

基於基因演算法的組合式多文件摘要方法

An Ensemble Approach for Multi-document Summarization using Genetic Algorithms

陳俊章 Chun-Chang Chen, 鍾羽函 Yu-Hang Chung, 楊正仁 Cheng-Zen Yang

元智大學資訊工程學系

Department of Computer Science & Engineering

Yuan Ze University

Email: {ccc15,yhc17,czyang}@syslab.cse.yzu.edu.tw

范植昇 Jih-Sheng Fan

中央研究院資訊科學研究所

Institute of Information Science, Academia Sinica

Email: fann1993814@iis.sinica.edu.tw

李肇元 Chao-Yuan Lee

妍發科技股份有限公司

Y-Fa Technology Inc.

Email: jason.lee@yfa.com.tw

摘要

多文件摘要研究在文件摘要工作中是一個重要的研究項目。透過多文件摘要，將可以減少人們閱讀許多內容相似但相同主題文件的時間。本研究中提出一個基於基因演算法的組合式多文件摘要模型，以四個網路摘要模型，以及四個機率主題網路摘要模型分別建構組合式多文件摘要模型。在這兩個組合式摘要模型中，以基因演算法找出最佳組合權重。在以DUC 2004年至2007年四個資料集的實驗中，這兩個組合摘要模型在ROUGE-1，ROUGE-2，ROUGE-SU4的評分項目上，相較於個別摘要模型都有最好的表現。

Abstract

Multi-document summarization is an important research task in text summarization. It helps people to reduce much time in reading articles of similar contents but with the same topics. In this study, we propose an ensemble model based on genetic algorithms. Two ensemble summarization models are thus

constructed, one for four network summarization models, and the other for four probabilistic topic network models. These two ensemble models use genetic algorithms to find the optimal weights. We use the datasets of DUC 2004 to DUC 2007 for performance evaluation. The experimental results show that these two ensemble models can achieve the best performance in ROUGE-1, ROUGE-2, and ROUGE-SU4 than other standalone network models and standalone probabilistic topic network models.

關鍵字：多文件摘要，組合摘要模型，基因演算法，效能評估

Keyword: Multi-document summarization, ensemble summarization models, genetic algorithms, performance evaluation

一、緒論

文件摘要 (Text Summarization) 技術能從文件中擷取出重要資訊彙整成摘要，是相當受到重視的研究[7,11]。依照處理的文件數量，文件摘要技術可分為單文件摘要(Single Document Summarization)技術[4]與多文件摘要(Multi-document Summarization) 技術[1,2,3,5,9,13,14,15]。近年來許多研究都在研討多文件摘要技術的發展。透過將許多相關的文件進行分析，產生精簡的摘要，能幫助人們快速了解重要資訊。在過往多文件摘要研究中，已提出許多方法。例如MEAD 使用Centroid-based Summarization 的技術，來計算語句的重要性[9]。Xiong與Luo提出以Latent Semantic Analysis (LSA)的技術來提升多文件摘要的效能[12]。Erkan與Radev提出LexRank[3]，以PageRank的網路模型計算語句重要性。Yang等人以Probabilistic Latent Semantic Analysis (PLSA)計算語句主題並結合網路模型來計算語句重要性[13]。

然而過往研究大多只集中討論在多文件摘要模型中如何使用某一類型特徵中哪一種特徵能被使用來分析這些文章的語句重要性，沒有考慮如何綜合這一類型特徵中的多種特徵來分析語句重要性。因此在本研究中，我們提出一個基於基因演算法(Genetic Algorithms)的組合式多文件摘要模型，以提升多文件摘要的效能表現。在過往多文件摘要研究中，Chali 等人以SVM (Support Vector Machines) 也曾提出一個組合式的摘要模型[2]。然而這個SVM組合式摘要模型則是針對特定方式的使用者需求來進行多文件摘要，並不是通用型多文件摘要(Generic Multi-document Summarization)。本研究所提出的

組合式多文件摘要模型則沒有考慮使用者需求，是通用型的多文件摘要方式。

為了評估所提出的組合式多文件摘要模型，我們使用DUC (Document Understanding Conference) 2004年至2007年四年的資料集進行實驗。在以四個網路摘要模型(Degree Centrality [3], Normalized Similarity-based Degree Centrality [15], PageRank Centrality [3], iSpreadRank Centrality [14])構成的組合摘要模型中，無論是在ROUGE-1, ROUGE-2或ROUGE-SU4這些評分項目上，相較於個別單一摘要模型都能夠得到最好的表現。在以四個機率主題網路摘要模型(PL-Degree, PL-NSDC, PL-PageRank, PL-iSpreadRank)[13]構成的組合摘要模型中，在ROUGE-1, ROUGE-2或ROUGE-SU4這些評分項目上，相較於各別機率主題網路摘要模型，也能夠得到最好的表現。

