2. 相關研究

本論文將自動文件摘要的研究範疇略分為三大類，分別為「文件摘要方法的研究」、「摘要評估方法的研究」以及「摘要的呈現方法」。在大部份人的認知上，談到自動文件摘要的研究很直覺地就連想到是研究自動產生文件摘要的方法。但是，其實不盡然是如此。自動文件摘要的研究不像是語音辨識的研究，有公認的標準答案與評估方法為依據，然後只需比較召回率(Recall)、正確率(Precision)或是字錯誤率(Character Error Rate)就可以知道方法的優劣好壞。以標準答案來說，文件摘要並沒有絕對的標準答案，不同的人對不同文件的摘要認定就會不相同，因此，若給定一篇文件，由十位評估者所產生的摘要必定不盡相。於 [Gong et al. 2001; Sun et al. 2005]亦提到評估者對摘要認定的差異問題。因此，多數的摘要評估均是採用多份參考摘要來進行評估，以處理自動文件摘要的標準答案(參考摘要)不一的情形。美國的文件理解會議(Document Understanding Conference, DUC)針對文件摘要(Text Summarization)領域已經有提供標準的訓練集與測試集，可供不同文件摘要系統的比較，但是對於語音文件摘要還沒有一套標準的訓練集與測試集供研究使用。另一方面，目前台灣在中文語音文件摘要的研究，亦缺少一套一致性的測試資料可供使用。

由於文件摘要的標準答案不一的問題，以及文字文件摘要與語音文件摘要在評估上的問題與差異，有很多學者亦提出了不同的方法來解決摘要評估的問題[Hori et al. 2004; Zhu and Penn 2005]。但是，不同的評估方法對於不同的摘要結果的評估並不一致，如在[Zhu and Penn 2005]論文中針對不同評估方法的探討，提到不同的摘要方法在不同評估方式下摘要結果的排名並不相同，因此對於不同摘要模型的比較，我們就無法絕對的認定其優劣。自動摘要的方法不斷的被廣泛探討與研究，許多學者提出不同的摘要方法來提高摘要的正確率，但是若沒有一

套公平客觀且統一的評估方法，將無法正確地判定摘要研究成果的提昇。如此，除了摘要方法的研究，評估方法亦屬於自動文件摘要的重要研究範疇之一，目前已有許多文獻的探討。

本論文亦將摘要的呈現方法也列為自動文件摘要的重要研究內容之一，於本章最後一節有概略的討論與文獻回顧。呈現方式的研究，在過去只以文字文件摘要研究為主，以及多媒體影音技術還不發達的時代裡，並不會被特別討論。但是，隨著語音文件摘要的廣泛研究，吾人認為摘要的呈現方式將會持續地被研究與討論。語音文件摘要可以文字來呈現，方便使用者瀏覽、查詢，亦可以語音訊號來呈現，方便儲存與收聽。其中，語音訊號的呈現方式，牽涉到語音訊號的流暢度、訊號串接或合成等技術，雖然這些技術可被歸類為訊號處理的範疇；但是，由於與文件摘要的研究極為相關，論文中亦作簡單的概述。

本論文的研究主軸旨在探討自動文件摘要的方法。本章將針對相關的研究內容作簡單的回顧。將於於第一節中先概述自動文件摘要的歷史背景；然後，第二節介紹過去學者所提出的摘錄式摘要方法；第三節將針對常見的評估方法作簡單介紹；最後，概略地探討目前語音文件摘要的呈現方式及相關的研究。

2.1 自動文件摘要的歷史背景概述

自動文件摘要最早起源於 1950 年晚期(如圖 2.1 的文件摘要歷史追溯概略圖所示)，由 Luhn [Luhn 1958]在 1958 年提出表面層次方法(Surface-level Approach)應用於文字文件摘要(Text Summarization)中，主要是使用一些語幹特徵(Thematic Features)，例如詞的頻率(Word Frequency)。從 1961 年起，有學者開始提出實體層次的方法(Entity-level Approach)[Climenson et al. 1961]，嘗試使用句法分析(Syntactic Analysis)的方式來產生摘要；直到 1969 年，位置的特徵(Location Features)才開始有學者使用於文件摘要上[Edmundson 1969]。在 1970 年的早期與後期，分別有學者對表面層次方法及實體層次方法，提出進一步的探討與廣泛的

