals on German soil.	
bin laden	Mamdouh Mahmud Salim, jailed, A suspected top aide, bin Laden
A suspected top aide of bin Laden, Mamdouh Mahmud Salim, is jailed in Germany pending a U.S. extradition request, raising concern about reprisals on German soil.	
Mamdouh Mahmud Salim	bin Laden, Mamdouh Mahmud Salim, jailed, A suspected top aide, Germany

A suspected top aide of bin Laden, Mamdouh Mahmud Salim, is jailed in Germany pending a U.S. extradition request, raising concern about reprisals on German soil.	
第三群	
u.s. authorities	U.S. authorities, charge, Salim
U.S. authorities charge Salim helped finance, train and arm members of a terrorist organization, including the alleged bombers of the U.S. embassies in Kenya and Tanzania.	
salim	finance, U.S. authorities, charge, Salim, helped
U.S. authorities charge Salim helped finance, train and arm members of a terrorist organization, including the alleged bombers of the U.S. embassies in Kenya and Tanzania.	

表 8 概念分群範例

表 8 內都是出自於同一篇文章，分群的時候使用 K-Means，在這個例子中是分成五群，由於是 10 篇相同主題的新聞文件去分類，在這篇文章中恰好只有分到三群中，也就是裡面的意思可以跟其他篇的文章分成三群。

第一群中的句子大概都具有發現、找尋的意思。這篇文章中有六個概念都被分到第一群之中，這個類別都是在講官方組織去搜尋恐怖份子或是恐怖事件，因此屬於該叢集的兩個句子都是在描述這方面的事件 officials→find→terroristic event。

第二群中的概念是針對恐怖份子的主角賓拉登(Bin Laden)，在這篇文章中出現兩個有賓拉登的句子，其中一個句子被歸在叢集裡面，另外一個句子因為做名詞化的時候，"Saudi dissident Osama Bin Laden" 五個字當成了同一個片語，以致於沒有當成是同一個概念，不過在此群中仍然聚集了絕大部分跟賓拉登相關的句子。

第三群中環繞在控訴引發恐怖事件的恐怖份子，這篇文章中只有一句話被歸到該叢集之中，該叢集的大致意思為美國官方控告恐怖份子引起眾多的恐怖事件，該叢集大概可以看出 officials→charge→terrorist 的關係。

在概念分群的過程中，相似的概念的確可以被群聚在一起，不過由於分群

的過程中很多概念因為與目前的中心點沒有相似的字彙，無法判斷與各中心點的相似程度，因此結果中很多概念沒有被分群。在這個例子中，經過前置處理後有 1,943 個概念，但是只有 573 個概念有被分群到，共有 29.49% 概念被分群，平均有 2/3 的概念會在分群中被過濾掉。

在概念分群之後，將句子跟分群完的概念作對應，找到能夠代表每個句子的叢集。我們嘗試了兩種對應方法：第一，判斷句子中的字彙出現在哪個叢集中的字數最多，則歸類到該叢集。第二，判斷句子中的概念出現在哪個叢集中的字數最多，則歸類到該叢集。

<p>(1) $SIM_{s,i} = \text{Words Match}$ $= \text{sim}(\text{Match_Word}, \text{Cluster } j) / L_of_S$</p> <p>(2) $SIM_{s,i} = \text{Concepts Match}$ $= \text{sim}(\text{Match_Vector}, \text{Cluster } j) / L_of_S$</p> <p>Match_Vector : vector included in this sentence $\text{sim}(\text{Match_Word}, \text{Cluster } j)$: number of word appear in cluster_j $\text{sim}(\text{Match_Vector}, \text{Cluster } j)$: distance between vector and centroid of cluster L_of_S : Length of Sentence</p>

方程式 4 句子對應到叢集的方式

第一個方式只是單純去判斷句子中有多少字出現在該叢集裡面，叢集裡原本只有包含概念，不過我們嘗試著把描述概念的字彙也加進叢集裡，這樣的作法是希望能夠增加句子對應到字彙的數量，避免一句話裡只有少數幾個字彙出現在叢集內，對應的字彙數量多一點，比較容易判斷句子應該是屬於哪一叢集。

第二個方法只判斷句子中的概念在哪个叢集中，由於概念是以編成向量的方式在做 K-means 分群，因此每個向量都可以找出與所屬叢集的相似度，也就是離中心點的距離。當句子中有向量出現在叢集之中時，會以該向量的離中心

點的相似度當作句子跟這叢集的相似度。以圖 10 來當作範例：

假設維度為 2 的情形下做範例，比較方便瞭解，但是實際上維度可能在 2000 上下，此句子恰好有三個向量，分別對應到兩個叢集，對應到叢集 1 只有一個概念，但是該概念是叢集 1 的中心，兩個向量對應到叢集 2，但是對應到的概念是叢集 2 比較外圍的概念，也就是說與中心點的相似度較小。

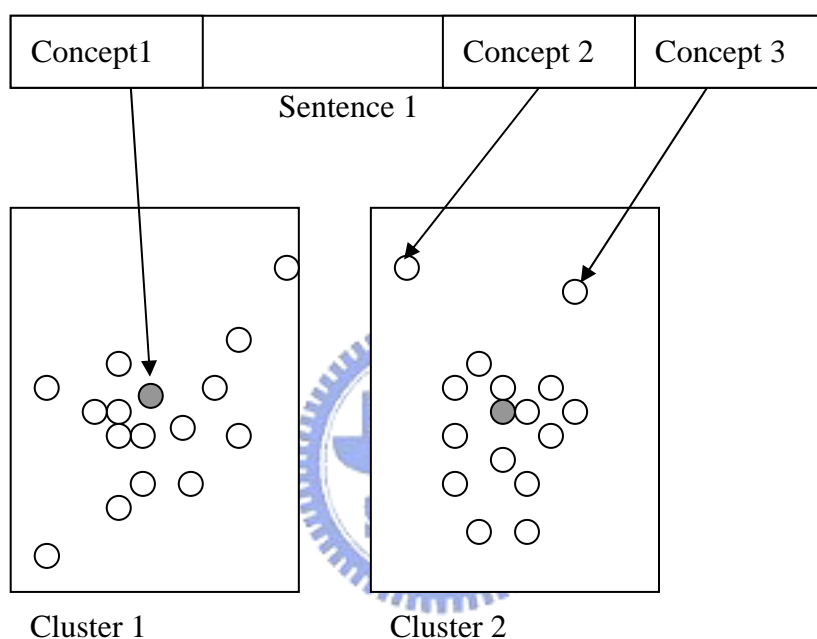


圖 10 句子對應到叢集

假設概念 1 對叢集 1 的相似度為 1、概念 2 對叢集 2 的相似度為 0.1、概念 3 對叢集 2 的相似度為 0.15，依照之前的公式可以算出句子對叢集 1 的相似度為 $(1/\text{句子長度})$ ，對叢集 2 的相似度為 $((0.1+0.15)/\text{句子長度})$ ，因此雖然在句子 1 中有比較多的字彙出現在叢集 2，但是在方法 2 的計算下，會對應到叢集 1。

比較這兩種方法發現方法 2 能融入語意的相似程度，更能夠表現出句子跟叢集之間的關係，在第五章評估中也有做了兩個方法的比較，方法 2 的確能夠選出更適當的句子來表現在摘要之中。

我們針對兩種方法作了句子對應到叢集，圖 11 和圖 12 以同一篇新聞文件來當作範例，兩個對應方法對句子的分群影響：

BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday. (第一群)

