以概念分群爲基礎之新聞事件自動摘要

劉政璋
國立交通大學資訊科學系
gis92573@cis.nctu.edu.tw

葉鎮源
國立交通大學資訊科學系
jyyeh@cis.nctu.edu.tw

柯皓仁
國立交通大學圖書館
claven@lib.nctu.edu.tw

楊維邦
國立交通大學資訊科學系
國立東華大學資訊管理系
wpyang@mail.ndhu.edu.tw

摘要. 新聞事件自動摘要乃針對敘述相似事件的多篇新聞文章編製摘要內容，其目的爲幫助讀者過濾資訊並快速瞭解事件的來龍去脈，以節省閱讀大量新聞文件的時間，主要的研究議題爲偵測不同新聞文章中相似及相異的內容，以達到過濾重複資訊的目的。本論文以概念分群(Concept Clustering)爲基礎，偵測新聞事件所要表達的語意，進而挑選涵蓋豐富語意的語句爲摘要。過程爲：1) 利用前後文關係(Context)及語意網路(Semantic Network)，描述概念詞彙；2) 使用 *K*-Means 對概念詞分群，期萃取更精確的概念，同時解決語意歧異的問題；3) 根據概念分群結果，並利用語句資訊量、語句位置及語句概念等特徵，計算每個語句之重要性，最後挑選重要性高的語句作爲摘要內容。實驗中使用 DUC 2003 (Document Understanding Conference) 所提供的新聞事件進行評估。評估結果於 ROUGE-1 指標比平均成績還好，於 ROUGE-L 指標接近最好的結果。

1. 前言

近年來，由於電腦科技的迅速發展及網際網路的推波助瀾，資訊陸續被數位化，以利於網路流傳。數位化的發展亦使得資訊量大幅增加，使用者在獲取資訊上不再受限於少數的流通管道，反而可以輕易取得大量資料。在這種現象下，困難的反而是如何過濾掉不需要及重複的資訊，使得使用者可以快速找到真正所需的資訊。爲解決前述困難，可利用摘要系統的精簡性及去重複性，減少使用者閱讀時間，幫助使用者於短時間判斷及取得重要資訊。

文件摘要技術的作法大致可分爲兩類。第一類使用專業領域知識(Domain Knowledge)分析文章中的人、事、時、地、物等要素，第二類則是以統計分析(Statistical Analysis)方法直接從原文判斷語句重要性。使用專業領域知識來達成摘要系統可有效抽取出文件內的主題，但是需要事前由領域專家介入建立領域知識，包括語言知識、文件主題背景知識等，而自動化建立領域知識的方法目前還很難突破。本論文以資訊擷取(Information Retrieval)技術爲基礎，導入語意網路(Semantic Network)，同時改良原文抽取摘要語句的方式，提出一套描述概念(Concept)且能分辨語意(Semantics)的概念偵測方法。利用不同的概念，便可找出新聞事件中具豐富概念及語意的語句，達到產生符合原文主題摘要的目的。

本文中提出以概念分群(Concept Clustering)抽取新聞事件所提及的主題(Topic)及語意，並結合傳統特徵選取法(Feature Selection)計算語句的重要性及語意涵蓋度，藉此作爲挑選摘要語句的參考依據。以下簡單說明本文所提之摘要方法的流程：1) 利用前後文關係(Context)及語意網路

(Semantic Network)，描述概念詞彙；2) 使用分群法(本文採用 *K*-Means [11])對概念詞分群，以萃取更精確的概念，同時解決語意歧異的問題；3) 根據概念分群結果，利用語句資訊量、語句位置及語句概念等特徵，計算每個語句的重要性，最後挑選重要性高的語句作爲摘要內容。

本文共分成六節。第二節介紹與多文件自動摘要相關的研究；第三節介紹結合前後文與語意網路的概念描述法，並說明如何進行分群偵測及抽取概念；第四節針對先前選出的概念找出對應的語句，並以數個特徵值進行語句的權重分析；第五節說明實驗結果的分析討論，以驗證本文所提方法的可行性；最後一節是結論與未來可繼續發展的方向。

2. 相關研究工作

本節介紹幾種多文件摘要技術。MEAD [18]接受分群過後的文件集[1]，併考量以下三個特徵：1) 語句與群中心(Centroid)的相似度；2) 語句於文件中的位置，通常出現於文件首句的語句，可加重其重要性；3) 語句與所屬文件之首句的相似度。MEAD以線性組合(Linear Combination)結合上述三種特徵，綜合評估語句重要程度。一般而言，MEAD使用的首句加重計分法，比較適用於新聞文章[2]；如果文件集是爲其他領域，例如技術類的文件，則首句加重計分法要再調整才合適。