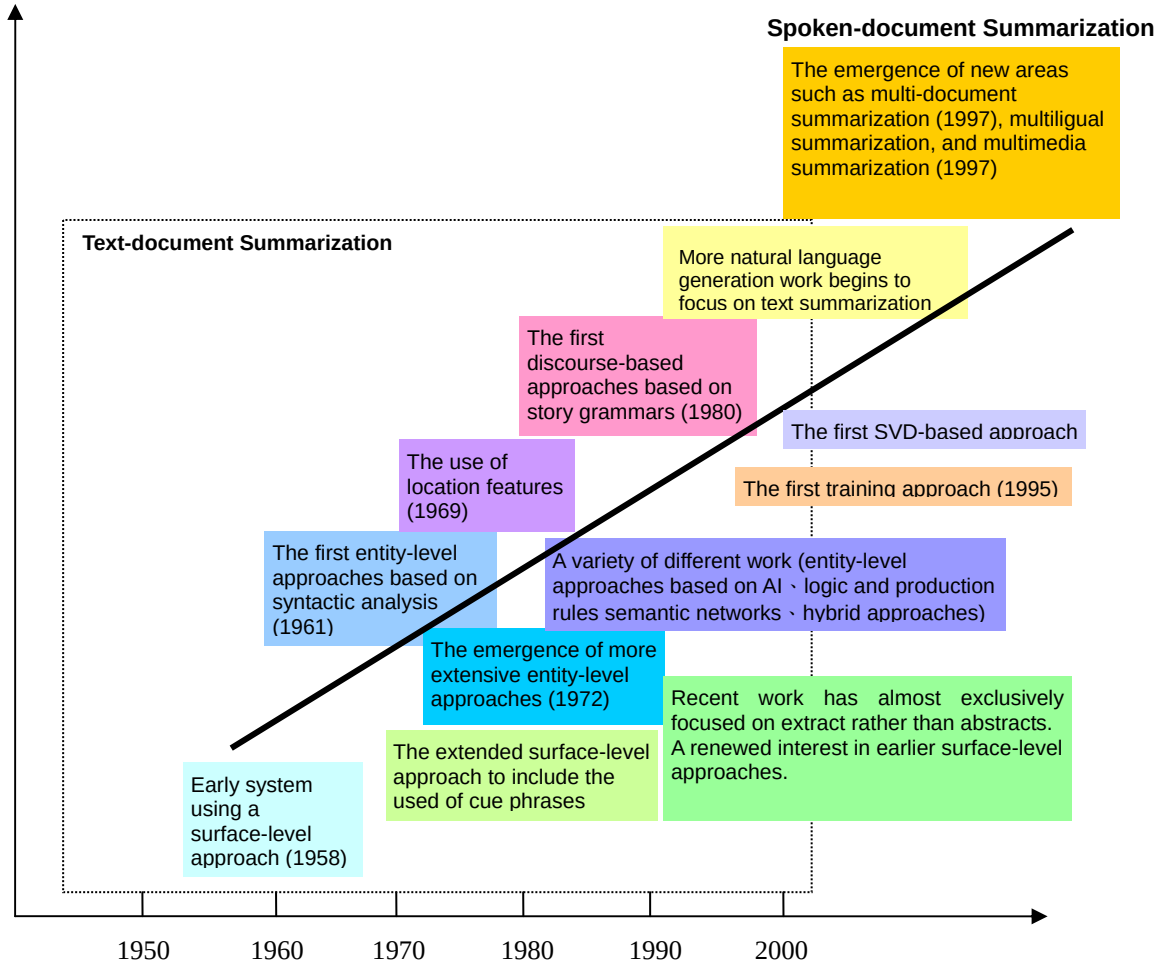

圖 2.1 文件摘要歷史追溯概略圖

使用，前者包括一些線索慣用語(Cue Phrases)的使用；同時，於 1975 年有了第一個自動摘要的商業性應用[Pollock et al. 1975]。1980 年開始，以會話為基礎的方法(Discourse-based Approaches)[vanDijk 1980]及各種不同方法的結合與運用也慢慢被探討，例如以人工智慧為基礎的實體層次方法、語意網路(Semantic Networks)、或混合方法(Hybrid Approaches)等。

然而，縱使這幾十年間陸續有學者提出不同摘要的方法及應用，自動文件摘要的研究並不活躍。一直到 1990 年後期，由於網際網路蓬勃發展、商業價值的提高，以及政府的關注，使得此領域重新復甦；同時，自動摘要的研究重心開始

著重於摘錄式摘要為主，自然語言處理的研究也漸漸將焦點集中於文件摘要上。此外，由於網際網路與多媒體技術的發展，各種不同形式的文件摘要的研究也慢慢出現，像是多文件摘要(Multi-document)[Salton et al. 1997]、多語言摘要(Multilingual Summarization)以及多媒體摘要(Multimedia Summarization)[Takeshita et al. 1997；Merlino et al. 1999]。因此，文件摘要研究開啟了另一個新的方向－語音文件摘要(Spoken Document Summarization)。

由於影音多媒體技術的發展與進步，文件內容已不再僅有文字，而是漸漸以動態的影音來呈現。其中，語音為多媒體影音內容中重要的訊息來源。因此，要針對大量且多元的多媒體影音內容，做自動化及系統化的處理，主要是以其中的語音資訊內容為考量，例如，語音文件的資訊檢索與摘要。然而，就文件摘要的研究來說，大量具時序性的語音訊號，例如：廣播新聞、演講錄音、語音郵件等，帶來了傳統文字文件摘要所沒有的額外問題，像是辨識錯誤、文句邊界定義錯誤、口語停頓、遲疑或是重覆的無意義詞等，造成過去的傳統方法無法正確摘錄出重要文句的問題。此外，如何利用語音文件中的豐富資訊(如音高、強度、重音及停頓時間等聲韻特徵)，來提昇文件摘要的精確率，亦成為新的研究課題。因此，正如圖 2.1 所示，從 1990 年後期開始，文件摘要的研究邁入了新的階段，除了文字文件摘要的新方法與概念持續被提出，例如 1995 年提出採用訓練方式的方法[Kupiec et al. 1995]及 2001 年有學者提出的相關評估及潛藏語意分析的方法[Gong et al. 2001]。另一方面，語音文件摘要的研究亦展開了一場激烈的競逐，許多的研究目前正被廣泛的探討中[Kikuchi et al 2003; Maskey et al. 2005]，這股熱潮並將持續著！

2.2 自動文件摘要方法

摘要依其形成方式、目的、文件形式或文件來源等，可細分為不同的摘要類別 [Ho 2003]。圖 2.2 以階層的方式表示摘要在基於不同考量與情況下的分類，以及類別

之間的關係，例如根據“文件的形式”摘要可分為文字文件摘要(Text Document Summary)及語音文件(Spoken Document Summary)摘要、根據“文件來源”可分為單一文件(Single Document)摘要及多文件(Multi-Document)摘要；同時，文字文件及語音文件的摘要可依文件來源分為文字文件摘要、語音文件摘要；或者，依摘要目的分一般性(Generic)文件摘要、以查詢為基礎(Query-based)的文件摘要等。而在某些類別之間是具有相關性的，如圖 2.2 所示，依文件形式來分類的文字文件及語音文件的摘要，可再依文件來源分為單一文字文件摘要、單一語音文件摘要、多文字文件摘要、多語音文件摘要；然後，單一文字文件摘要又可根據目的分為一般性單一文字文件摘要、以查詢為基礎的單一文字文件摘要。因此，圖 2.2 以階層的方式來表達這樣的關係。

如圖 2.2 所示，摘要依的形成方式可分成摘錄式(Extractive)與非摘錄式(Non-extractive or Abstractive)兩類 [Hove et al. 1998]。摘錄式摘要是依據設定之摘要比例(Summarization Ratio)從原文件中選出重要的文句、段落或章節來組成摘要。其中，摘要比例是摘要長度佔全文件長度的比例，例如 20%的摘要比例就是擷錄出長度約為原文件 20%的摘要內容。此種方式可能發生摘要內容中文句或段落間，語意或內容的不連貫或不流暢；此外，選取之文句、段落中亦可能包含有某些不重要的綴詞或不相關的內容。非摘錄式摘要是依文件內容中之主題或概念而重寫成的摘要，摘要內容可能含有非原文件中所使用但語意相同的詞或片語，其優點為摘要內容連貫具可讀性。相較於摘錄式摘要方法，非摘錄式摘要的內容更為接近由人所撰寫文章，因此需考慮資訊擷取(Information Extraction)及語意(Semantics)、文法(Grammars)與對話理解(Discourse Understanding)等自然語音處理(Natural Language Processing)領域的知識，才能產生確切表達文件主題並且內容流暢的摘要。此外，非摘錄式摘要的評估標準定義困難，若採用一般的評估方法，例如索引(詞)相配(Term Matching)的評估方式，縱使摘要內容與文件主題相符，亦可能因其包含有非標準答案(參考摘要)中的詞或片語，造成較低的摘要正確率；而若是請測試者主觀地來評估摘要內容的優劣，可能會因每個人在認知