本論文其餘內容安排如下：第二節將介紹多文件摘要的相關研究。第三節將說明基於基因演算法的組合式多文件摘要模型的設計。第四節將說明以DUC2004年至2007年四個資料集進行實驗的結果。最後，第五節是本論文的結論。

二、 相關研究

在過往多文件摘要技術相關研究上，Radev等人在2000年提出MEAD多文件摘要器[9]。他們使用Centroid-based Summarization (CBS) 質心計算技術，利用一個主題偵測追蹤(Topic Detection and Tracking, TDT) 的集群計算方法找出文件集的質心。他們將MEAD的質心值與Positional value 和First-sentence overlap三種特徵萃取出的摘要經過實驗比較後，實驗結果顯示質心值的效能最好。

潛在語義分析(Latent Semantic Analysis, LSA)也曾被使用在多文件摘要技術中。例如Ozsoy等人於2010年提出二種基於LSA的方法來改進Steinberger與Jezek的方法[10]，處理土耳其文多文件摘要[8]。他們發現所提出的方法中Cross作法能夠得到很好的效果，在實驗中比其他LSA作法的摘要模型都好。Hachey等人於2006年討論SVD對多文件摘要的效果[5]。他們發現SVD加上字詞共現特徵後，雖然摘要效果與TF-IDF的作法相比之下不分軒輊，但是比單純使用字詞共現特徵的效果要好。

Erkan與Radev於2004年提出Degree Centrality 與LexRank兩種網路摘要模型[3]。他

們利用網路模型計算語句重要性。在網路模型中，每個節點代表一個語句。節點之間的連結，則由語句TF-IDF 向量的Cosine Similarity 來決定。如果相似度超過一個門檻值，則兩個語句節點之間存在連結關係，建構出文章語句之間的關係網路。Degree Centrality 方法會計算每個節點的對外連結數，對外連結數越多的節點代表越重要。計算方式如式 (1)，其中 α 為相似度門檻值， s_i 是語句， $\text{deg}(s_i)$ 是 s_i 的Degree Centrality：

$$\text{deg}(s_i) = \sum_{j:i \neq j \text{ and } \text{sim}(s_i, s_j) \geq \alpha} 1 \quad (1)$$

但是當一些不重要的節點因彼此連結而有高的連結數時，Degree Centrality會將這些不重要的語句也納入摘要，降低摘要結果的品質。因此Erkan等人利用網路模型PageRank來計算出語句的重要性。因此對於 s_i 的PageRank分數 $PR(s_i)$ ，計算方式如下：

$$PR(s_i) = \frac{d}{N} + (1 - d) \sum_{s_v \in \text{adj}[s_i]} \frac{PR(s_v)}{\text{deg}(s_v)} \quad (2)$$

其中 N 是語句關係圖節點總數， d 是阻尼因子，通常在0.1-0.2之間。 $PR(s_i)$ 是節點 s_i 的LexRank向心性， $\text{adj}[s_i]$ 是 s_i 相鄰的節點集合， $\text{deg}(s_v)$ 是節點 s_v 的Degree Centrality分數。

2008年Yeh 等人提出另一個網路摘要模型iSpreadRank [14]。iSpreadRank利用了擴散活化理論，考慮節點的連結數量，並且考慮這些節點彼此之間的影響性。在iSpreadRank中，首先會將語句相似度網路轉成一個Sentence-by-Sentence矩陣 A 。

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix} \quad (3)$$

其中 $a_{i,i} = 0$ ， $a_{i,j}$ 代表兩語句的相似度。接著使用擴散活化理論來擴散語句重要性給其他語句。網路中的每個節點都有一個活化級別。在離散時間內iSpreadRank反覆地更新所有節點的影響性，直到終止條件被滿足。在每次迭代中，所有節點在時間點 t 的iSpreadRank Centrality $X(t)$ 計算方式如下：

$$X(t) = X(0) + MX(t - 1), M = \sigma R^T \quad (4)$$

其中 $X(0)$ 是一個 $n \times 1$ 的向量，代表每個語句的初始向心性分數。 σ 為一擴散因子，設為0.7，控制節點影響性的程度。 R 是一隨機矩陣 (Stochastic Matrix)，其中 $r_{i,j} = \frac{a_{i,j}}{\sum_k a_{i,k}}$ ：

$$R = \begin{bmatrix} r_{1,1} & \cdots & r_{1,n} \\ \vdots & \ddots & \vdots \\ r_{n,1} & \cdots & r_{n,n} \end{bmatrix} \quad (5)$$