McKeown et al. [17]認爲主題相關的文件集中，存在有許多不同的小主題(Theme)。他們的方法，分爲三個部分：1) 主題辨識(Theme Identification) [8]以語句爲最小單位，透過分群技術將文件中的主題抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [3]將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) 將所摘錄出來的重要字詞重新組合以產生流暢的摘要。他們主要考慮以下特徵以決定兩段落的相似度，進而利用分群法將找出主題，即相似段落的集合：

- Word co-occurrence：假如兩個段落有許多相似的字，則可視爲相似。
- Matching noun phrases：利用 LinkIt [26]判斷是否擁有互相關聯的名詞片語群組。
- WordNet synonyms：使用 WordNet [27]找出同義詞組。
- Common semantic classes for verb：判斷具有同一語意的動詞詞組。

接著利用 Information Fusion 的技術，從主題中萃取出具有代表性的詞組或片語。同時依照出現在文章中的次序，對片語排序。最後，藉由 FUF/SURGE [9]自然語言產生器生成完整語句。

MMR (Maximal Marginal Relevance) [4]適用於單文件摘要，其概念乃是對所挑選出與 Query 相關的語句重新排序，以符合具有最大相關度及最大差異度的特性。MMR-MD [10]延伸 MMR 的概念，可有效降低摘要中具有相同涵義的語句(即，減少重複性資訊)。MMR-MD 同時考慮到時間順序、專有名詞、對主題的相似度以及代名詞的 Penalty。其挑選段落的依據如下：

---

[1] MEAD 接受相關的文件集，以產生摘要。然此處所提及之相關文件集，實爲考慮 loosely-related documents。
[2] 此類文章通常於第一段第一句說明整篇文章的重點。因此，首句之重要性必須加重考慮。

$$MMR-MD \overset{def}{=} Arg \max_{P_{ij} \in R/S}[\lambda Siml(P_{ij}, Q, C_{ij}) - (1-\lambda) \max_{P_{ij} \in S} Sim_2(P_{ij}, P_{nm}, C, S)]$$

公式 1: MMR-MD [10]

其中，$Sim_1(P_{ij},Q,C_{ij})$計算 P_{ij} 與 Q 的相似度，同時衡量與段落所在的文件群的相關度；
$Sim_2(P_{ij},P_{nm},C,S)$計算 P_{ij} 與 P_{nm} 的相似度，其中 P_{nm} 爲一以挑選出之段落。

MMR-MD 目的在於使摘要中的段落儘可能的相似於 Query，但其所選到的段落間要儘可能的不相似。由於與 Query 相似度高的段落，彼此之間的重複性可能也高，而與段落相似度稍低的段落，彼此之間的重複性可能也低，透過適當的 λ 值可以找到兼具主題但又不會有過多重複性的段落爲摘要。

Mani et al. [15]將文件表示成圖形(Graph)，其中，每個節點代表一個關鍵詞(Term)，節點與節點間用不同的關係連接起來，包含 1) 片語關係(PHRASE)；2) 形容詞關係(ADJ)；3) 同義關係(SAME)；4) 關聯關係(COREF)。首先，賦予每個節點一權重(Weight)，權重值初始爲該關鍵詞的 TF-IDF 值。接著，利用 Spreading Activation 演算法，透過節點間相連的連結權重變更節點的權重值，以找出與 Query 相關的節點。接著，比較兩兩文件圖形模型的相似度(Commonality)及差異性(Difference)。他們提出 FSD (Find Similarities and Differences)演算法，以找出兩圖形中相似或差異的節點。最後，透過分析 Similarities 及 Differences 集合中的關鍵詞，計算語句的重要性，並挑選重要的語句爲摘要結果。

3. 概念分群及抽取

本節說明如何以統計方法及分群技術由新聞文件集中推導出事件概念群[3]。首先介紹如何選取重要的概念詞，並利用概念詞的前後文(Context)與語意網路(Semantic Network)作爲其描述；接著說明利用分群法將概念詞分群，以導出新聞事件中的主題。

圖 1 說明本文所提方法之架構。步驟一爲前置處理；步驟二挑選具有代表性的名詞及名詞片語當作候選的概念詞；步驟三根據候選概念詞之前後文及事先建立的語意網路，描述該候選概念詞，以得到一向量表示式；步驟四針對候選概念詞作分群，可得到概念相似之概念群；步驟五依據語句中關鍵詞的資訊，將語句對應到概念群中，得到語句與概念群的關連；步驟六根據語句與概念群的關聯，同時考量文章結構的關係，計算 3 個與語句相關及 2 個與概念群相關的特徵值；步驟七則依據步驟六計算之特徵值，以線性組合的方式，計算位於同一概念群中語句的重要性，該重要性可作爲摘要語句挑選的依據。