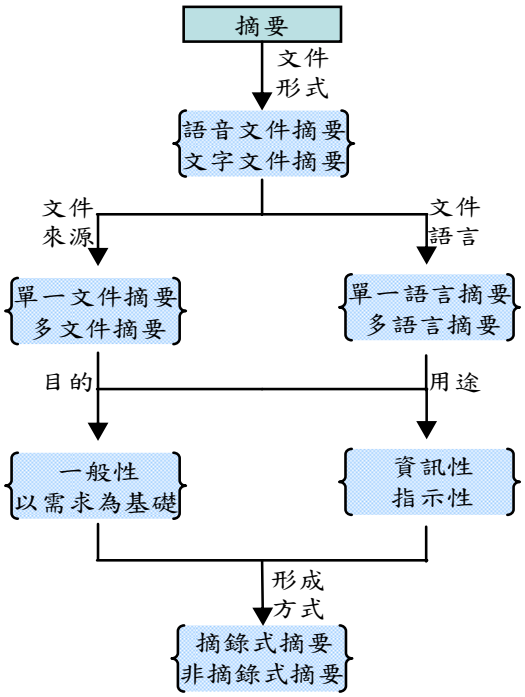

圖 2.2 文件摘要細分類圖

上的不同而差異甚大。因此，非摘錄式摘要較摘錄式摘要存在著更多的困難與挑戰，目前自動語音文件摘要的研究以摘錄式摘要為主。本論文亦屬於摘錄式摘要方法的研究。

摘要依其目的可分為一般性(Generic)摘要與以查詢為基礎(Query-based)的摘要，前者呈現文件全面性的主題，摘要內容以涵蓋整篇文件所有重要主題為主；後者則根據使用者或特定的查詢來呈現文件中與查詢相關的摘要內容。摘要依其用途可分為資訊性(Informative)或指示性(Indicative)摘要。資訊性摘要嘗試表達文件中重要的資訊內容；而試圖要將摘要內容作為提供文件分類資訊的屬於指示性摘要。此外，根據文件來源可分為單一文件與多文件摘要，針對一篇文件內容所擷錄的摘要屬單一文件摘要，歸納多篇主題相近的文件所產生的共同摘要內容為多文件摘要。依文件的語言分為單一語言(Monolingual)與多語言(Multilingual)摘要，多語言摘要指對不同語言但主題相近的多篇文件產生一個以單一語言表示的摘要。或者，因文件形式的不同分為語音文件與文字文件摘要兩類，文字文件摘

要是針對以文字內容為主的文件所產生的摘要，而語音文件摘要則是以語音辨識後的自動轉寫文件，其中可能包含辨識錯誤及某些無義意的內容。相較於文字文件摘要，語音文件摘要包含有許多困難點，例如辨識錯誤或因辨識錯誤所造成的語意、詞素特徵(Lexical Features)統計值(如詞頻)的錯誤，以及文句邊界定義的問題。但是相對地，語音文件也帶來了更多可利用的語音資訊，例加重音、聲調、語音辨識信心度分數等。

本論文以一般性單一中文語音文件摘錄式摘要的研究為主，並將摘錄式摘要視為文句之重要性排序(Ranking of Sentence Importance)的問題，當某一文句對於整篇文件的重要性越高或是相關性越大，則優先選取為文件的摘要內容。下面各小節中將針對過去文獻中所探討的摘錄式摘要方法作簡單的概述與回顧。本論文將不同的摘錄式摘要方法概略地分成三類：以文件結構為基礎、以統計值為基礎，以及以機率生成模型為基礎的摘錄方法。

2.2.1 以文件結構為基礎的摘錄方法(Document Structure-based Approach)

依摘要單位元(如詞、文句)所在的位置(Location)來決定其重要性。一般來說，位於重要段落的摘要單位元可能有較高的重要性及代表性，如第一段、最後一段或位於標題如「簡介、目的、結論」的段落被視為重要。巴氏[Baxendale 1958]在其研究文獻中表示，有 85%的段落其主題句(Topic Sentences)出現在段落的最前面，而 7%的段落主題句出現在最後面。[Hajime and Manabu 2000]中所提及的"前導(Lead)"摘要方法，即是根據摘要比例選取文件前 N%的部分作為摘要。

其它的相關研究，例如 Hirohata 亦作過類似的研究分析[Hirohata et al. 2005]，他指出重要的文句多半落於文章的簡介與結論二個段落中，並且利用文句位置的資訊結合其他摘要方法來選取重要的文句。在[Maskey et al. 2003]論文中利用廣播新聞特定的文件結構與段落排列的方式，以廣播新聞中段落的所在位置(Position)、段落長度及每一段落的語者類別(Speaker Type)等結構特徵作為摘要特徵。

此種摘要方式簡單、直覺，但是僅適用於特定的結構內容且遵循著某種編排方式的文件。對於某些沒有特定結構編排的文件，或是缺乏文件結構的語音文件內容可能不適用。

2.2.2 以統計值為基礎的摘錄方法(Statistic-based Approach)

基於索引特徵統計值的摘要方法，是根據摘要索引特徵的統計值，如詞頻(Term Frequency, TF)、反文件頻(Inverse Document Frequency, IDF)、語言學分數(Linguistic Score)、辨識信心度(Confidence Score)及聲學特徵值(Prosodic Features Values)，來計算文句的重要性。可依文句重要性分數的估測或計算方式不同分為：

A. 相似度估測(Similarity Measure)

向量空間模型(Vector Space Model, VSM)[Ho 2003]、相關評估(Relevance Measure)[Gong et al. 2001]、最大臨界相關(Maximal Marginal Relevance, MMR)[Murray et al. 2005]均屬於此種摘要方法。概念是將文件中每一文句及整篇文件內容視為一 L 維向量，向量的每一維度代表某個索引特徵(詞、字或音節)在文句或是文件中的統計值；然後，根據每一文句向量 $\vec{S}_i$ 與整篇文件向量 $\vec{D}$ 的相似度(例如，餘弦值)來決定重要文句的選取。當文句與文件之間的相關程度越大，則表示此文句重要性越高。相似度的計算方式通常是採用餘弦估測法(Cosine Measure)，以文句及文件在向量空間上的夾角大小(如圖 2.3 所示)，作為二者之間的相關程度。

A.1：向量空間模型(Vector Space Model)

向量空間模型原應用於資訊檢索(Information Retrieval)[Salton et al. 1968]，用來估測使用者查詢(User Query)與文件(Document)之間的相關程度，藉此決定文件與查詢的相似度排名。[Ho 2003]將向量空間模型延伸應用於文件摘要中，並且探討不同權重值計算方式的索引特徵值對摘要正確率的影響。在向量空間摘要模型