擴散過程的終止條件設定為時間 t 與時間 $t - 1$ 的節點影響性差異總和是否小於一個門檻值 ϵ 。若小於則結束擴散過程。計算方式如下，在[14]中 $\epsilon = 0.0001$ ：

$$\sum_i |X_i(t) - X_i(t - 1)| \leq \epsilon \quad (6)$$

當iSpreadRank計算結束時，網路處於穩定狀態，每個節點用數字權重標記它們的最終重要程度。iSpreadRank便依照權重高低排名來選擇語句加入摘要。

2013年，Yeh等人討論了不同的網路模型多文件摘要技術[15]。他們同時討論兩個不同的關係網路：連續權重語句關係網路與離散權重語句關係網路，在連續權重語句關係網路中，節點之間的Cosine Similarity 值若高於門檻值，便為節點連結的權重，若沒有高過門檻值，則為0。在離散權重語句關係網路中，節點之間的連結關係則成為一種0或1的二元關係。接著，他們討論五種節點向心性計算方法在這二種關係網路上的表現，分別為：Degree，Normalized Similarity-based Degree，HITS，PageRank 及iSpreadRank。實驗結果顯示，連續權重語句關係網路較離散權重語句關係網路更佳，五種節點向心性的計算以iSpreadRank具有最高的效能。

機率式潛在語意分析(Probabilistic Latent Semantic Analysis, PLSA) 改善LSA不符合統計觀點和無法解決一詞多義的缺點，以機率計算並找出文件中潛在的重要主題，再以機率計算並找出符合該潛在主題的字詞，進而找到含有這些字詞的語句來當作摘要。2014年，Yang等人提出了機率主題網路模型(Probabilistic Topic-based Network Models)的多文件摘要技術[13]。在機率主題網路模型中，語句的主題特徵向量成為語句關係網路節點，再用不同的網路模型來處理主題網路。在實驗中分別討論PLSA與LDA (Latent Dirichlet Allocation)主題模型的效果。實驗結果顯示PLSA所強化的機率主題網路摘要模型能比原有的網路模型有更好的效能，且在不同的資料集中都有名列前茅的表現。

2009年Chali等人提出一個組合式摘要模型(Ensemble Summarization Model)[2]。這個組合摘要模型使用Support Vector Machines (SVM)的組合方法，並使用多個SVM建構一個交叉驗證委員會機制(Cross-Validation Committees, CVC) 來進行摘要。其策略是藉

由使用其他SVM的輸出來校正其中一個SVM的錯誤。實驗結果證明組合的SVM摘要模型效能高於單一SVM摘要模型。

基因演算法(Genetic Algorithms) 是一種用於解決最佳化問題的搜索算法。對於一個最佳化問題，GA用染色體代表可能的解答，透過選擇、交配、突變等過程，來尋找最佳解。Bossard和Rodrigues在2011年提出Clustering Based Sentence Extractor for Automatic Summarization (CBSEAS)多文件摘要方法，將多文件摘要技術與GA結合[1]。CBSEAS依據與文件質心的相似度來排列所有語句，語句的相似度以Jaccard方式計算，接著建構相似矩陣，利用Fast Global K-means來將語句分成數個集合，並以四個加權方式：與用戶查詢或質心相似度、與集群中心相似度、重要的句子分數、得分句與期望句長之間的長度差異，來為每個集合選擇一個最具代表性的摘要。實驗結果顯示GA可以找到最佳的參數組合以提高摘要效能。

三、 組合式多文件摘要模型

在基因演算法組合式多文件摘要模型中，這個摘要模型的處理流程將說明如下。

(一)、組合式多文件摘要處理流程

圖1是本研究所提出的基因演算法組合式摘要模型的處理流程。將多個單一摘要模型進行組合，然後由基因演算法來找出最佳權重。在處理流程中，首先會將文件進行前處理。前處理的項目分別為Sentence Segmentation，Tokenization，Stopword Removal 以及 Stemming。然後按照不同的單一摘要模型分別計算出語句的分數。基因演算法組合摘要模型則依照線性組合公式，用基因演算法找出最佳組合權重，重新得到語句的最後加權分數。最後則用CSIS演算法[9]進行語句篩選，產生最後的摘要。