Police, including agents of an elite anti-terrorist unit, checked several suspects during raids in an industrial zone and other sites Tuesday, but no arrests were made, said Eva Schuebel, spokeswoman for the Federal Prosecutor's Office in Karlsruhe. (第二群)

The agency said it had received ``credible information'' that Middle East terrorists had stockpiled arms and explosives at the sites, but none were found. (第二群)

The agency's investigation is continuing, Schuebel said. (第一群)

``For now, we can no longer speak of an immediate threat to the U.S. Embassy," she said. (第二群)

The embassy had no comment. (第一群)

Security was tightened at U.S. installations worldwide after the Aug. 7 bombings of the U.S. embassies in Kenya and Tanzania. (第三群)Saudi dissident Osama Bin Laden is wanted by U.S. officials for those attacks. (第三群)

A suspected top aide of bin Laden, Mamdouh Mahmud Salim, is jailed in Germany pending a U.S. extradition request, raising concern about reprisals on German soil. (第二群)

U.S. authorities charge Salim helped finance, train and arm members of a terrorist organization, including the alleged bombers of the U.S. embassies in Kenya and Tanzania. (第二群)

In September, German police stepped up security at the U.S. consulate in Hamburg after receiving a tip about a possible threat. (第三群)

Source:d30005t\APW19981104.0772.xml

圖 11 句子對分群結果的對應情形 1

在上述例子中，對整個 DUC2003 的其中一類新聞文件分成五群，一個類別包含 10 篇文章，上面的這篇新聞文件中，只佔了三群。整篇新聞所要表達為找尋恐怖份子的行動。

第一群中包含了三句話，三句話中大致可以看得出是在講跟美國大使館 (U.S. Embassy) 有關的事件，包括了跟大使館有關的恐怖攻擊、發表的意見。

第二群包含的句子數量較多，包含了五句，且可以看出為該篇新聞報導的主題，此群大致是圍繞在警方搜查了恐怖份子可能出現的地方(anti-terrorist、suspect、credible information)，還有關於嫌疑犯 Mamdouh Mahmud Salim 的描述。

第三群包含了三個句子，三個句子大致可以看出是對於恐怖攻擊事件的發生，因而提高了安全等級(Security was tightened、stepped up security)，內容偏向對於恐怖攻擊做了哪些措施。

<p>BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday. (第一群)</p> <p>Police, including agents of an elite anti-terrorist unit, checked several suspects during raids in an industrial zone and other sites Tuesday, but no arrests were made, said Eva Schuebel, spokeswoman for the Federal Prosecutor's Office in Karlsruhe. (未分群)</p> <p>The agency said it had received ``credible information'' that Middle East terrorists had stockpiled arms and explosives at the sites, but none were found. (第一群)</p> <p>The agency's investigation is continuing, Schuebel said. (未分群)</p> <p>``For now, we can no longer speak of an immediate threat to the U.S. Embassy," she said. (未分群)</p> <p>The embassy had no comment. (未分群)</p> <p>Security was tightened at U.S. installations worldwide after the Aug. 7 bombings of the U.S. embassies in Kenya and Tanzania. (未分群)Saudi dissident Osama Bin Laden is wanted by U.S. officials for those attacks. (未分群)</p> <p>A suspected top aide of bin Laden, Mamdouh Mahmud Salim, is jailed in Germany pending a U.S. extradition request, raising concern about reprisals on German soil. (第二群)</p> <p>U.S. authorities charge Salim helped finance, train and arm members of a terrorist organization, including the alleged bombers of the U.S. embassies in Kenya and Tanzania. (第三群)</p> <p>In September, German police stepped up security at the U.S. consulate in Hamburg after receiving a tip about a possible threat. (未分群)</p> <p>Source:d30005t\APW19981104.0772.xml</p>

圖 12 句子對分群結果的對應情形 2

由方法 2 作出來的對應方式，全篇文件中可以被分群的句子數量變少了(如圖 12)，不過叢集中的句子語意更集中。每個叢集中的句子跟上一步驟處理概念分群的結果具有語意上的符合，叢集內的概念主題可參考表 8 的說明。

從句子層次比較兩個方法，方法一的對應由於把描述概念的字彙也加入對應的條件，因此幾乎文件集內的每一個句子都可以找到對應到的叢集，造成了每一個叢集內的句子數量多，但是句子的語意可能不是與叢集的概念相似，因為可能只有對應到描述概念的字彙，並不是對應到概念本身。相比之下，方法 2 的對應可以有效的過濾掉語意不夠符合叢集的句子，雖然剩下的句子數量較少，但是再經由後面的特徵選取時，可以提升選取適合摘要句子的效率。



第四章 語句語意權重摘要

透過叢集方法的切割，可以判斷出語意相近的概念並找出對應這些概念的句子，在第三章中也舉出例子可以看出語意相近的句子被叢集在一起。在同一叢集裡面的句子雖然語意近似，但是仍需要一些條件來找出哪些句子最能代表整個叢集。由句子特徵來挑選句子的方法在很多研究中被提出來，藉由不同的方法抽取出不同的特徵，且假設這些特徵彼此是獨立的並且可以藉由整合這些特徵來判斷句子的重要程度[14]。

依照之前的前置處理以及分群動作，這一節中定義了三個與句子、兩個與叢集有關的特徵，利用這些特徵來分辨出同一個叢集裡句子的重要性。

第一節 句子的特徵

本論文以句子內字彙的 TF*IDF、句子出現在文章中的位置、句子與所屬叢集的相似度這三個特徵量去計算句子的重要性。下面詳細介紹這三個特徵：

4.1.1 TF*IDF

把句子裡面所有字彙的 TF*IDF 加總，再除以句子的長度。算法如下：

$$S_{tfidf} = \left(\sum_{i=1}^m TF \times IDF_i \right) / sentence_length$$

由於 TF*IDF 可以算出字彙在文件中所帶的語意強度，因此選取這個特徵是希望能夠藉由字彙的語意來計算句子所包含的語意強度，但是為了避免句子長度越長使得 TF*IDF 的總和越高，因此除以句子長度加以正規化 (Normalization)。下面以文件中一個句子當作範例：

BONN, Germany (AP) _ German police raided several locations near Bonn after receiving word of a terrorist threat against the U.S. Embassy, but no evidence of a planned attack was found, officials said Wednesday.

Source : d30005t\APW19981104.0772.xml

term	German police	raided	several locations	Bonn	receiving	word
tf*idf	4.47734	4.47734	5.17048	5.17048	2.97326	5.17048
term	a terrorist threat	the U.S. Embassy	no evidence	a plan attack	found	officials
tf*idf	5.17048	2.97326	5.17048	5.17048	3.56105	4.47734
term	Wednesday					
tf*idf	2.97326					

表 9 Tf-IDF 範例

表 9 的例子在這個特徵中可以得到

$S_{tfidf} = (4.47734 + 4.47734 + 5.17048 + 5.17048 + 5.17048 + 2.97326 + 5.17048 + 5.17048 + 3.56105 + 4.47734 + 2.97326) / 13 = 3.5549$ 。每個句子都會經過同樣步驟的計算當作之後挑選摘要的一個特徵。

4.1.2 句子在文章中出現的位置

一個段落的第一句或是最後一句通常會涵蓋較多上層的語意，也就是涵蓋的語意範圍廣，這對於摘要有很大的幫助。句子會有階層式的組織，通常在首句或尾句會有關鍵性的資訊隱含在其中[12]，尤其是在段落的首句或尾句常常會有「in conclusion」、「summarily」等字彙出現，句中有出現這種字彙將會加重這個句子的語意廣度。所以如果句子出現在這個位置的話，會加重這句話的權重，加重的比例會在第五章實驗分析來討論。