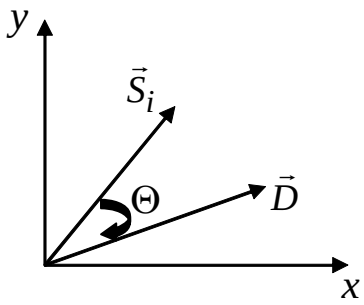

圖 2.3 向量空間模型示意圖

中，摘錄式摘要被視為一個檢索的問題，重要文句的選取順序以其與文件內容的相關程度而定；也就是將文件內容當成查詢去檢索最相關的 n 個文句，最後依相似度將文句串起來作為文件的摘要。下面式(2-1)與(2-2)分別代表整篇文件 D 與每一文句 S_i 的 L 維向量表示式：

$$\vec{D} = (\phi_1, \phi_2, \ldots, \phi_j, \ldots, \phi_L) \tag{2-1}$$

$$\vec{S_i} = (\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,j}, \ldots, \phi_{i,L}) \tag{2-2}$$

其中 ϕ_j 、 $\phi_{i,j}$ 分別代表文件 D 向量中索引特徵 j 的統計值(或權重)及文句 S_i 向量中索引特徵 j 的統計值。統計值通常使用詞頻(Term Frequency ,TF)或是詞頻與反文件頻乘積(Term Frequency Multiplied by Inverse Document Frequency, TF-IDF)來計算。詞頻，以符號 tf_j 表示，是基於索引特徵 j 於文件 D 或文句 S_i 中的出現次數來計算，根據索引特徵 j 出現的頻率來決定其重要性，當索引特徵 j 出現於文件 D 或文句 S_i 中次數愈多即代表愈重要。一般的計算方式可如式(2-3)或是式(2-4)所示，其中 $freq_j$ 表示索引特徵 j 於文件 D 或文句 S_i 中的出現次數：

$$tf_j = 1 + \log(freq_j) \tag{2-3}$$

$$tf_j = \frac{freq_j}{\max_l freq_l} \tag{2-4}$$

反文件頻，以符號idf_j表示，為索引特徵 j 的文件頻率(Document Frequency)之倒數。計算公式如下：

$$idf_j = \log\frac{N}{n_j} \tag{2-5}$$

其中，N 表示文件集中的文件總數，n_j 表示索引特徵 j 出現於文件集中的文件數。文件頻率是計算索引特徵 j 於文件集中出現過的文件篇數，用以表示索引特徵 j 是否具有鑑別力與資訊量。若索引特徵 j 在文件集中的每一篇文件都出現過(例如中文詞：在、的)，則代表它不是重要的詞且不具鑑別力，應降低其權重；反之，若索引特徵 j 僅在少數文件中出現過，代表它可能是重要的關鍵詞，代表某種主題或涵意。詞頻反文件頻是結合詞頻及反文件頻二種資訊，表示當索引特徵 j 越傾向於“在某文件或某文句中出現頻率高且不常出現在其他文件中(較具鑑別力)”越具重要性。公式表示如下：

$$tf\_idf_j = tf_j \cdot idf_j \tag{2-6}$$

[Ho 2003]中表示詞頻較適用於文字文件摘要，而語音文件摘要採用詞頻反文件頻作為權重會有較好的結果。

文件 D 與每一文句 S_i 的相似度，可由下面的餘弦估測法(Cosine Measure)來計算：

$$Sim(\vec{D},\vec{S}_i) = \frac{\vec{D}\cdot\vec{S}_i}{\left|\vec{D}\right|\times\left|\vec{S}_i\right|} = \frac{\sum_{j=1}^{L}\phi_j \times \phi_{j,i}}{\sqrt{\sum_{j=1}^{L}\phi_j{}^2}\times\sqrt{\sum_{j=1}^{L}\phi_{j,i}^2}} \tag{2-7}$$

其中，$\vec{S}_i$ 為文句 S_i 的向量表示式，$\vec{D}$ 為文件 D 的向量表示式，ϕ_j、$\phi_{j,i}$ 分別代表索引特徵 j 於文件 D 向量中及文句 S_i 向量中的統計值，L 為向量的維度。向量空間摘要模型會優先選取與文件 D 相似度(餘弦值)最高的文句作為摘要。

A.2：相關評估(Relevance Measure)

相關評估**[**Gong et al. 2001**]**與向量空間模型的方法類似，但是相關評估除了考慮

文件與文句之間的相似度外，亦考慮文句間概念主題重覆性的問題。相關評估方法的步驟與流程可參考圖 2.4 所示。其作法是將文件 D 中每一個未被摘錄的文句

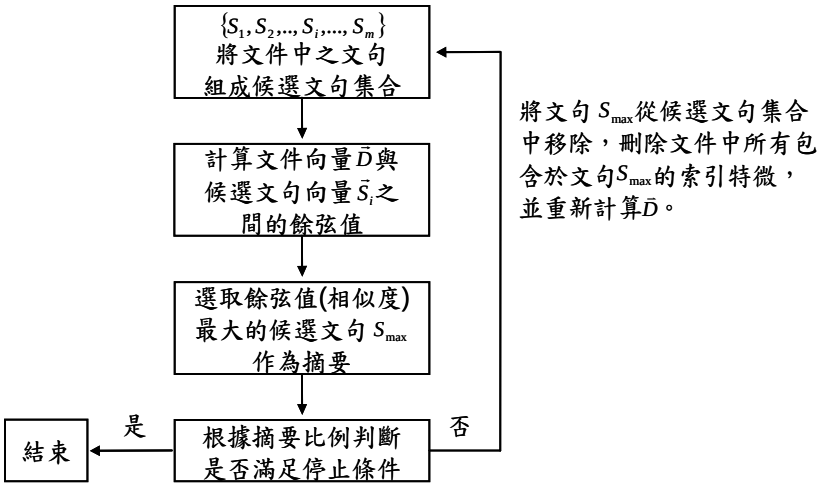

圖 2.4 相關評估方法的處理流程示意圖

S_i，分別與文件 D 計算二者之間的相似度並作排名，然後選出相似度最大的文句 $S_{\max}$ 作為摘要，同時亦將文件 D 中包含有文句 $S_{\max}$ 中的索引特徵全部移除，重新計算文件 D 的向量。接著重覆對未被選取的文句進行相同的步驟直到取出所需之 n 個文句後停止。相關評估方法的目的是為了選取出能夠代表文件不同主題的文句，透過移除文件中已摘錄文句中所包含的索引特徵，來減少摘要內容主題的重覆性。

A.3：最大臨界相關(Maximal Marginal Relevance, MMR)

最大臨界相關[Murray et al. 2005]在概念上與相關評估相同，均是考慮被選取文句間主題重覆性的問題，但是在作法上並不相同。最大臨界相關是分別計算欲選取文句與文件之間的相似度，及欲選取文句與已選取文句之間的相似度，希望所摘錄出之文句能與文件的相似度越大(越能表達整篇文件主題)，並且也與已選取文句間越不相似(避免主題重覆性)。其公式如下：

$$S^{MMR}(i) = a(Sim(\vec{S}_i, \vec{D})) - (1-a)(Sim(\vec{S}_i, \overrightarrow{Summ})) \quad (2\text{-}8)$$