在組合的單一摘要模型中，分別考慮近年來有不錯表現的四個網路摘要模型，以及四個經過PLSA強化後的機率主題網路模型。這四個網路主題模型分別是Degree Centrality (Degree)摘要模型[3]，Normalized Similarity-based Degree Centrality (NSDC)摘

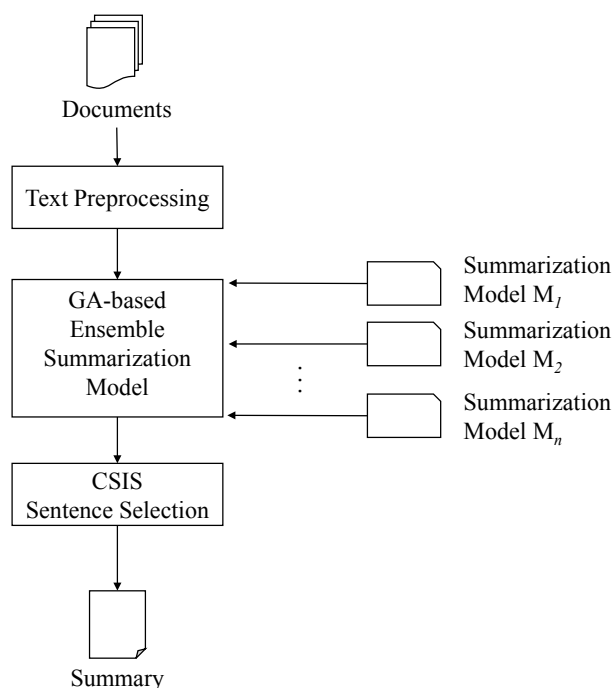


圖1、基因演算法組合摘要模型處理流程

要模型[15]，PageRank摘要模型[3]，以及iSpreadRank摘要模型[14]。它們的PLSA 主題強化摘要模型分別是PL-Degree，PL-NSDC，PL-PageRank 以及PL-iSpreadRank [13]。

(二)、文件前處理

在進行多文件摘要前，每份文件都需要先進行文本的前處理。由於文件敘述中會有一些缺乏意義的詞彙，例如冠詞或Be動詞等所謂的停用詞(Stopword)，這些詞通常無法表現文件摘要的語句重要性，因此在前處理中會被移除。以下是前處理中進行的項目：

- Sentence Segmentation：文件中的每一個語句先需要被斷句分開。本研究使用NLTK (Natural Language Toolkit)¹工具來進行斷句。
- Tokenization：本研究使用NLTK以語句為單位，將語句中的單詞取出，以bag-of-words的方式來表示該語句。
- Stopword Removal: 我們使用Onix²停用詞表將語句中的停用詞移除。
- Stemming:在經過停用詞移除後，本研究再使用Porter Stemmer³進行詞根還原。

¹ <https://www.nltk.org/>

² <http://www.lextek.com/manuals/onix/stopwords1.html>

³ <https://tartarus.org/martin/PorterStemmer>

經過上述前處理，如果語句的字數少於3個，該語句因為含有字詞過少，因此該語句極有可能缺乏重要語意，在本研究中將會捨棄而不予考慮。

(三)、組合成員之單一摘要模型

1、語句關係網路建構

在本研究中，所組合的單一摘要模型為四個網路摘要模型與四個機率主題網路摘要模型。因此將先建構出語句關係網路(Sentence Relationship Network)。在網路摘要模型中，語句之間的關係是用語句之間的相似度(Sentence Similarity) 來考慮。在本研究中，語句以TF-IDF (Term Frequency-Inverse Document Frequency) 的向量表示。兩個語句 s_i 和 s_j 之間的相似度用Cosine Similarity 來計算，如式(7)所示。

$$\text{sim}(s_i, s_j) = \frac{\bar{s}_i \cdot \bar{s}_j}{|\bar{s}_i| \times |\bar{s}_j|} \quad (7)$$

在建構語句相似網路時，如果 $\text{sim}(s_i, s_j) < 0.1$ ，因為相似度太低，因此在網路模型中這兩個語句便不會有連結關係。

在機率主題網路摘要模型中，是以每個語句的主題字詞作為網路節點。語句之間的相似度是以主題機率的相似度來考慮。因此對於語句 s_i 和 s_j ，在主題空間 z 的情況下，它們的主題特徵向量 $\vec{v}_i = \langle p(z_1|s_i), p(z_2|s_i), \dots, p(z_K|s_i) \rangle$ ， $\vec{v}_j = \langle p(z_1|s_j), p(z_2|s_j), \dots, p(z_K|s_j) \rangle$ 兩者的相似度用Cosine Similarity 計算，如式(8)：

$$\text{sim}(v_i, v_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \times |\vec{v}_j|} \quad (8)$$