4.1.3 句子與所屬的叢集相似度

第三章第四節中提出兩種句子對應到分群的方法，第一種是比較句子與叢集中有共同出現的字彙數量，第二種是比較句子中出現概念與叢集的相似度。

這兩種方法都可以取出一個特徵，來計算句子與所屬的叢集有多大的代表性。

因為摘要的目的之一是要精簡原文，我們使用分群的方法來找出概念，基於精簡的原則，每個叢集也應該找出具有代表性的句子才能達到精簡化的目的，所以取用這個特徵就是希望能夠找出代表性的句子。

在句子對應叢集方法一中，計算相似度的公式如方程式 5

$$S_{sim} = match_words_i / sentence_length$$

Match_Words：計算與叢集 i 內有多少字彙是一樣的

i：句子所對應到的叢集 i

Sentence_Length：句子的長度，裡面有多少字彙

方程式 5 相似度特徵計算

方法一的相似度取決於所對應到的字彙數目，因為在對應的方式中是用字數的對應，所以這個方法也是採用同樣的方式來計算相似度，再加上除以句子長度作正規化。不過在第五章實驗的結果可以發現以這種方式來計算相似的話的，會發生有很多句子所對應到的字彙數量是一樣的，會使得這個方法計算出的權重不具有太多意義。

在句子對應叢集的方法二中，計算相似度的公式如方程式 6

$$SIM_{s,j} = \frac{1}{\sum_{i \in s} (distance(concept_i, cluster_j)) \times L_s}$$

concept：句子有對應到叢集的概念

distance(concept, cluster)：取向量到叢集中心的距離

L:句子的長度

方程式 6 相似度特徵計算 2

方法 2 中是取決於向量對應的叢集與其中心點之間的距離，使用這個方法可以比較哪些句子比較接近該叢集的中心點。在多維度的向量中，使用歐基理得距離(Euclidean Distance)可以更精確地找出哪些向量接近中心點，每個句子將可以更清楚地分出代表叢集的高低。

摘要的目的之一，是要能夠代表全篇文章的意思，因此選用這個特徵是希望能夠找出代表叢集內概念的句子，之後的實驗評估也會針對這兩個方法作分析。

第二節 叢集的特徵

在第一節中提出了三個在句子階層的特徵，由於是基於分群方法來找出概念，因此提出在叢集階層選用特徵，希望在這個階層中的特徵可以有效的加入分群時的一些狀態。叢集階層中選了兩個特徵：叢集內含的概念多寡、叢集與整體中心的距離。



4.2.1 叢集內含的概念多寡

經過前置處理完的字彙形成概念，再對概念作分群，每一個叢集內含的概念數量都不會一樣，包含越多的概念數量，表示原文件集提到的許多概念都在同一個叢集，可以看成包含概念越多的叢集，其包含的概念也是佔原文件集中的多數，Chen [12]也選擇叢集內概念的數目多寡來當作特徵，越多概念的叢集，權重應該越高。

4.2.2 叢集與中心點的距離

分群完之後的結果，依照向量的分佈情形可以找出全部向量的中心點。每一個叢集中越靠近中心點的給予越高分。以圖 13 來解釋這個特徵：

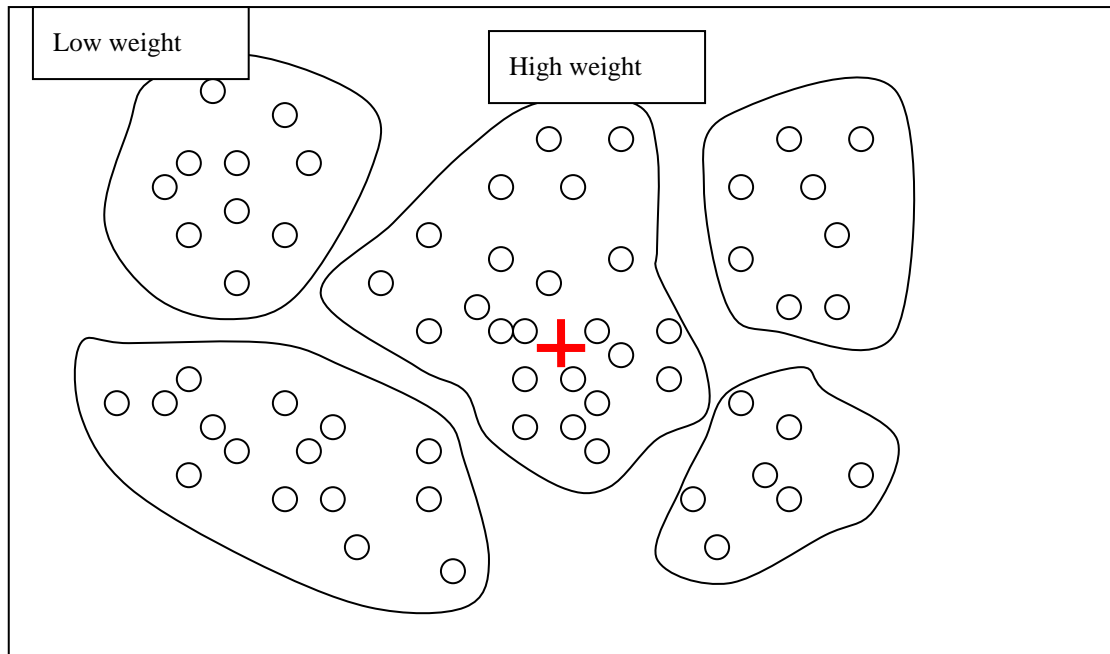


圖 13 叢集特徵圖例

選擇這個特徵是因為在中心點附近的叢集，越有可能涵蓋其他叢集的意思，在順序上應該要比其他遠離中心點的叢集要重要，因此加入這一個叢集，期望能加強涵蓋性越大的叢集。

第三節 句子挑選

綜合上述的五個特徵可以得到一個權重總和：

$$sentence_weight = \alpha(C_{length}) + \beta(S_{tfidf}) + \gamma(C_{distance}) + \theta(S_{location}) + \lambda(S_{sim})$$

「 C_{length} 」為叢集內的向量個數；「 S_{tfidf} 」為句子內字彙的 TF*IDF 總和；「 $C_{distance}$ 」為句子所屬叢集距離全體向量質心的距離倒數；「 $S_{location}$ 」為句子所在位置；「 S_{sim} 」為句子與所屬叢集的相似度。

這五個特徵值都是互相獨立且彼此可以互相補足所欠缺的特性，句子的權重將會依照不同的比例把上述五個特徵加總。基本上是按照該權重的總和來挑選句子順序當作摘要。但是加入叢集的順序來作判斷。叢集的順序是依照叢集內的最高句子得分的做排序。挑選的時候從第一個叢集內選出最高分，再來選