其中 $Sim(\vec{S}_i, \vec{D})$ 表示文件與文句 S_i 之間的相似度，$Sim(\vec{S}_i, \overrightarrow{Summ})$ 為文句 S_i 與已選取文句集合之間的相似度，計算方式如式(2-7)所示。a 為權重參數，可設定為固定值，或是隨已選取文句數目不同而改變。

最大臨界相關及相關評估除了考慮文句與文件之間的相關程度外，亦考量主題重覆性的問題，為了選取代表文件不同主題的文句。但是，當文件中僅包含單一主題，則選出第一個重要文句後，可能會造成之後所選取的文句而反與真正的文件主題很不相關。又或者是當作為評估依據的參考摘要內容沒有考慮主題重覆性問題時，即使自動摘要系統真正選出代表不同主題的文句，也可能因內容與參考摘要有不一致的準則，而得到較差的摘要結果。

B. 計算文句特徵值分數

將文件中的每一文句以一連串索引特徵(如詞、字、或音節等單位)表示，計算文句中索引特徵的統計值的總和，作為文句的重要性分數，並以此分數做為文句選取的依據。文句中每個索引特徵的統計值，可以根據不同的權重參數值將不同的摘要特徵作結合(Combination)。如式(2-9)所示：

$$Score(w_j) = \lambda_1 F_1(w_j) + \lambda_2 F_2(w_j) + ... + \lambda_m F_m(w_j) \quad (2\text{-}9)$$

$Score(w_j)$ 表示索引特徵 w_j 的統計值分數，由不同摘要特徵值 $F_1, F_2, \ldots, F_m$，例如詞重要性分數、語言模型分數、辨識信心度，依不同權重值作加總。λ_m 為不同摘要特徵所對應的權重值。文句的重要性分數，可考慮是否以文句長度做正規化處理，若不經正規化處理則表示文句的選取偏向於長句優先；反之，文句的重要性則以平均索引特徵統計值為依據。

日本學者 Furui 等人所提出的語音文件摘要方法[Kikuchi et al 2003, Furui et al. 2004]，便是以文句的平均索引特徵統計值作為文句的重要性分數，來選取重

要的文句。方式是將文件中的每一文句以詞串表示，然後結合詞的語言學分數(Linguistic Score)、重要性分數(Significance Score)及辨識信心度分數(Confidence Score)，計算每一文句的平均詞統計值：

$$S_i = \frac{1}{J}\sum_{j=1}^{J}\left\{L(w_j) + \lambda_I I(w_j) + \lambda_C C(w_j)\right\} \tag{2-10}$$

J 為文句 S_i 的長度，$L(w_j)$ 代表詞 w_j 的語言學分數，$I(w_j)$ 代表詞 w_j 的重要性分數，$C(w_j)$ 代表詞 w_j 的辨識信心度，λ_I 與 λ_C 分別代表重要性分數與辨識信心度的權重值。其中語言學分數即為句中每一個詞的 n-連語言模型對數機率，如式(2-11)所示，是為了降低語言學上不自然或不流暢的文句字串之分數。通常採用二連語言模型(bigram)或三連語言模型(trigram)。

$$L(w_j) = \log P(w_j \mid .....w_{j-1}) \tag{2-11}$$

重要性分數表示文句中每一個詞的重要程度，計算方式如下：

$$I(w_j) = f_j \log\frac{F_A}{F_j} \tag{2-12}$$

其中 f_j 為詞 w_j 在文件中出現的次數，F_j 為詞 w_j 在文件集中出現的次數，F_A 表文件集中所有詞出現次數的總和。有意義詞是指像名詞、動詞、形容詞等較具有意義詞性的詞。辨識信心度分數是對於辨識結果可靠度的一個估測，此估測值為詞的辨識事後機率(Posterior Probability)，是以計算語音辨識所產生的詞圖(Word Graph)中，包含詞 w_j 的所有詞序列(Word Sequence)事後機率與詞圖中所有詞序列事後機率之比例，作為詞 w_j 的辨識信心度分數。計算公式如下：

$$C(w_j) = \frac{\sum_{W^* \in \overline{W}_{\Sigma^*}} P(X \mid W^*)P(W^*)}{\sum_{W \in \overline{W}_{\Sigma^X}} P(W)P(X \mid W)} \tag{2-13}$$

其中，$\overline{W}_{\Sigma^X}$ 代表所有可能的語音辨識結果詞序列(詞圖中所有的詞序列)集合，$\overline{W}_{\Sigma^*}$

代表所有包含有詞 w_j 的詞序列集合，W^* 為語音辨識結果中包含有詞 w_j 的詞序列；$P(W)$ 為詞序列 W 的語言模型機率，$P(X|W)$ 代表假設詞序列 W 為辨識結果的情況下產生 X 的機率。式(2-13)表示語音辨識所產生的所有可能詞序列中包含有詞 w_j 的詞序列之事後機率值。

另外，[Wu et al. 2005; Huang et al. 2005] 也採用類似的方式，使用不同摘要特徵值的分數，經由動態規劃演算法(Dynamic Programming Algorithm)為自動轉寫文件產生出精簡的文件摘要。考慮韻律學信心度(Prosodic Confidence)、辨識信心度、詞重要性分數、三連詞語言模型及語意相依信心度(Semantic Dependency Score)五種分數，期望在某特定摘要比例下，透過動態規劃演算法找出最適當的摘要結果，同時使得下列分數總和是最高的：

$$Summ(Y)=\sum_{n=1}^{N}\left\{\lambda_A A(w_n)+\lambda_C C(w_n)+\lambda_R R(w_n)+\lambda_L L(w_n\mid w_{n-2},w_{n-1})+\lambda_B B_{SDG}(w_{n-1},w_n)\right\} \tag{2-14}$$

$Summ(Y)$ 表示經由動態規劃演算法所產生出的摘要結果 Y 的分數總和，X 為摘要結果 Y 的長度，$A(w_n)$ 代表詞 w_n 韻律學信心度，$C(w_n)$ 代表辨識信心度，$R(w_n)$ 代表詞重要性分數，$L(w_n\mid w_{n-2},w_{n-1})$ 代表三連詞語言模型分數，$B_{SDG}(w_{n-1},w_n)$ 代表語意相依度分數。λ_A、λ_C、λ_R、λ_L 及 λ_B 為權重參數。