[3] 概念爲單一或多個字詞所組成的集合。此集合可視爲一個概念性的描述，並定義該概念的範圍。透過此集合，可作爲系統理解概念語意的媒介。

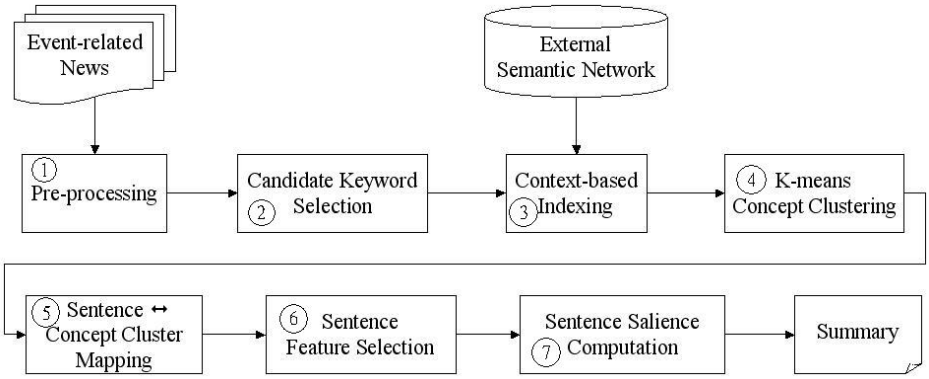

圖 1: 系統架構圖

3.1. 結合前後文與語意網路的概念描述法

首先，進行前置處理(Preprocessing)，其作用爲避免雜訊干擾，降低統計數據的參考性。此步驟包含斷詞切字(Tokenization)、詞性標記(Part-Of-Speech)、詞幹還原(Stemming)、小寫化(Lowercasing)、刪除停用字(Stopword Removing)及片語化(Chunking)等。本文中，斷詞切字及詞性標記採用 NLP Processor [20]；詞幹還原採用 Porter 演算法[24]；片語化則利用統計方法計算詞性組合機率，以辨別是否爲可組合片語；停用字的部份針對 DUC 2003 所提供的文件集設計，共有 309 個字，其中絕大多數爲介係詞、指代詞、連詞及助詞。

前置處理後僅保留名詞(Noun)及名詞片語(Noun Phrase)作爲可能的候選概念(Concept Candidate)，原因乃是名詞比其他詞性含有更多語意[1] [13]。本論文同時計算每個候選詞的 *tf-idf* 值[28]，進一步過濾不具代表性的字詞。最後，再由概念候選詞中挑選一般名詞、複數名詞、專有名詞、複數專有名詞等字詞，作爲最後所保留的概念候選詞集合。

接著說明如何導出概念候選詞的表示法。[2]提到絕大多數描述同一事件所伴隨出現的字詞，其語意皆很相似。[5]亦提到除考慮單一字詞的重要性外，更不能忽略出現在重要詞彙前後文的影響力；例如，condemn(譴責)及 intensively(強烈)經常一起出現，此兩關鍵詞可用來描述彼此。基於這個想法，本研究利用候選概念詞的前後文描述該字詞。作法上以出現在候選概念詞前後分別爲 *N* 及 *M* 個字作爲描述字彙集合，同時限制挑選的範圍爲一個完整的語句。在 *N* 與 *M* 的設定上，由於在[5]中提及人類的短暫記憶通常爲 7±2 個字詞，因此，實作上設定 *N* 及 *M* 各爲 5。取最小值最主要是要讓前後文的涵蓋範圍小一點，使相鄰的不同概念在描述的內容不致於有過多重複，以免影響到概念分群的結果。

表 1 爲下例中「the U.S. Embassy」的描述法，以 several locations, Bonn, receiving, word, a terrorist threat, the U.S. Embassy, no evidence, a planned attack, found, officials, Wednesday 作爲索引詞(Indexing Description)。此描述概念的方式是希望利用前後文的關係，讓概念的語意更加明顯，以方便進行分群的時候能更精確計算兩兩概念的相似度。

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

表 1: 以前後文描述後選概念範例[4]

Concept	Indexing Description	Length
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