擇第二個叢集內的最高分，直到選完全部叢集內的最高分。再來重新選擇第一個叢集，但是這次選擇第二高分，其後依此類推。



第五章 實驗結果分析與評估

自動摘要的成效評估，可分為直接(Intrinsic)與間接(Extrinsic)兩種方式[14]。直接的評估需先定義出一組理想的摘要準則或答案，然後跟系統取出的摘要做比較。尤其是給人閱讀的摘要，其評估準則有重點涵蓋率(Coverage)、可讀性(Readability)、連貫性(Coherence)、凝聚性(Cohesion)、組織性(Organization)及摘要長度等，因此文句中的連接詞(Conjunction)、代名詞(Pronoun)、前後文照應詞(Anaphor)等需做適當的修詞(Rhetoric)處理。間接的方式則無須具備理想的摘要答案，而是評估自動摘要的結果在其他相關應用的成效。例如，以問答的方式，比較使用者分別閱讀全文與閱讀摘要後，回答問題的成績來比較自動摘要的成效；或者無需人工介入，將原來以全文進行的自動分類或主題檢索，以摘要來取代全文，求出摘要的分類或檢索成效，全自動的比對出各種自動摘要的效果。

本摘要系統使用的文件集為DUC2003(Document Understanding Conferences 2003)[1]，文件內容是英文的新聞文件，分成30個類別，每個類別中有10篇相同主題的新聞，評估方式是與人工作出的摘要做比較，每個類別的摘要以100字為限，請不同的專家對同一類別作三篇摘要，再將系統自動產生的摘要跟這三篇作比較。比較的項目可以分成重要詞彙涵蓋率、可讀性、連貫性等三類：

1. 重要詞彙涵蓋率記作「ROUGE-N」：計算自動摘要有多少N字詞與人工摘要一樣。
2. 連貫性記作「ROUGE-L」：計算自動摘要的句子中，與人工摘要有多少字彙是出現在同一句話內，由此可以判斷句子的語意是否連貫。
3. 可讀性記作「ROUGE-W」：在前一個項目「ROUGE-L」是比較同一句話內出現相同字彙個數，但是沒有考慮是否連續，在這一個項目加重有連續性的權重。

從這幾個項目可以看自動摘要的好壞。之後的實驗將按照這幾個項目去比較自動摘要的效能。

第一節 實驗步驟

第三、四章實作了自動摘要系統，其中有許多地方需要以實驗證明我們的想法，以及一些變數需要最佳化以增加系統的效能。總結上述的實作步驟，整理了一些用於實驗的變數：

步驟	變數	說明
前置處理	TF*IDF 門檻 (Threshold)	前置處理後需要過濾掉多少字彙當作候選概念
描述概念	前後文長度	3.2 節中討論描述字彙的長度，透過 ROUGE 來評估選擇的長度
	加入語意網路方法 1	3.3 節中加入語意網路來描述概念字彙，以 ROUGE 來評估
	加入語意網路方法 2	同上
分群	分群數量	K-means 分群法需要先設定 K 值，在 ROUGE 設定 100 字以內的摘要下，K 值應該如何選擇
	句子對應叢集方法 1	3.4 節中提出如何判斷句子所屬的叢集
	句子對應叢集方法 2	同上
句子權重	權重比例調整	五個特徵應該照何種比例計算才能挑出最適當的句子，以 ROUGE 來評估

表 10 預計的實驗變數說明

在這一章中所提出的數據都是依照 DUC2003 的評估工具「ROUGE」(Recall-Oriented Understudy for Gisting Evaluation)[1]來作評估，變數的優劣以「ROUGE」中 ROUGE-N、ROUGE-L 兩個值來觀察，至於每個變數所影響的內容在之前的章節都有範例以及觀察說明。

實驗方法採用簡易的貪婪演算法(Greedy Method)，貪婪演算法可以取出每一個局部最佳解，期望由此導致全域最佳解[11]。利用此種方法調整變數，在實作上比較有效，可以減少搜尋空間。演算法採用逐步向前選擇(Stepwise Forward Selection)，初始時認定沒有一個變數是良好的，每經過一次實驗將會

保留一個最好的變數，並在下一次調整變數時，保留之前調整過最好的，直到所有變數調整完畢。

第二節 實驗結果

我們最先選擇了權重比例進行最佳化，先選擇該變數的原因是希望在後面的實驗都可有一個最佳的權重比例去作實驗。調整的方法是先變換一個變數，固定其他四個，以下為調整步驟以及使用 ROUGE 評估的結果，評估數據以 ROUGE-1、ROUGE-L 兩個最常見的數值來表示。