在建構主題機率相似度網路時，也會設定門檻值來判斷語句之間的相似度關係是否存在。

2、單一摘要模型

在語句關係網路中，每個節點是各個語句的TF-IDF向量。在主題機率關係網路中，每個節點是各個語句的主題機率特徵向量。在本研究中，組合成員則是基於這兩種不同向量 上的四種網路摘要模型：Degree, NSDC, PageRank與iSpreadRank。Degree, PageRank與iSpreadRank如第2節中所介紹，在此便不重複說明。NSDC則是基於Degree模型但是考慮所有節點的相似度，然後就相似度總和來做正規化。在NSDC中，對語句 s_i 而言，它

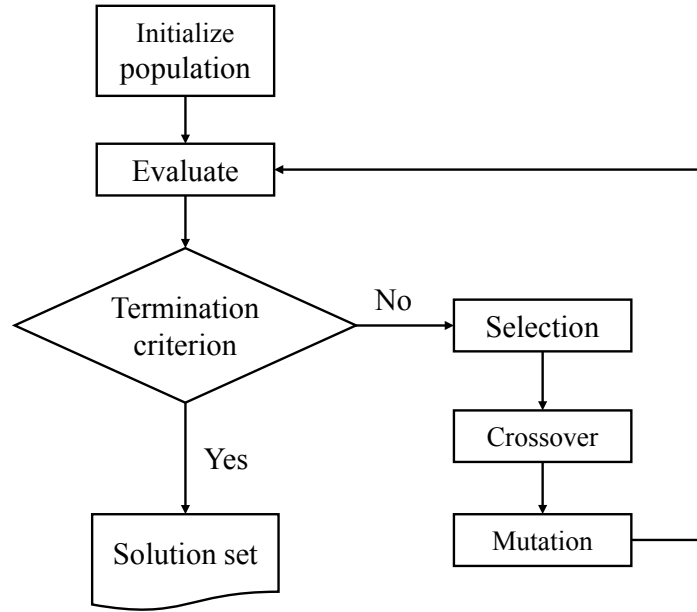


圖2、基因演算法流程

的NSDC Centrality $sim deg(s_i)$ 計算如下：

$$sim deg(s_i) = \sum_{j:i \neq k \text{ and } sim(s_j, s_k) \geq \alpha} \frac{sim(s_i, s_j)}{\sum_{k:j \neq k \text{ and } sim(s_j, s_k) \geq \alpha} sim(s_j, s_k)} \quad (9)$$

(四)、基於基因演算法的組合摘要模型

在組合摘要模型中，組合的方式是每個單一摘要模型 M_j 先各別算出語句 s_i 語句重要性分數 I_{ij} ，然後用一個線性組合公式來算出語句 s_i 的最後排序分數 R_i 。在線性組合公式中，每個單一摘要模型的 I_{ij} 會分別乘上一個加權權重 w_j 。因此線性組合的計算公式如下：

$$R_i = \sum_j w_j \times I_{ij} \quad (10)$$

為了要使語句排序的計算得到最佳結果，我們使用基因演算法來求得最佳組合權重。我們以摘要的目標函式作為GA適應函式。本研究分別以ROUGE-1, ROUGE2和ROUGE-SU4做為最佳化的目標函式來求得最佳的組合權重。

基因演算法的流程如圖2所示，以下將說明各個步驟。

1. Initialize population:

本研究以 $K=50$ 組染色體來進行計算。每組染色體有 N 個基因， N 為單一摘要模型總數，代表單一摘要模型 M_j 的權重 w_j 。 w_j 的初始值範圍是 $0 \leq w_j \leq 1$ ，取到小數點後4位，同時 $\sum_j w_j = 1$ 。在初代染色體中，保留代表各個單一摘要模型 $(1, 0, \dots, 0), (0, 1, \dots,$