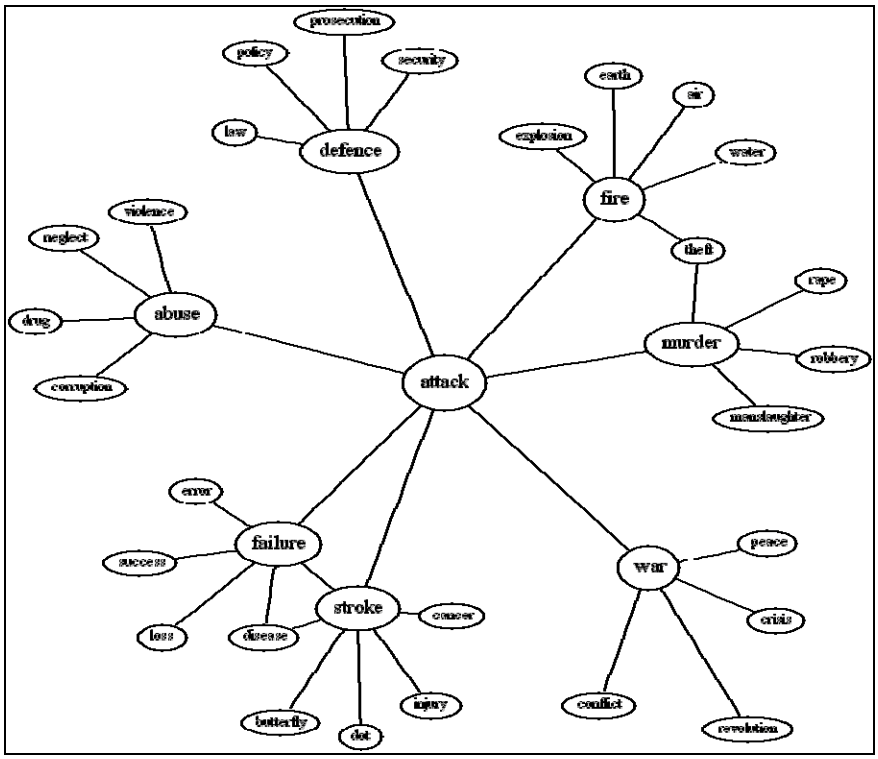

圖 2: 以 attack 爲中心之語意網路範例[12]

爲了加強描述字詞的語意，本研究在以前後文描述概念時亦加入語意網路。我們使用 Informap [12]建立新聞事件的語意網路，Infomap 以共現 (Co-occurrence) 的原則來判斷字與字之間語意的相關性。首先，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(Window)，在這個範圍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。Infomap 利用共現矩陣，並使用奇異值分解(Singular Value Decomposition)降低字向量的維度， 最後計算兩兩詞間的相似度，並建立語意網路。圖 2 爲一以 attack 爲中心的語意網路範例，透過語意網路，可找出與 attack 語意相關的字詞。例如，直接相關的字詞有 defense、abuse、failure、stroke、war、murder 及 fire。

本論文提出兩種整合語意網路於前後文描述的方法。第一種方法是在前後文中只取與概念於語意網路中有關聯字詞，以美國大使館(the U.S. Embassy)此一概念爲例，於表 1 中，描述字詞的集合包含 Bonn、word、found、officials、Wednesday 等五個字彙，然而，透過語意網路的連接分

[4] 爲方便說明，此例及本文中所提及之 Indexing Description 皆列出未經過前置處理的關鍵詞。

析，發現「the U.S. Embassy」與上述五個字並沒有相關，因此在表 2 中便去除此五個字彙。由此可看出加入語意網路後，可消除多餘的字詞，使得描述概念的字詞更精確。

表 2: 加入語意網路方法 1 的範例

Concept	Indexing Description	Length
the U.S. Embassy	several locations, receiving, a terrorist threat, the U.S. Embassy, no evidence, a planned attack	6

第二種方法則是希望能夠突顯與概念語意有關的字詞，且不至於影響到原本描述字詞的組成，我們保留了原始用以描述概念的所有前後文字彙，但加重在語意網路上與概念相關的字彙。以美國大使館(the U.S. Embassy)此一概念爲例，由於 several locations, receiving, a terrorist threat, the U.S. Embassy, no evidence, a planned attack 爲語意網路中與 the U.S. Embassy 有直接相關的關鍵詞，因此在表 3 中將其 TF-IDF 值加上一常數 X，以達到加重權重的目的。

由上述兩例可知，方法一與方法二的差別爲，方法一根據語意網路刪除不重要的索引詞，而方法二則是保留所有的索引詞，但加重在語意網路中與概念相關詞之權重值。另外，方法一的描述雖然比較能夠貼近概念的語意，但其描述的字彙分佈比較散，且去除了不在語意網路內的字彙，使得描述的字彙數目少於原本的描述字彙，因此，方法一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。

表 3: 加入語意網路方法 2 的範例

Concept	Indexing Description
the U.S. Embassy	several locations (5.1705+***X***), Bonn (5.1705), receiving (2.9733+***X***), word (5.1705), a terrorist threat (5.1705+***X***), the U.S. Embassy (2.9733+***X***), no evidence (5.1705+***X***), a planned attack (5.1705+***X***), found (3.5611), officials (4.4773), Wednesday (2.9733)

3.2. 利用分群技術抽取主題概念

本文中分群的對象，是經過 3.1 處理後之概念向量，分群方法則採用 *K*-Means [11]。考量新聞事件可再細分爲地點、對象、影響結果等特性，分群的結果可視爲文件中所提及的主題概念。

概念分群之後，便要將語句與概念群作連結，以期找到能夠代表每個語句的概念群(亦即，該語句所要表達的語意及相關主題)。本文提出兩種對應方法。第一，判斷語句中的字詞出現於哪個概念群中的字數最多，則歸類到該概念群。第二，判斷語句中的概念出現在哪個概念群中的字數最多，則歸類到該概念群。公式 2 爲語句對應到概念群的判斷依據。