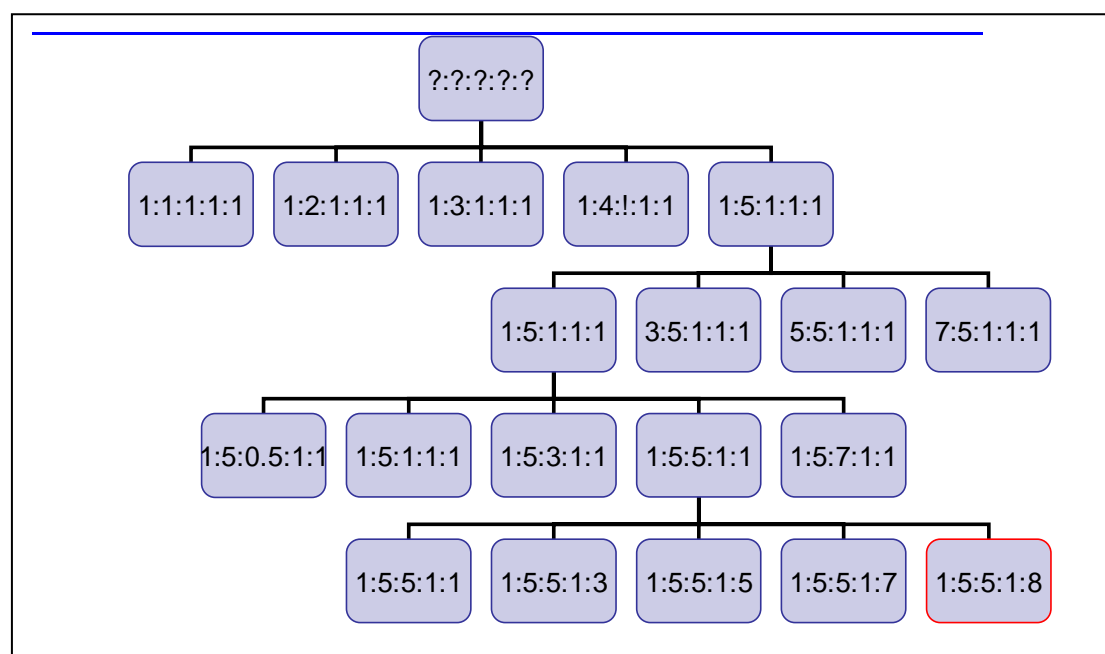


圖 14 權重比例調整 1

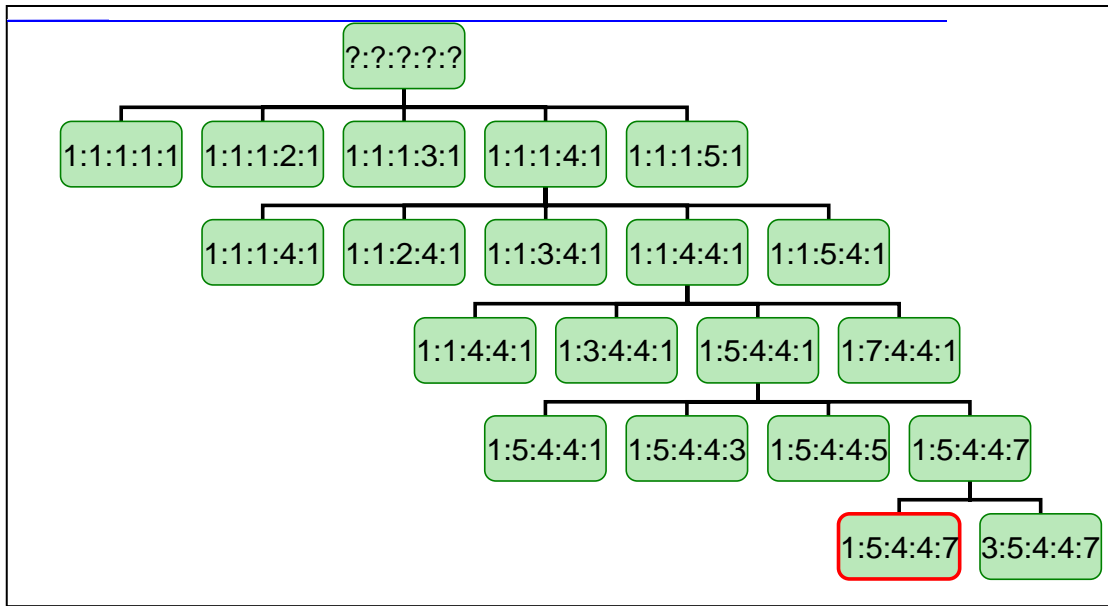


圖 15 權重比例調整 2

圖 14 和圖 15 兩個調整變數的步驟是從不一樣的特徵開始調整，希望藉由不一樣的起始點避免落入單一的區域最佳值，由圖 14 和圖 15 可以找到「1:5:5:1:8」、「1:5:4:4:7」這兩組變數對 ROUGE 而言是最佳的權重比例。表 11、表 12 為詳細的數值。

比例	ROUGE-1	比例	ROUGE-1	比例	ROUGE-1
1:1:1:1:1	0.27868	1:5:1:1:1	0.28784	1:5:0.5:1:1	0.28794
1:2:1:1:1	0.281943	3:5:1:1:1	0.28359	1:5:1:1:1	0.28458
1:3:1:1:1	0.28518	5:5:1:1:1	0.27490	1:5:3:1:1	0.28269
1:4:1:1:1	0.29130	7:5:1:1:1	0.27270	1:5:5:1:1	0.29562
1:5:1:1:1	0.29266			1:5:7:1:1	0.29709
比例	ROUGE-1	比例	ROUGE-1	比例	ROUGE-1
1:5:5:1:1	0.29983				
1:5:5:1:3	0.29794				
1:5:5:1:5	0.29284				
1:5:5:1:8	0.31785				

表 11 調整權重比例 ROUGE-1 數值，從第二變數開始調整

比例	ROUGE-1	比例	ROUGE-1	比例	ROUGE-1

1:1:1:1:1	0.27868	1:1:1:3:1	0.30127	1:1:4:3:1	0.31028
1:1:1:2:1	0.29390	1:1:2:3:1	0.29380	1:3:4:3:1	0.30095
1:1:1:3:1	0.30127	1:1:3:3:1	0.29852	1:5:4:3:1	0.29658
1:1:1:4:1	0.28986	1:1:4:3:1	0.31028	1:7:4:3:1	0.30323
1:1:1:5:1	0.28928	1:1:5:3:1	0.28897		
比例	ROUGE-1	比例	ROUGE-1	比例	ROUGE-1
1:1:4:3:1	0.31028				
1:1:4:3:3	0.29597				
1:1:4:3:5	0.30164				
1:1:4:3:7	0.29331				

表 12 調整權重比例 ROUGE-1 數值，從第四個變數開始調整

比較這一步驟的實驗中最低與最高的數據，可以發現在這一階段兩組數據分別提升 14.05%、11.33%，接下來將會固定權重比例這個變數，調整其他變數。

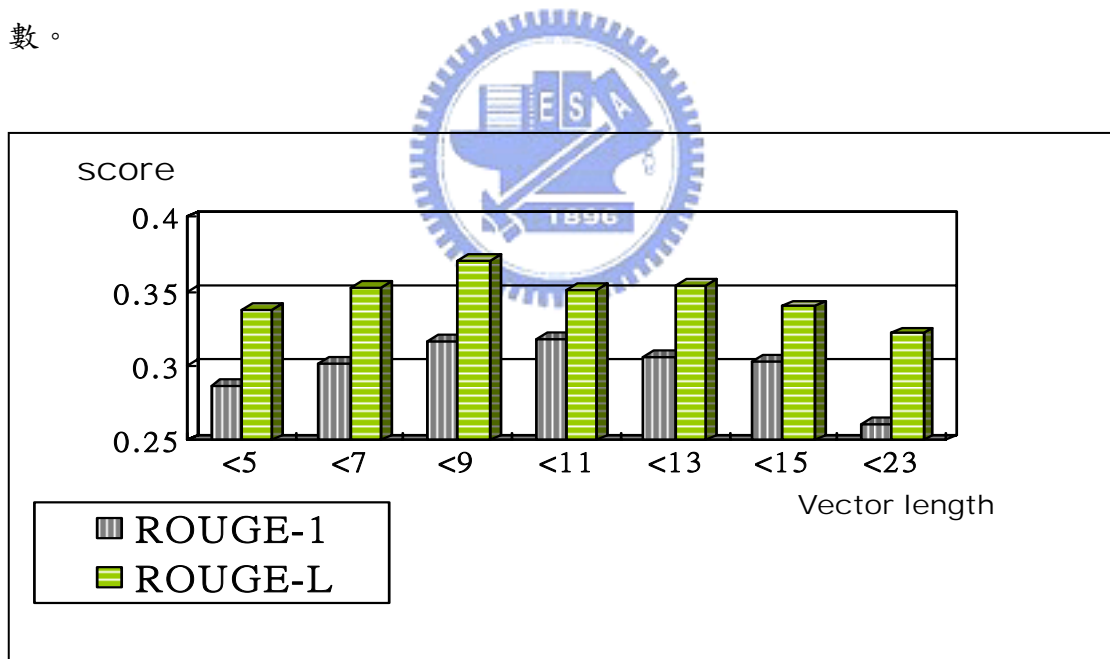


圖 16 調整向量長度變數

第二個調整的變數，是向量編成的長度，也就是用來描述概念所用的前後文。在 3.2.2 中有提到前後文的長度對描述概念的可能性，以及由參考文獻得到的適當描述長度。圖 16 中 ROUGE-1 最高情形出現在向量長度為 11 之內，ROUGE-L 最高出現在向量長度為 9 之內。評估的結果與[12]提到的資料吻合，