韻律學信心度包含能量估測(Energy Measure)與音高估測(Pitch Measure)。人在說話時通常會因為要特別強調重要或關鍵的詞而提高音量，語音訊號的能量可以反映出這樣的“強調”，某個詞所對應的能量越高可能代表其越重要，計算公式如下[Wu et al. 2005]：

$$e(w_n,f_{n_s},f_{n_e})=\frac{1}{f_{n_e}-f_{n_s}}\sum_{l=f_{n_s}}^{f_{n_e}}\frac{(e_l-e_{\min})}{(e_{\max}-e_{\min})} \tag{2-15}$$

$e(w_n,f_{n_s},f_{n_e})$ 代表詞 w_n 的能量值，f_{n_s}、f_{n_e} 分別表示詞 w_n 的起始與結束音框

(Frame)，$f_{n_e} - f_{n_s}$ 為詞 w_n 所佔的音框總數，e_l 為第 l 個音框的能量，$e_{\max}$ 與 $e_{\min}$ 分別表示語音文件中最大及最小的能量值。同樣地，一般人在說話時都會有高低起伏，對於重要的地方可能會特別提八度音來表示，音高即是用來表示人說話的聲調(例如低沉或高亢)，計算公式如下[[Huang et al. 2001；Wu et al. 2005]：

$$Ph_l = 80\log_{10} F_0(l) \tag{2-16}$$

$$Ph(w_n, f_{n_s}, f_{n_e}) = \frac{1}{f_{n_e} - f_{n_s}} \sum_{l=f_{n_s}}^{f_{n_e}} 24 \times \frac{(Ph_l - Ph_{\min})}{(Ph_{\max} - Ph_{\min})} \tag{2-17}$$

其中，音高值會先取對數值(Logarithmic Value)，如式 2-16 所表示，其中 $Pitch(l)$ 為第 l 個音框的音高值。式 2-17 中，$Ph(w_n, f_{n_s}, f_{n_e})$ 表示音高值以文件中最小與最大值作正規化處理後的結果，f_{n_s} 、f_{n_e} 分別表示詞 w_n 的起始與結束音框(frame)，$f_{n_e} - f_{n_s}$ 為詞 w_n 所佔的音框總數，$Ph_{\max}$ 與 $Ph_{\min}$ 分別表示語音文件中最大及最小的音高。最後，將能量估測值與音高估測值以線性組合的方式結合，作為聲韻特徵分數：

$$A(w_n) = \alpha_1 e(w_n) + \alpha_2 P(w_n) \tag{2-18}$$

其中，α_1 與 α_2 為權重參數，用來結合能量估測值與音高估測值。

在[Wu et al. 2005; Huang et al. 2005]中，詞重要性分數的定義與式(2-12)不同；其考慮了詞頻反文件頻的資訊，如式(2-5)，以及與代表文句主題的關鍵詞($w_b^{t^*}$)之間的相似度來計算重要性分數。當詞在文句中出現頻率高、不常出現在其他文件中，也與文句的概念主題越相近，代表越重要且重要性分數越高。計算方式如下：

$$R(w_n) = \max_b \left\{ Sim(\vec{w}_n, \vec{w}_b^{t^*}) \cdot tf\_idf_n \right\} \tag{2-19}$$

tf_idf_n 表示詞 w_n 的詞頻反文件頻，其敘述與計算方式可參照式(2-3)至式(2-6)。相似度分數 $Sim(\vec{w}_n, \vec{w}_b^{t^*})$ 的計算如式(2-7)所示，是以餘弦估測方法來計算潛藏語

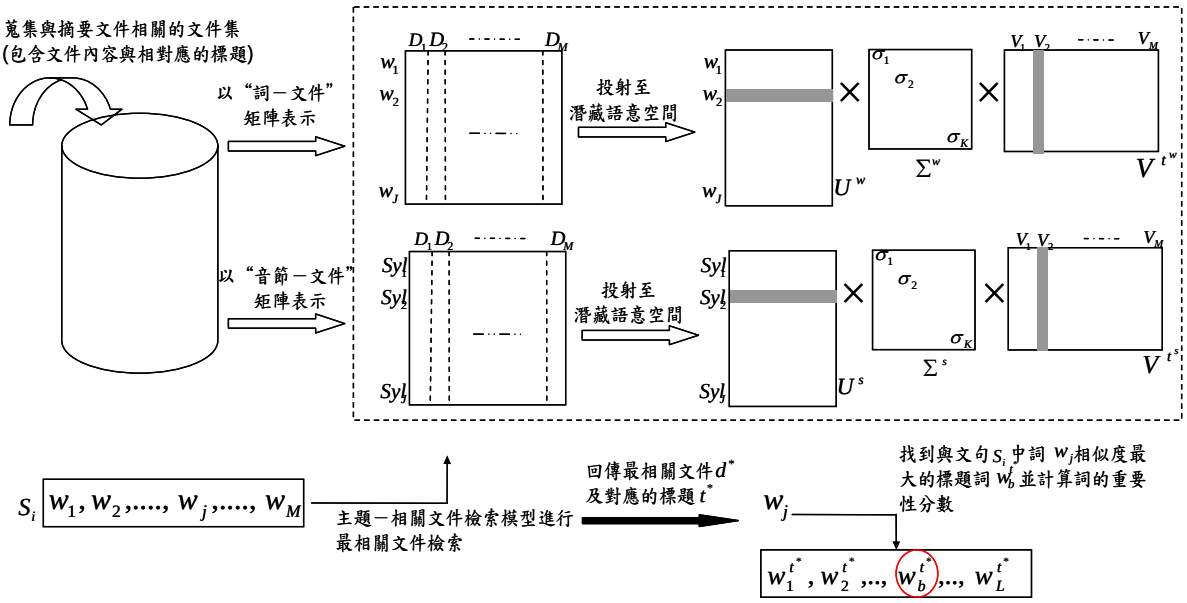

圖 2.5 詞重要性分數計算流程圖

意空間(Latent Semantic Space)中，詞 w_n 向量 $\vec{w}_n$ 與詞 $w_b^{t^*}$ 向量 $\vec{w}_b^{t^*}$ 之間的相似度。其中詞 $w_b^{t^*}$ 代表在概念上與文句最相近的關鍵詞，詞 $\vec{w}_b^{t^*}$ 為詞 $w_b^{t^*}$ 於潛藏語意空間中的向量表示式。作法如圖 2.5 所示，先將訓練文件集(其中每一文件有其所對應的標題資訊)分別以“詞－文件”矩陣 A^w 及“音節－文件”矩陣 A^s 表示，然後透過奇異值分解(Singular Value Decomposition, SVD)將二個矩陣投射到低維度的潛藏語意空間，分別以 A^{w^*} 與 A^{s^*} 表示；接著以“主題－相關文件檢索模型(Topic-Related Document Retrieval Model)” [Manning and Schutze 1999]，從文件集中找出與文句 S_i 最相關的文件 d^* 及所對應的標題 t^*，同時找出標題 t^* 中與詞 w_n 最相似的詞 $w_b^{t^*}$ (經由餘弦估測)。最後，詞 w_n 的重要性分數就是詞 w_n 與詞 $w_b^{t^*}$ 之間的相似度