簡單的說，第一種方法只單純判斷語句中有多少字出現在該概念群裡，概念群中原本只包含概念，然而本文亦嘗試把描述概念的字彙也加進概念群裡；這樣的作法是希望能夠增加語句對應到字詞的數量，避免一句話裡只有少數幾個字詞出現在概念群內，且對應的字詞數量越多，越容易判斷語句屬於哪一概念群。第二個方法判斷語句中的概念在哪個概念群中，由於概念是以編成向量的方式做 *K*-Means 分群，每個向量都可以找出與所屬概念群的相似度，也就是離中心點的距離。當語句裡有向量出現在概念群之中時，會以該向量離中心點的相似度當作該語句跟此概

念群的相似度。

(1)$SIM_{s,i}$ = Words Match
= sim (Match_Word, $Cluster_j$) / L_of_S

(2)$SIM_{s,i}$ = Concepts Match
= sim (Match_Vector, $Cluster_j$) / L_of_S

Match_Vector：concept vector included in this sentence
sim (Match_Word, $Cluster_j$)：number of word appear in $cluster_j$
sim (Match_Vector, $Cluster_j$)：distance between vector and centroid of $cluster_j$
L_of_S：length of the sentence

公式 2: 語句對應到概念群的方式

比較上述兩個方法，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語句都可以找到對應到的概念群，造成了每一個概念群內的語句數量多，但是語句的語意可能不是與概念群的概念高度相似，造成此現象的原因可能爲只對應到描述概念的字彙，並不是對應到概念本身。方法 2 的對應則可以有效的過濾掉語意不符合概念群的語句，雖然對應後每個概念群包涵蓋的語句數目較少，雖然剩下的語句數量較少，但是再經由後面的特徵選取時，亦可有效提升選取適合摘要語句的效率。

4. 語句語意權重摘要

透過概念群及其中概念字詞與語句關鍵詞的相似關係，可將每個語句對應至語意相近的概念群。然而，位於同一概念群中的語句彼此語意近似，仍需要透過其他條件以判斷哪個語句最能代表該概念群。由語句特徵挑選重要語句的方法在很多研究中被提出來，藉由抽取不同的特徵，可以整合這些特徵以判斷語句的重要程度[16]。本文考量 3 個與語句相關及 2 個與概念群相關的特徵計算位於同一概念群中語句的重要性。

4.1. 語句相關特徵

- TF*IDF

  考慮語句中所有字詞的 TF*IDF 總和，並除以語句長度以正規化(Normalization)。

$$S_{tfidf} = \left(\sum_{i=n}^{m} TF \times IDF_i \right) \Big/ sentence\_length$$

- 語句於文件中出現的位置

  位於首句或尾句的語句通常具有關鍵性的語意資訊[5]。因此，當語句位於此位置時，則加重此語句的權重。

- 語句與所屬的概念群的相似度

  3.2 節提到兩個不同的對應方式，分別爲比較語句與概念群中共同出現的字彙數量及比較語

句所包含概念與概念群的相似度。本文針對此兩種對應方式，提出不同計算相似度的方法。

方法一：採用所對應到的字彙數目計算相似度，並除以語句長度作正規化，如公式 3。然而，由實驗中發現以這種方式來計算相似度，會發生有很多語句所對應到的字彙數量是一樣的情形，導致這個方法所計算出的權重不具有辨別性。

$$S_{sim} = match\_words_i / sentence\_length$$

match_words: 計算與概念群 *i* 內有多少字彙是一樣的
i: 語句所對應到的概念群 *i*
sentence_length: 語句的長度

公式 3: 相似度特徵計算方法 1

方法二：相似度的計算取決於向量對應的概念群與其中心點的距離，如公式 4。此方法可比較哪些語句比較接近該概念群的中心點。在多維度的向量中，使用歐基理得距離(Euclidean Distance)可更精確地找出哪些向量接近中心點，每個語句將可以更清楚地分出代表概念群的重要性。

$$SIM_{s,j} = \frac{1}{\sum_{i \subset s} \left(dis\tan ce(concept_i, cluster_j) \right) \times L_s}$$

concept: 語句有對應到概念群的概念
distance(concept, cluster): 取向量到概念群中心的距離
L_s: 語句長度

公式 4: 相似度特徵計算方法 2

4.2. 概念群相關特徵

- 概念群內含的概念多寡

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群。當包含越多概念的群，其權重應該越高[5]。

- 概念群與中心點的距離

分群的結果，依照向量的分佈情形可以找出全部向量的中心點。每一個概念群中越靠近中心點的給予越高分。在中心點附近的概念群，越有可能涵蓋其他概念群的意思，在順序上應該要比其他遠離中心點的概念群要重要，亦能加強涵蓋性越大的概念群重要性。