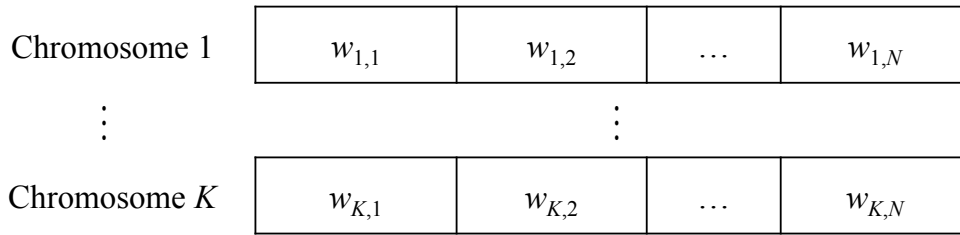


圖3、基因演算法中的染色體

0),..., (0, 0, ..., 1)的染色體。其餘染色體是隨機產生。圖3是此 K 組染色體的情況。

2. Evaluate:

在基因演算法評估時，會針對適應函式ROUGE-1，ROUGE-2，ROUGE-SU4 三種不同的摘要評分方式找出各自的最佳組合權重。GA 會以迴圈方式進行評估。中斷條件是當下面兩個條件其中之一滿足，GA便會結束。

- 連續5次的適應函數結果差距 ≤ 0.0001 。
- 基因產生迴圈已進行1000次迭代。

3. Selection:

如果GA中止條件沒有滿足，GA便會進行Selection，Crossover和Mutation。在Selection中，如果該染色體的適應函式分數 $F_k \geq F_{avg}$ ，就會被保留下來。如果 $F_k < F_{avg}$ ，該染色體被目前所找到的最佳染色體所代替。 F_{avg} 是所有染色體的平均適應函式分數。

4. Crossover:

GA依據交配率決定是否進行交配。本研究設定交配率為0.5，進行 K 次交配，以單點交配法進行。交配時會隨機選2組染色體，並隨機決定1個切斷點，將這2組染色體切

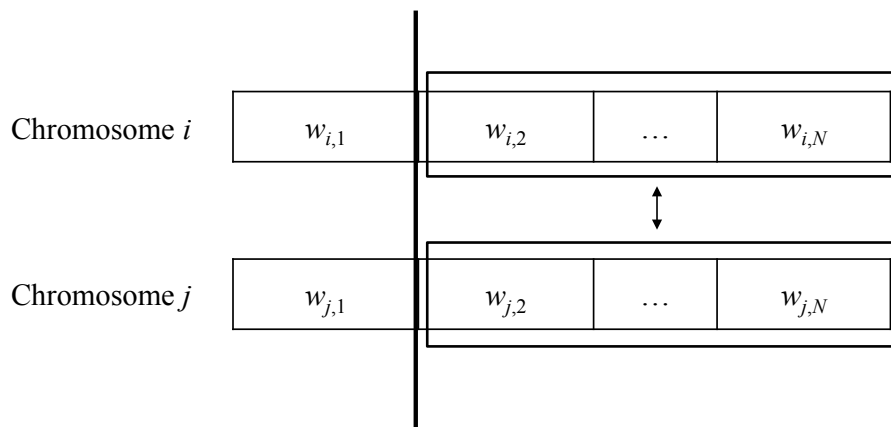


圖4、染色體的單點交配方式

Input: a rank list L of sentences
Output: the summary Sum
Initialize:
 set $Sum = \emptyset$
Summarize:
 while the required compression rate R is not met
 $s_c \leftarrow$ the candidate sentence having the highest score in L
 if $\max_{s_j \in Sum} sim(s_c, s_j) < C_R$
 add s_i to Sum
 else
 omit s_i
 endif
 remove s_j from L
 output the summary Sum

圖5、CSIS語句選取方式

成兩部分，再將切開的部分重新組合成一對新的染色體。交配的方式如圖4所示。

5. Mutation:

GA交配完成後進入突變流程。本研究中突變率設0.25，希望能盡可能突破Local Optimal的情況。突變方式為單點突變，GA會隨機選擇該染色體的任一個基因突變，取一個在0到1之間的小數來取代，精確度到小數點後4位。

(五)、語句選擇

經由組合計算後，將得到的語句分數排序，然後按此順序挑選摘要語句。為了去除重複多餘的資訊，我們採用CSIS (Cross-Sentence Information Subsumption)[9]來檢查候選語句是否與已選的摘要內容重複。CSIS演算法如圖5所示，相似度計算採用Cosine Similarity， R 是摘要大小的限制， Sum 是摘要集合， C_R 門檻值決定語句是否重複，本研究設定為0.7。 L 是語句的排序列表。在計算相似度時，在網路摘要模型中是用語句的TF-IDF向量來計算，在機率主題網路摘要模型中，則是用語句的主題機率特徵向量來計算。

四、實驗結果及分析

(一)、實驗環境設定

表1、DUC資料集內容

	DUC 2004	DUC 2005	DUC 2006	DUC 2007
文件集數	50	50	50	45
每個文件集的文件數	10	25-50	25	25
摘要長度要求	665 bytes	250字	250字	250字