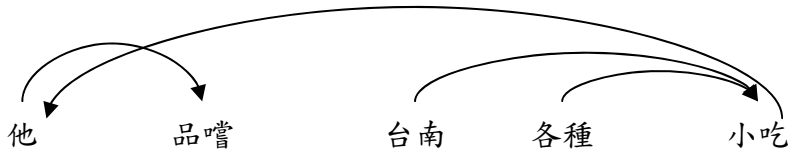

圖 2.6 詞與詞語意相依關係圖

乘上 w_n 之文件反文件頻。

三連詞語言模型分數，是為了評估文句中詞序列的語意相依性，與[Kikuchi et al 2003; Furui et al. 2004]中使用的語音學分數之意義相同。[Wu et al. 2005; Huang et al. 2005]使用三連詞語言模型，計算方式如式(2-20)：

$$L(w_j \mid w_{j-2}, w_{j-1}) = p_1 \frac{F(w_{j-2}, w_{j-1}, w_j)}{F(w_{j-2}, w_{j-1})} + p_2 \frac{F(w_{j-1}, w_j)}{F(w_{j-1})} + p_3 \frac{F(w_j)}{\sum_z F(w_z)} \tag{2-20}$$

為結合三連詞語言模型、二連詞語言模型及單元詞語言模型的機率。其中 $F(\cdot)$ 代表某 N 連詞序列出現的次數， p_1 、 p_2 、 p_3 為權重參數且 $p_1 + p_2 + p_3 = 1$。

最後，語意相依度分數是為了找出詞與詞之間的語意關係，如圖 2.6 所示。是採用語意相依文法(Semantic Dependency Grammar，SDG)來找出詞與詞之間的語意或文法關係，關於詳細的作法可參照[Manning and Schutzeet al. 1999; Wu et al. 2005]。

下面，列出目前較常被使用的摘要特徵，可被細分為以下幾大類：

I. *結構特徵*：利用文件的結構資訊作為摘錄重要文句的資訊 [Maskey et al. 2003]，例如文句所在位置，若文句位於重要段落則有較高權重。或是廣播新聞中不同語者(Speaker)的資訊、文句長度等。

II. *統計量特徵*：如詞頻、反文件頻、詞頻反文件頻等統計資訊 [Ho 2003]，或是利用統計資訊所計算的詞重要性分數，如式(2-12)與式(2-19)的計算

方式 [Wu et al. 2005；Huang et al. 2005; Kikuchi et al 2003; Furui et al. 2004]。

III. *語言學的分數*：以文件或文句中的文法、語意等自然語言相關的資訊作為特徵值，如類專有名詞 [Maskey et al. 2005]的使用，對於文句中的類專有名詞可給予較高的權重分數，或將文句中出現類專有名詞的個數佔全文句長度的比例作為一種摘要特徵 [Maskey et al. 2005]，或是如[Kikuchi et al 2003]中所提出的語意相依性的分數作為摘要特徵。

語言模型分數：使用訓練文件集所計算的語言模型(Language Model)分數[Kikuchi et al 2003；Furui et al. 2004]。例如二連語言模型(Bigram Language Model)或三連語言模型(Trigram Language Model)。

IV. *辨識信心度*：用來估測辨識結果可靠度，可作為摘要中重要的資訊，以選取可靠、正確的辨識結果 [Furui et al. 2004]。

V. *聲韻特徵*：由語音訊號中所取得的資訊，例如重音、強度、音高、文句訊號的持續時間(Duration)[Maskey et al. 2005]。

C. 潛藏語意概念估測(Latent Semantic Analysis Measure)

目的在於找出與文件的語意概念最相近的文句。在很多語言(例如中文)中，都有遇到同義字、詞的問題，不同的字、詞有許多相同語意但不同形式的字、詞。若僅以詞的形式來估測與文件最相近的文句，而不是從詞意上來看，可能會錯失了真正與文件主題或概念相似的文句[Furnas et al. 1988]。

C.1：潛藏語意分析模型

VI. 因此，從 2001 年開始便有學者[Gong et al. 2001]提出了潛藏語音概念的方法應用於文件摘要上。[Gong et al. 2001]所提出的潛藏語意分析模型(Latent Semantic Analysis, LSA)的摘錄方法，是將文件以“索引特徵－文句”矩陣表示，如圖 2.7 中最左邊的矩陣，矩陣中每一行代表文件中的某一文句，而文句中每一維度代表某

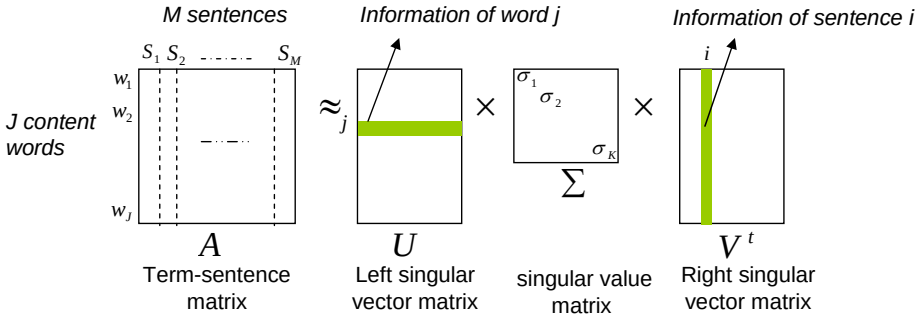

圖 2.7 奇異值分解(SVD)圖示

個索引特徵(詞、字或音節)在文句的統計值，例如詞頻、詞頻反文件頻，可以式(2-6)計算。然後透過奇異值分解(Singular Value Decomposition, SVD)將矩陣投射到低維度的潛藏語意空間中，如此，文件中的每一文句即可被表示為潛藏語意空間中的一個向量，每一文句中的每一維度代表某個潛藏語意概念所對應的索引值。如圖 2.7 所示，經奇異值分解後，每一奇異值及其對應的奇異向量(Singular Vector)代表一概念(值越大越重要)，文件中每一文句可由右奇異矩陣轉置的行向量表示，可依所對應奇異值由大至小，從右奇異向量(右奇異矩陣轉置的列向量)取出含最大索引值的文句作為文件的摘要。

C.2：以奇異值分解為基礎之維度化簡法

另外，在 2005 年[Hirohata et al. 2005]提出另一種使用潛藏語意概念的摘錄方法－以奇異值分解為基礎之維度化簡法(Dimension Reduction based on SVD，DIM)。同樣是將原文件以“索引特徵－文句”矩陣表示，再經奇異值分解將矩陣投射至低維度的潛藏語意空間後，原文句可以潛藏語意空間中的向量來表示，如圖 2.8 所示，原 M 維的文句向量以 N 維的潛藏語意向量來表示(N<M)，然後再根據 N 維