4.3. 語句重要性

綜合上述的五個特徵可以得到一個權重總和，如公式 5。「C_{length}」爲概念群內的向量個數；「S_{tfidf}」爲語句內字彙的 TF*IDF 總和；「$C_{distance}$」爲語句所屬概念群距離全體向量質心的距離倒數；「$S_{location}$」爲語句所在位置；「S_{sim}」爲語句與所屬概念群的相似度。

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

公式 5: 計算語句權重總和公式

5. 實驗結果分析與評估

自動摘要的成效評估，可分爲直接(Intrinsic)與間接(Extrinsic)評估兩種方式[16]。直接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。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。本摘要系統使用的測試文件集爲 DUC 2003 (Document Understanding Conferences 2003) [6]，文件內容是英文的新聞文件，分成 30 個新聞事件，每個事件中約有 10 篇相同主題的新聞，DUC 2003 並請不同的專家對同一類別作三篇摘要。評估方法是將系統自動產生的摘要與 DUC 2003 的專家所作出的摘要比較，每個事件的摘要以 100 字爲上限。效能評估採用 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [22]，主要比較的項目爲 ROUGE-N、ROUGE-L，分別代表「自動摘要有多少 N 字詞與人工摘要一樣」及「自動摘要與人工摘要有多少字彙是出現在同一語句」。

本文所提出的方法有諸多變數需要最佳化以調整系統效能，表 4 列出所有可能的變數。

表 4: 實驗變數說明

步驟	變數	說明
描述概念	前後文長度	描述字彙的長度(即 *N* 與 *M*)
	加入語意網路方法	加入語意網路後描述概念字彙的方法
分群	分群數量	*K*-means 分群法需要先設定 *K* 值
	語句對應概念群方法	如何判斷語句所屬的概念群
語句重要性	權重比例調整	五個特徵以何種比例計算才能挑出最適當的語句

首先對權重比例進行最佳化，先選擇該變數的原因是希望之後的實驗都可有一個最佳的權重比例。調整的方法是先變換一個變數，同時固定其他四個。最後我們所調整出計算語句重要性之特徵權重比例 $\alpha:\beta:\gamma:\theta:\lambda$ 爲 1:5:5:1:8。

圖 3 調整的變數是概念向量的長度，也就是用來描述概念所用的前後文長度。評估的結果發現 ROUGE-1 最高情形出現在向量長度爲 11 之內，ROUGE-L 最高出現在向量長度爲 9 之內，與[5]提到的資料吻合。亦即，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，這區間的字也最爲相似，實驗結果也比其他超過區間的長度爲高。圖 4 爲調整分群數量變數的實作評估，由圖 4 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 作爲分群的數量。圖 5 爲調整語意網路關係權重的結果，在最好的情況下，加入語意網路可以比沒有加入語意網路改善約 7%。這個數據顯示出適當地加入語意網路可以有效地提升摘要品質。結果中也顯示分群數目在 5 群、20 群時互有高低，不過我們只取最高值，因此在這一結果中決定將語意網路加重之常數值 *X* 爲 1，分群數目則設定爲 5。