依照人類書寫以及閱讀習慣在看到某個字時，會記憶到前 7 ± 2 個字彙，這區間的字也最為相似，實驗結果也比其他超過區間的長度為高。

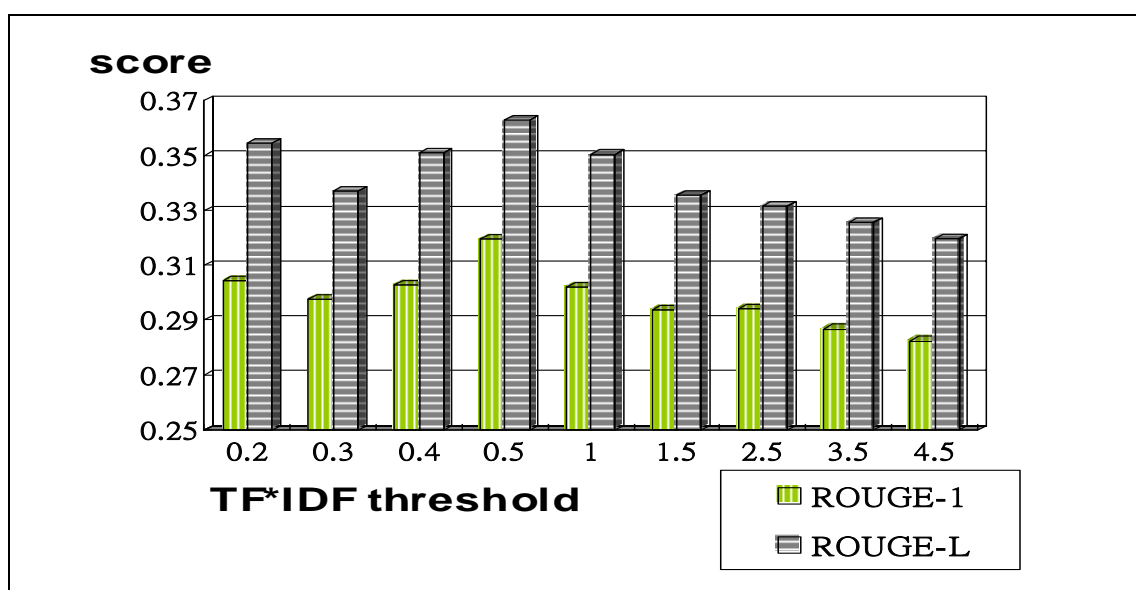


圖 17 調整 TF*IDF 門檻變數

第三個調整的變數是 TFIDF 的門檻值，這個變數將會影響候選概念的多寡以及概念分佈在文件集的密集度，由評估的結果可以得到(如圖 17)，門檻值如果設定偏高，例如設定在 4.5 時，雖然可以減少 75% 的候選字數且處理速度較快，但是因為資訊量的不足可能過濾掉應該成為摘要的候選字，所以造成評估的分數較低。降低門檻值所得到的評估分數較好，但是 TFIDF 的門檻值設定在 0.5 的時候僅過濾掉 5% 的候選字數，因此在這一個步驟中，對整體的效能並沒有太多的提升。

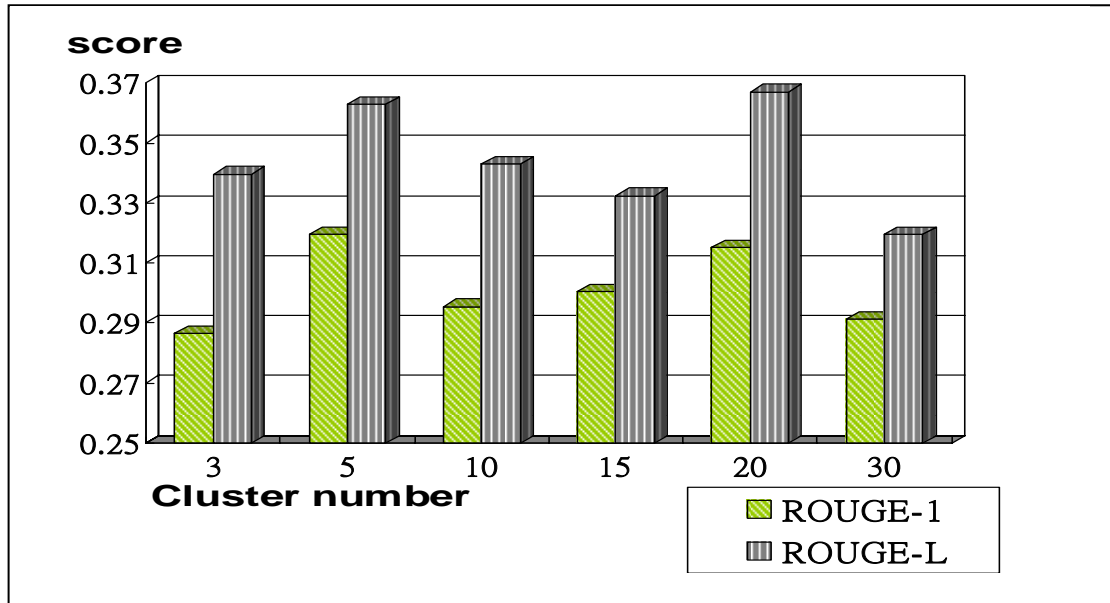


圖 18 調整分群數量變數

透過圖 18 調整分群數量變數的實作評估，用以瞭解對於 DUC2003 的新聞文件應該分為多少群比較適合，由於評估程式是針對 100 字數的短摘要，加上我們是依照叢集來挑選句子，因此叢集數目也與評估的方式有關。由圖 18 得知在 5 群時 ROUGE-1 的分數最高，在 20 群的時候 ROUGE-L 的分數最高，因此之後會分別利用 5、20 來去當作分群的數量，繼續調整下一個變數。

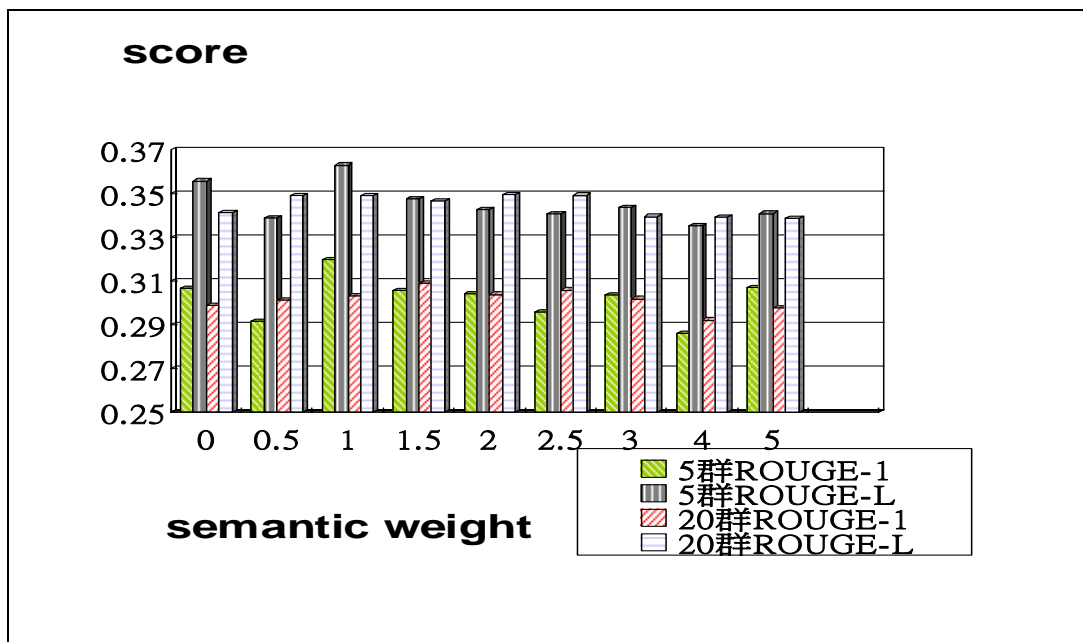


圖 19 調整加入語意網路變數

加入語意網路是為了增加描述概念的精準度，在 3.3 節中有詳細的描述，圖 19 中是以在 3.3 節中提到的方法二來調整語意網路關係的權重。從結果中得知有加入語意網路描述的結果最好的情形可以比沒有加入語意網路的改善約 7%，這個數據顯示出適當的加入語意網路是可以有效的提升摘要品質。結果中也顯示分群數目在 5 群、20 群之中的時候是互有高低，不過我們只取最高值，因此在這一結果中將取了語意網路值加”1”，分群數目取 5。