表2、實驗參數設定

主題數目(τ)	256 (DUC 2004,DUC 2005), 128 (DUC 2006,DUC 2007)
主題機率相似度門檻值(γ)	0.05 (NSDC:2007; iSpreadRank:2005, 2006), 0.2 (Degree:2004, 2007; NSDC:2006; PageRank:2004), 0.25 (Degree:2005, 2006; NSDC:2004, 2005; PageRank: 2005, 2006, 2007; iSpreadRank:2004, 2007)
字詞向量相似度門檻值(S_t)	0.1
PageRank阻尼因子(d)	0.15
iSpreadRank擴散因子(σ)	0.001
CSIS篩選門檻值	0.7

本實驗採用的資料集是來自於DUC (Document Understanding Conference)所提供DUC 2004- 2007共4個資料集。它們的特性如表1所示。評估標準是採用ROUGE (Recall-Oriented Understudy for Gisting Evaluation)作為評估工具[6]，ROUGE將人工所產生的摘要作為參考答案，以 n -gram為單位，計算人工摘要和 機器所產生摘要的同樣所包含的字詞比例。本實驗將針對ROUGE-1，ROUGE-2，ROUGE-SU4來進行探討。

在不同的網路模型中，各項實驗參數設定如表2所示。這些參數的設定是參考自[3]和[13]研究中的參數設定。

(二)、網路模型組合實驗

在網路摘要模型的組合實驗中，本研究所提出的基因演算法組合摘要模型(GA-DNPI) 與另外四個網路摘要模型Degree，NSDC，PageRank，iSpreadRank 進行效能比較。表3到表6是這五個摘要模型在四個資料集中的效能表現。我們可以看到GA-DNPI在這四個資料集中都有最好的表現。

(三)、機率主題網路模型組合實驗

在機率主題網路模型的組合實驗中，本研究所提出的基因演算法組合摘要模型(GA-PL-DNPI)與另外四個機率主題網路摘要模型PL-Degree, PL-NSDC, PL-PageRank,

表3、DUC 2004效能比較

	ROUGE-1	ROUGE-2	ROUGE-SU4
Degree	0.36264	0.08471	0.12636
NSDC	0.35068	0.06862	0.11376
PageRank	0.36482	0.08348	0.12539
iSpreadRank	0.36874	0.08882	0.12885
GA-DNPI	0.37009	0.08963	0.12993

表4、DUC 2005效能比較

	ROUGE-1	ROUGE-2	ROUGE-SU4
Degree	0.36444	0.06843	0.12342
NSDC	0.36255	0.06244	0.11850
PageRank	0.36891	0.06801	0.12455
iSpreadRank	0.36494	0.06693	0.12228
GA-DNPI	0.37235	0.06912	0.12575

表5、DUC 2006效能比較

	ROUGE-1	ROUGE-2	ROUGE-SU4
Degree	0.39174	0.08388	0.13943
NSDC	0.39216	0.07673	0.13450
PageRank	0.39867	0.08497	0.14187
iSpreadRank	0.39568	0.08491	0.14032
GA-DNPI	0.39929	0.08550	0.14204

表6、DUC 2007效能比較

	ROUGE-1	ROUGE-2	ROUGE-SU4
Degree	0.41209	0.09618	0.15271
NSDC	0.40988	0.08974	0.14645
PageRank	0.41561	0.09784	0.15332
iSpreadRank	0.41263	0.09741	0.15260
GA-DNPI	0.41781	0.09838	0.15422

PL-iSpreadRank 進行效能比較。由於在[13]中，主題數目也會影響效能，因此我們在實驗中，單一的機率主題摘要模型會用在 Rouge-1表現不錯的主題數目來處理。在實驗中，主題數目以64，128，256來考慮。表7到表10是這五個摘要模型在 DUC2004-DUC2007四年資料集中的效能表現，我們可以看到GA-PL-DNPI 也都有最好的表現。

五、 結論

多文件摘要是一件相當受到重視的研究工作，因為能夠幫助人們迅速掌握關鍵資料。本研究提出了基於基因演算法的組合摘要模型來處理自動多文件摘要問題，使用基因演算法來求得最佳組合。在 DUC 2004-2007資料集的實驗中，相較於個別單一摘要模型，

表7、DUC 2004不同主題數目效能比較

	τ值	ROUGE-1	ROUGE-2	ROUGE-SU4
PL-Degree	256	0.38440	0.09010	0.13395
PL-NSDC	64	0.37878	0.08472	0.13024
PL-PageRank	256	0.38387	0.08978	0.13379
PL-iSpreadRank	64	0.38087	0.08297	0.13020
GA-PL-DNPI	256	0.38466	0.09165	0.13515