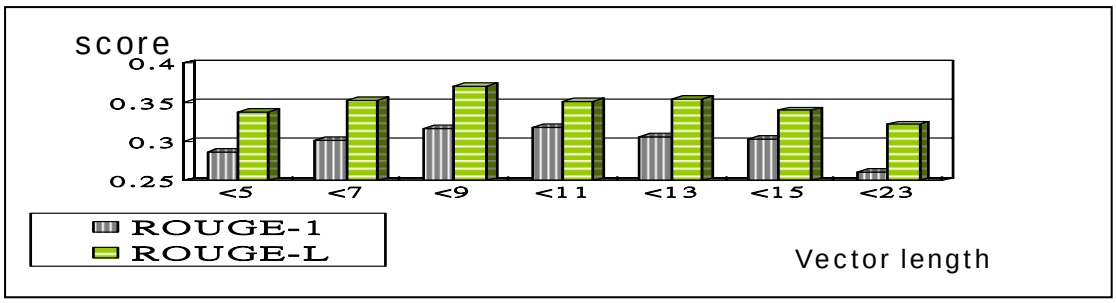

圖 3: 調整概念向量長度變數

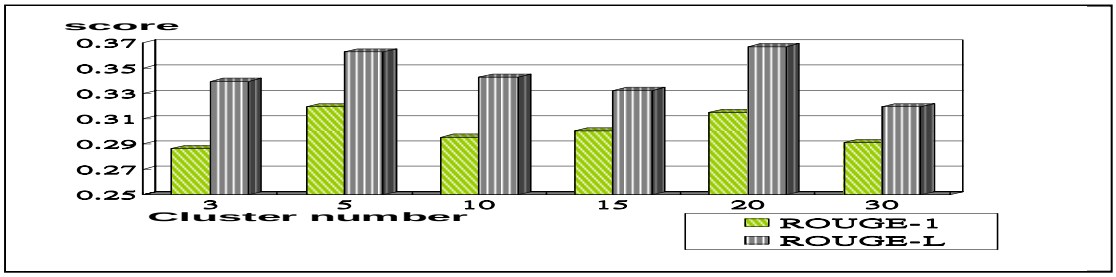

圖 4: 調整分群數量變數

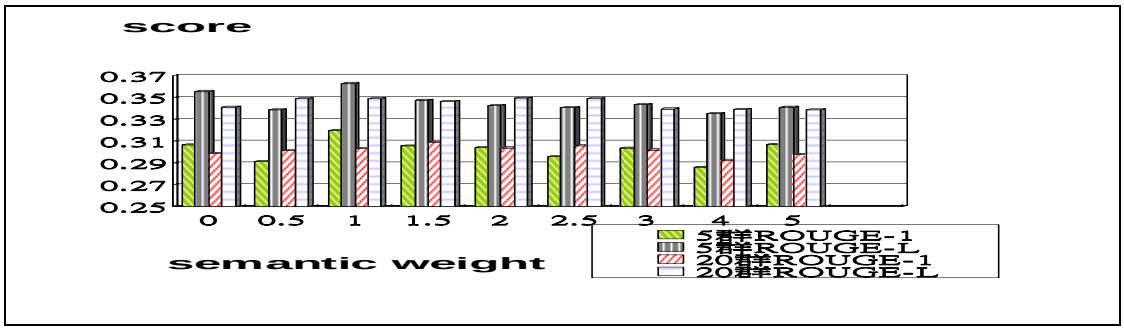

圖 5: 調整加入語意網路變數

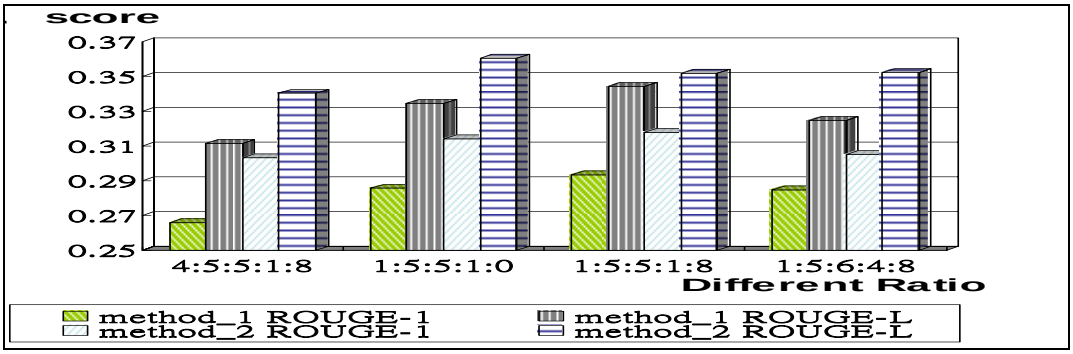

圖 6: 兩種加入語意網路方法的比較

圖 6 比較在 3.1 中提出的兩個加入語意網路的方法，方法一是只用有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字彙的權重。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%。推估原因有二：第一，以方法一描述概念時，描述的字彙會比較少，因爲描述概念的字彙必須與概念共同出現在語意網路

中；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 7 中比較語句對應到概念群的兩個方法。方法二使用向量距離來決定語句該對應到哪個概念群，方法一是只比對出現的字彙數量來決定對應到哪個概念群。從圖 7 可以觀察出在不同的特徵比例下，使用方法二的效果較好。

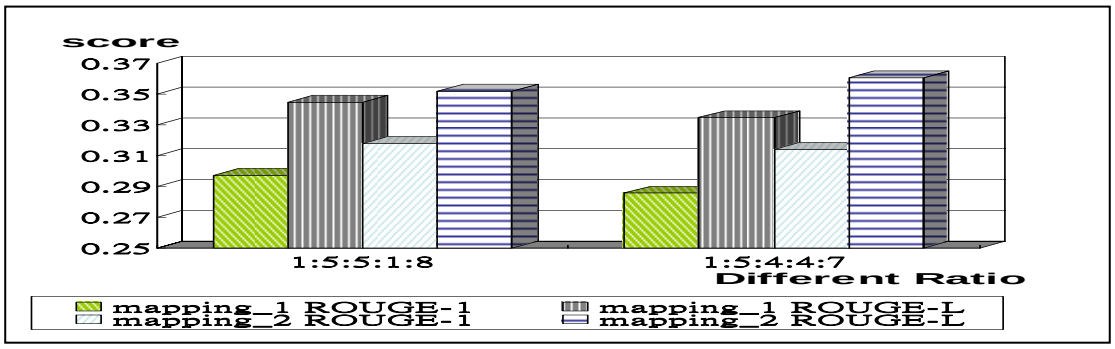

圖 7: 兩種語句對應概念群方法的比較

表 5 列出參加 DUC 2003 的其他系統及專家建立的摘要使用 ROUGE 進行評估的分數。「average of human summarizers」這列的數值是以人工針對三十個新聞事件做出的摘要，經過 ROUGE 評估工具所得到的成績，亦即評估專家間對於摘要內容看法的一致性。由此可知，以人工建立的摘要所得到的召回率(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及觀點不盡一樣。「best system」代表的是參加 DUC 2003 年摘要比賽的結果中最佳的數據，「worst system」則是代表比賽結果中最差的數據。「average of all DUC 2003 systems」則是所有參加 DUC 2003 系統的平均值。本論文所設計的系統在以 ROUGE-1 評比時位於平均以上，但是當考慮 ROUGE-L 時，本文所提的方法則非常接近於最佳的系統。