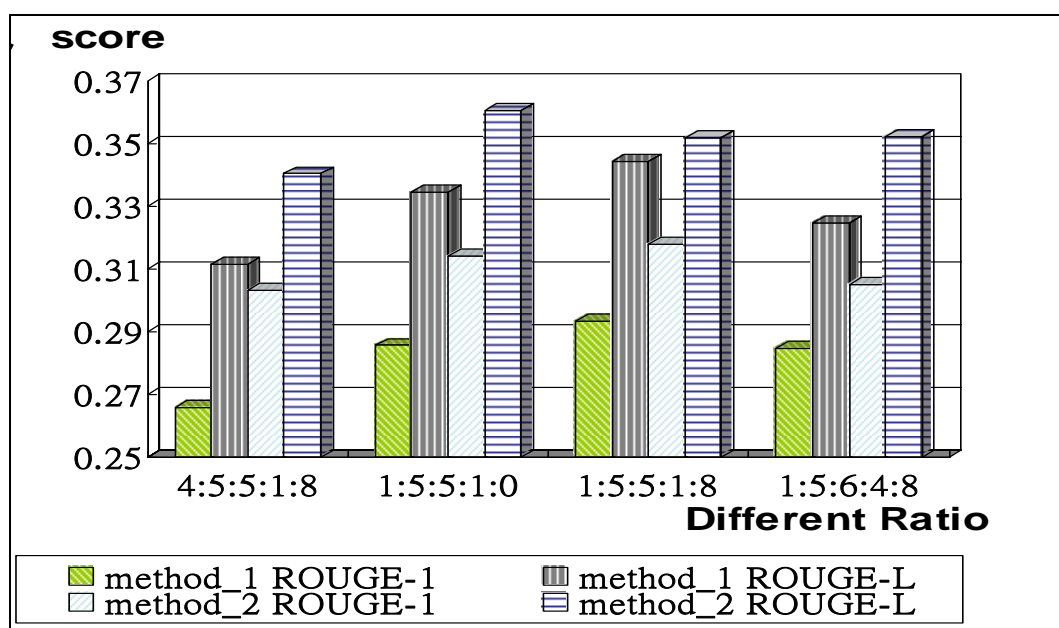


圖 20 兩種加入語意網路方法的比較

圖 20 中比較在第三章第三節中提出的兩個加入語意網路的方法，方法一是只用有同時出現在語意網路上的字彙來描述概念，方法二是使用語意網路來決定是否要增加描述時的權重。由圖 20 中的結果可以發現方法二在各種變數的情況下都比方法一要好，最極端的情況下可以相差 19.6%。原因有二，第一，方法一去描述概念時，描述的字彙會比較少，因為必須共同出現在語意網路中才可以去描述；第二，描述的字彙可能會離所要描述的概念距離過遠，在方法二中用來描述的字彙距離概念都在 4 個字的距離之內，第二點在之前的實驗也說明了使用距離過遠的字彙來描述效果並不好。

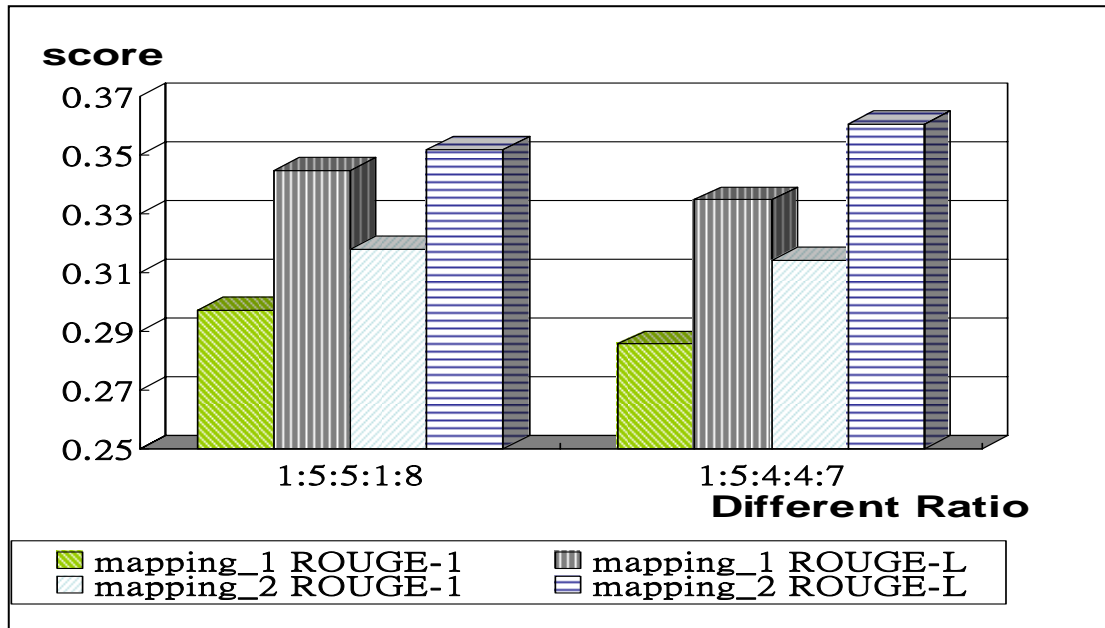


圖 21 兩種句子對應法的比較

圖 21 中比較了在第三章第四節中提出的兩個方法。方法二使用向量距離來決定句子應該對應到那個叢集，比方法一中只比對出現的字彙數量來決定要準確多了，在圖 21 的結果中也可以清楚的看出在不同的特徵比例下，方法二都較方法一好。

表 13 中列出了其他研究人員以及專家建立的摘要使用 ROUGE 進行評估的分數。

Summarizer	ROUGE-1	ROUGE-L
our summary	0.32404	0.381149
average of human summarizers	0.4030	0.4202
best summary	0.36842	0.38668
worst summary	0.23924	0.28194

表 13 ROUGE 分數比較(部分數值取自 DUC[1])

「average of human summarizers」這列的數值是以人工針對三十個新聞類別做出的摘要，經過 ROUGE 評估工具所得到的成績。可得知以人工建立的摘

要所得到的回收率(Recall)約在 40%左右，這個數值也代表不同專家對相同的文件集所摘要的內容不盡一樣，用到相同的詞的機率有 40%。本論文所設計的系統在 ROUGE-L 數據與目前其他研究結果的最佳結果幾乎是一樣的，加上 ROUGE 有些許誤差值存在，測試評估有 95%可信度(Confidence)所以並沒有什麼差距。「best summary」代表的是參加 DUC2003 年摘要比賽的結果中最佳的數據，「worst summary」則是代表比賽結果中最差的數據。在 ROUGE-1 中相差較多顯示以我們的方法在概念擷取上仍可以有進步的空間，但是在流暢度上面表現是不錯的。