表8、DUC 2005不同主題數目效能比較

	τ值	ROUGE-1	ROUGE-2	ROUGE-SU4
PL-Degree	256	0.38510	0.07269	0.13239
PL-NSDC	64	0.37368	0.06276	0.12402
PL-PageRank	256	0.38456	0.07145	0.13221
PL-iSpreadRank	64	0.37473	0.06389	0.12393
GA-PL-DNPI	256	0.38524	0.07289	0.13264

表9、DUC 2006不同主題數目效能比較

	τ值	ROUGE-1	ROUGE-2	ROUGE-SU4
PL-Degree	128	0.41315	0.08642	0.14696
PL-NSDC	128	0.41296	0.08748	0.14743
PL-PageRank	128	0.41355	0.08629	0.14703
PL-iSpreadRank	128	0.41172	0.08795	0.14708
GA-PL-DNPI	128	0.41355	0.08854	0.14785

表10、DUC 2007不同主題數目效能比較

	τ值	ROUGE-1	ROUGE-2	ROUGE-SU4
PL-Degree	128	0.42890	0.10150	0.15860
PL-NSDC	64	0.42889	0.10153	0.15933
PL-PageRank	128	0.42800	0.10516	0.16069
PL-iSpreadRank	64	0.42793	0.10065	0.15916
GA-PL-DNPI	128	0.43133	0.10645	0.16273

都能夠在ROUGE-1，ROUGE-2 和ROUGE-SU4這些評分項目上得到最好的表現。

六、 參考文獻

- [1] A. Bossard and C. Rodrigues, "Combining a Multi-Document Update Summarization System –CBSEAS- with a Genetic Algorithm," in *Proceedings of the 2nd International Workshop on Combinations of Intelligent Methods and Applications (CIMA 2010)*, 2011, pp. 71–87.
- [2] Y. Chali, S. A. Hasan, and S. R. Joty, "A SVM-Based Ensemble Approach to Multi-Document Summarization," in *Proceedings of the 22nd Canadian Conference on Artificial Intelligence (Canadian AI 2009)*, 2009, pp. 199–202.

- [3] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality As Saliency in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, Jul. 2004.
- [4] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, 2001, pp. 19–25.
- [5] B. Hachey, G. Murray, and D. Reitter, "Dimensionality Reduction Aids Term Co-occurrence based Multi-document Summarization," in *Proceedings of the Workshop on Task-Focused Summarization and Question Answering (SumQA '06)*, 2006, pp. 1–7.
- [6] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.
- [7] I. Mani, *Automatic Summarization*. John Benjamins Publishing Co., 2001.
- [8] M. G. Ozsoy, I. Cicekli, and F. N. Alpaslan, "Text Summarization of Turkish Texts Using Latent Semantic Analysis," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*, 2010, pp. 869–876.
- [9] D. R. Radev, H. Jing, and M. Budzikowska, "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies," in *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, vol. 4, 2000, pp. 21–30.
- [10] J. Steinberger and K. Ježek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation," in *Proceedings of the 5th International Conference on Information Systems Implementation and Modelling*, Apr. 2004, pp. 93–100.
- [11] J.-M. Torres-Moreno, *Automatic Text Summarization*, J.-C. Pomerol, Ed. John Wiley & Sons, Inc., 2014.
- [12] S. Xiong and Y. Luo, "A New Approach for Multi-document Summarization Based on Latent Semantic Analysis," in *Proceedings of the 2014 7th International Symposium on Computational Intelligence and Design (ISCID '14)*, 2014, pp. 177–180.
- [13] C.-Z. Yang, J.-S. Fan, and Y.-F. Liu, "Multi-document Summarization using Probabilistic Topic-based Network Models," *Journal of Information Science and Engineering*, vol. 32, no. 6, pp. 1613–1634, Nov. 2016.
- [14] J.-Y. Yeh, H.-R. Ke, and W.-P. Yang, "iSpreadRank: Ranking Sentences for Extraction-based Summarization Using Feature Weight Propagation in the Sentence Similarity Network," *Expert Systems with Applications*, vol. 35, no. 3, pp. 1451–1462, Oct. 2008.
- [15] J.-Y. Yeh, W.-P. Yang, H.-R. Ke, and P.-C. Cheng, "Extraction-based News Summarization Using Sentence Centrality in the Sentence Similarity Network," *Journal of Information Management*, vol. 21, no. 3, pp. 271–304, 2014.