表 5: ROUGE 分數比較 (部分數值取自 DUC 2003 [6])

Summarizer	ROUGE-1	ROUGE-L
our system	0.32404	0.381149
average of human summarizers	0.4030	0.4202
best system in DUC 2003	0.36842	0.38668
average of all DUC2003 systems	0.31102	0.34652
worst system	0.23924	0.28194

6. 結論與未來研究方向

本文提出兩個新的想法，第一爲使用前後文資訊及語意網路描述隱藏在文件中的概念；第二爲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。同時，以概念分群爲基礎，並考量語句特徵，來計算語句的重要性，以挑選重要性高的語句作爲摘要內容。本研究所提出的技術有下列幾項特點：1) 以詞頻爲基礎，無須事前訓練；2) 透過共現矩陣建立語意網路，無須專家以人工建立；3) 分群可擷取重要主題概念，並對應語句與概念群關聯；4) 特徵選取包含一般性特徵(Surface Feature)以及加入概念分群的語意特徵。

未來，我們希望針對以下幾點作改進。首先，本論文以擷取語句爲基礎，然而測試文件集中大多爲長語句，以產生 100 字內的短摘要而言，僅能挑選到約 5 個語句，若考量語句壓縮，可納入更多語句，增加摘要內容的多元性。第二，*K*-Means 分群會因爲 *K* 值的不同而影響摘要品質，未來可考慮不同的分群法，如階層式分群(Hierarchical Clustering)。最後，片語化的過程，仍然會有相似的片語卻被當成不一樣的片語，例如 Saudi dissident Osama Bin Laden 與 Bin Laden 皆爲恐怖份子首腦賓拉登，但是沒有經過關聯及指代(Anaphora)處理會被誤認是不一樣的名詞，因此若加入指代關係處理，相信對於以名詞當作候選概念的擷取方式，可以增加準確度。

致謝

本研究由國科會計畫 92-2213-E-009-126-及部份由 93-2213-E-009-044-補助。

參考文獻

[1] 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.
[2] R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization," *ACL/EACL Workshop on Intelligent Scalable Text Summarization*, 1997 Page(s): 10 – 17.
[3] Barzilay, R., McKeown, K. R., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA (pp. 550-557).
[4] Carbonell, J., & Goldstein, J. (1999). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), Melbourne, Australia (pp. 335-336).
[5] 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)* Volume1, Oct. 2002 Page(s):489- 493.
[6] DUC 2003 (Document Understanding Conferences). Available at http://www-nlpir.nist.gov/projects/duc/guidelines/2003.html.
[7] Elhadad, M. (1993). Using argumentation to control lexical choice: a functional unification implementation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.
[8] Eskin, E., Klavans, J., & Hatzivassiloglou, V. (1999). Detecting similarity by applying learning over indicators. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA.
[9] FUF/SURGE. Available at http://sal.jyu.fi/Z/3/FUF-SURGE.html.
[10] Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 40-48).
[11] J. Han and M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2000.
[12] Information Mapping Project. Available at http://infomap.stanford.edu.
[13] 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.
[14] 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.
[15] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.
[16] 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.

[17] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence* (*AAAI'99*), Orlando, FA, USA (pp. 453-460).
[18] MEAD. Available at http://tangra.si.umich.edu/clair/mead.
[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] NLPprocess- Text Analysis Toolkit. Available at http://www.infogistics.com/textanalysis.html.
[21] Robin, J. (1994). Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.
[22] ROUGE. Available at http://www.isi.edu/~cyl/ROUGE/.
[23] 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.
[24] The Porter Stemming Algorithm. Available at http://tartarus.org/~martin/PorterStemmer.
[25] L. Vanderwende and M. Banko and A. Menezes, "Event-Centric Summary Generation," In *Document Understanding Conference at HLT-NAACL*, Boston, MA, 2004.
[26] Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics in a document. In *Proceedings of Workshop on the Computational Treatment of Nominals*, *COLING-ACL*, Montreal, Canada (pp. 70-79).
[27] WordNet. Available at http://wordnet.princeton.edu/.
[28] 陳莉君(2003),"線上個人化參考文獻系統," 碩士論文, 國立交通大學資訊科學研究所, 新竹, 2003.
[29] 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.