類別	ID	ROUGE-1	ROUGE-L
法治類	d30003t	0.35063	0.39507
恐怖份子類	d30005t	0.27859	0.38121
中東情勢類	d30010t	0.32731	0.35775
天文景象類	d30012t	0.24443	0.27373
墜機事件類	d30016t	0.31527	0.32703
運動類	d30020t	0.23395	0.28419
選舉政治類	d30025t	0.31285	0.33994
外交類	d30028t	0.33821	0.42465
恐怖份子類	d30034t	0.29357	0.35089
外交航線類	d30040t	0.30562	0.34715
外交類	d30042t	0.29996	0.35292
外交類	d30044t	0.31581	0.38067
能源類	d30048t	0.36478	0.43923
民主政治類	d30050t	0.29632	0.35122
法治類	d30051t	0.31878	0.37256
刑事類	d30056t	0.35821	0.42973
災害類	d31001t	0.32405	0.38223
宗教類	d31002t	0.31572	0.38734
政治類	d31009t	0.34361	0.40287
天氣類	d31010t	0.34587	0.41917
天氣類	d31011t	0.36746	0.42456
刑事類	d31013t	0.37903	0.42989
災害類	d31022t	0.26978	0.30978
政治類	d31027t	0.37912	0.43031

災害類	d31028t	0.36601	0.40975
政治類	d31031t	0.28585	0.35312
科技類	d31033t	0.34481	0.39964
恐怖分子類	d31038t	0.34787	0.40612
政治類	d31041t	0.29571	0.35778
民主政治類	d31050t	0.33189	0.40396
Average		0.32403	0.381149

表 14 DUC2003 全部類別評估數值

表 14 的內容為我們自動摘要系統對每一個類別評估的分數。依照類別的不同，以及 K-means 分群法每次分群的結果皆不相同，因此最高分與最低分會有差距在 50% 左右，這顯示了影響摘要品質的變數非常多，加上所對照的人工摘要以及原文件的內容書寫有相當大的個人主觀因素在其中，也會影響摘要品質，不過整體的平均數值仍有水準之上，



第六章 結論與未來研究方向

本章總結整篇論文所提出的方法並描述未來研究方向。第一節說明本論文提出的方法對文章概念的抽取以及對概念分群的可行性，並說明方法的特色，第二節提出不足或是可擴充之處作為未來的研究方向。

第一節 結論

本摘要系統著重在兩個方面，第一為使用改良式概念描述法描述隱藏在文件中的概念；第二為對擷取出的概念進行語意分群，以解決語意歧異、語意重複的問題，並綜合上述作法以取得特徵並選取原文句子，依照合理的順序填入摘要中。

一般摘錄(Extract)方式的摘要系統在使用特徵選取原文片段時，多數取自於無語意關連的特徵，本論文的方法能在以摘錄為基礎之的情況下，提升特徵中的語意程度。



改良式概念描述法有幾項特點：

1. 與詞頻為基礎，無需事前訓練。
2. 透過共現矩陣建立語意網路，無需專家人工建立。

分群以及特徵選取特點：

1. 分群可以針對大量資料做整合或階層概念化，經過分群可以確定主題類別，並對文件集作適當的切割。
2. 特徵選取除了一般語言的重要特徵外，更增加分群結果的特徵在其中，使得在挑選句子的時候能包含更多的語意。

第二節 未來研究方向

本論文結合一般性摘錄、語意網路以及概念分群。並綜合上述加以定義所需的特徵。在實際的測試中能有效地挑選出重要的概念，但是仍有以下幾點可以改進，做為未來研究方向之參考。

1. 分群法

面對大量密集的文件時，K-Means 分群法可以有效且快速地進行分群，但是缺點在於初始化 K 個中心點時會影響到之後的分群結果，針對此一缺點可以改良分群法或是使用別的演算法來分群，例如：階層式分群(Hierarchy Cluster)、自我組織對應分群(Self-Organizing Maps, SOM)等，以避免每次產生的摘要品質變動。

2. 加入文法分析，壓縮句子

本摘要系統使用抽取原文句子的方式當作摘要，但在新聞文件中大部分的句子都是 30 字以上的長句子，對於產生 100 字之內的短摘要來說，選入的句子大都在 5 句之內，使用文法分析壓縮句子長度可以讓更多句子進入摘要，並增加摘要中的概念種類。

3. 代名詞指代(Pronominal Anaphora)

片語化的過程，仍然會有相似的片語但是被當成不一樣的片語。例如：Saudi dissident Osama Bin Laden, Bin Laden。兩個都是講恐怖份子的首腦賓拉登，但是沒有經過指代處理會被誤認是不一樣的名詞。因此代名詞指代對於以名詞當作候選概念的擷取方式，可以增加準確度。

參考文獻

- [1] DUC2003(Document Understanding Conferences). Available as
<http://www-nlpir.nist.gov/projects/duc/guidelines/2003.html>
- [2] NLPprocess- Text Analysis Toolkit. Available as
<http://www.infogistics.com/textanalysis.html>
- [3] MUC(Message Understanding Conferences). Available as
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/
- [4] Infogistics, POS -tag stands. Available as <http://www.infogistics.com/tagset.html>
- [5] Information Mapping Project. Available as <http://infomap.stanford.edu/>
- [6] The Porter Stemming Algorithm. Available as
<http://tartarus.org/~martin/PorterStemmer/>
- [7] R. Angheluta and R. De Busser and M.-F. Moens, “The Use of Topic Segmentation for Automatic Summarization,” In *Proceedings of the ACL-2002 Post-Conference Workshop on Automatic Summarization*,2002.
- [8] R. Barzilay and M. Elhadad, “Using Lexical Chains for Text Summarization,” *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997 Page(s): 10 - 17
- [9] W. Lam and K. S. Ho, “FIDS: an intelligent financial Web news articles digest system,” *IEEE Transactions on Systems, Man and Cybernetics, Part A*, Volume 31, Issue 6, Nov. 2001 Page(s):753 – 762
- [10] C. N. Silla Jr. and C. A. A. Kaestner and A. A.Freitas, “A Non-Linear Topic Detection Method for Text Summarization Using Wordnet,” *Workshop of Technology Information Language Human (TIL'2003)*, 2003.
- [11] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2000

- [12] F. Chen and K. Han and G. Chen, "An Approach to Sentence-Selection-Based Text Summarization," *IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, (TENCON '02) Volume 1*, Oct. 2002 Page(s):489- 493
- [13] J. Lin and E. Keogh and W. Truppel, "Clustering of Streaming Time Series is Meaningless," *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, CA, USA, 2003 Page(s): 56-65
- [14] D. McDonald and H.C. Chen, "Using sentence-selection heuristics to rank text segment in TXTRACTOR," *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*, Portland, Oregon, USA, 2002 Page(s): 28 - 35
- [15] 曾元顯, "中文手機新聞簡訊," 第十六屆自然語言與語音處理研討會, 台北, 2004年9月2-3日, 頁177-189.
- [16] Penn TreeBank. Available as www.cis.upenn.edu/~treebank/home.html
- [17] L. Vanderwende and M. Banko and A. Menezes, "Event-Centric Summary Generation," In *Document Understanding Conference at HLT-NAACL*, Boston, MA, 2004
- [18] SVDPACK. Available as <http://www.cs.utk.edu/~berry/projects.html>
- [19] U. Y. Nahm and R. J. Mooney, "Text Mining with Information Extraction," In *Proceedings of the AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 2002
- [20] 葉鎮源(2002), "文件自動化摘要方法之研究及其在中文文件的應用," 碩士論文, 國立交通大學資訊科學研究所, 新竹, 2002
- [21] C. S. Lee, Z. W. Jian and L. K. Huang, "A Fuzzy Ontology and Its Application to News Summarization," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics : Accepted for future publication* Volume PP, Issue 99, 2005 Page(s):859 - 880