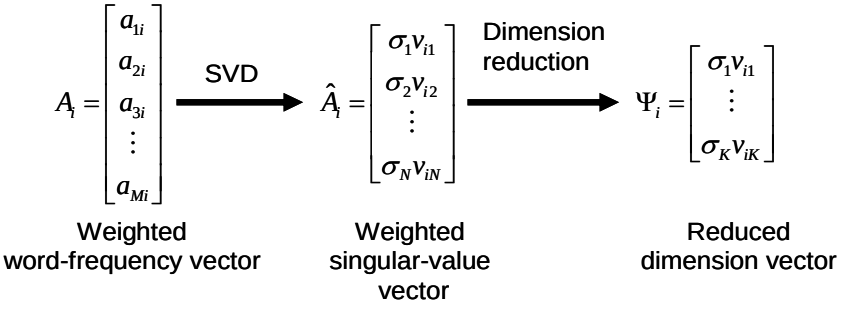

圖 2.8 以奇異值分解為基礎之維度化簡法

的潛藏語意向量來計算每一文句的分數，計算方式是取前 K 維索引值平方和的平方根，即是計算文句的向量範數(Norm):

$$S_i = \sqrt{\sum_{k=1}^{K} (\sigma_k v_{ik})^2} \tag{2-21}$$

最後，每一文句以所計算出來的範數作為重要性排名。

C.3：嵌入式潛藏語意分析模型(embedded LSA, eLSA)

[黃耀民 2005; 陳怡婷 et al. 2005]亦提出所謂的嵌入式潛藏語意分析模型(embedded LSA, eLSA)應用於文件摘要，嵌入式潛藏語意分析模型將被摘要的文件以“索引特徵－文句&文件”矩陣表示(如圖 2.9 所示)，經奇異值分解投射到低維度的潛藏語意空間，如此，每一文句以及文件本身皆可以潛藏語意空間中的向量來表示，最後文句的選取以每一文句與整篇文件在潛藏語意空間中向量的距離(採用餘弦值估測法計算)來決定。圖 2.9 中 A 表示索引特徵－文句&文件矩陣，矩陣 A 與 V^t 的最後一行向量分別代表整篇文件在原來向量空間與潛藏語意空間的表示式。

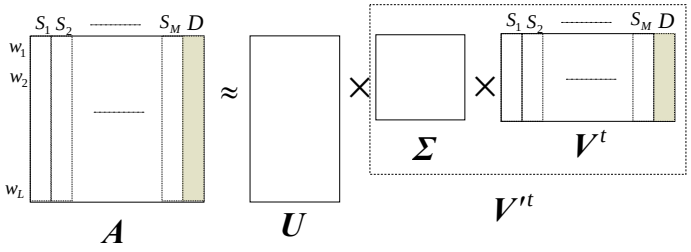

圖 2.9 嵌入式潛藏語意分析模型

D. 以分類器為基礎(Classifier-based)的摘要方法

將重要文句的選取視為一種分類的問題，利用訓練語料根據許多摘要特徵，例如結講特徵、統計量特徵、語言學模型、及聲韻特徵，訓練出具鑑別性的分類器，能夠正確的將文句分類成摘要文句類別與非摘要文句類別 [Koumpis et al. 2005]。最後，被分類為摘要文句類別的文句即可視為摘要內容。目前已有許多學者提出應用不同的分類器於自動文件摘要中，例如支援向量機(Support Vector Machine, SVM)[Zhu and Penn 2005] 、邏輯迴歸(Logistic Regression) [Zhu and Penn 2005]、高斯混合模型(Gaussian Mixture Model, GMM)[Murray et al. 2005] 及貝式網路分類器(Bayesian Network classifier) [Maskey et al. 2005]。

此種方法必需有標註摘要資訊的文件集以訓練分類器，通常有標註的訓練文件集取得不易，同時也可能會遇到訓練語料不足的問題。同時，所訓練出來的分類器可能會產生領域相依(Domain-Dependent)的問題，即使用科技相關的文件所訓練出來的摘要分類器，可能不適用於新聞文件上。此外，某些分類器的輸出只有 1 或 0(是或否的結果)，對於每一文句無法給予一個重要性的分數或機率值，因此無法對文句作重要性排名，也無法根據摘要比例來選取特定的句數作為摘要結果。

2.2.3 以機率生成模型為基礎的摘錄方法(Probabilistic Generative Model-based Approach)

機率生成模型以最大事後機率法(Maximum A Posteriori, MAP)為基礎，來表示文句與文件之間的關係。當給一篇欲摘要之文件 D 時，文件 D 與文句 S_i 的關係可以表示成：

$$P(S_i|D) = P(D|S_i)P(S_i) \tag{2-22}$$

機率值 $P(D|S_i)$ 為文句 S_i 生成文件 D 的可能性(Likelihood)，$P(S_i)$ 為文句 S_i 的事前機率值，可簡單地假設其為一致性分佈，使每一文句 S_i 的事前機率值相同。如此，每一文句與文件之間的相關程度，以文句機率模型生成文件的可能性來決定，其中文件中每一文句視為一個機率生成模型(Probabilistic Generate Model)，表示文句中索引特徵(音節、字、或是詞)的分佈，而文件中的索引特徵當成是一連串的觀測值(Observation)。例如，隱藏式馬可夫模型(Hidden Markov Model, HMM)與主題混合模型(Topical Mixture Model, TMM)皆屬於此種摘要方法。

A. 隱藏式馬可夫摘要模型

我們過去將隱藏式馬可夫模型用於中文語音資訊檢索(Spoken Document Retrieval, SDR)，獲得不錯的成果[Chen et al. 2004; 陳怡婷 et al. 2005]；之後，我們亦將其延伸應用至語音文件自動摘要[黃耀民 2005; 陳怡婷 et al. 2005]。隱藏式馬可夫模型將每篇文件 D 中每一文句 S_i 視為一個機率生成模型，如圖 2.10 所示，每一文句 S_i 對於每一個索引特徵 w(可為音節、字、詞或數個音節的組合)都會產生一個機率值。被摘要的文件與此文句的相關程度，是藉由文件中所有的索引特徵 w 在文句 S_i 發生可能性來決定。文句 S_i 可以看成有兩個狀態(States)，分別對應到兩個機率分佈，其中 $P(w|S_k)$ 為索引特徵 w 在文句 S_i 發生的機率；$P(w|C)$ 為索引特徵 w 在整個文件集(或語料)發生的機率，用來平滑化(Smoothing)索引特徵 w 在文句 S_i 發生的機率，同時表示索引特徵 w 在語言裡的統計資訊。因此，當文件中的索引特徵在此文句的機率分佈值連乘積越高，則文件與此文句的

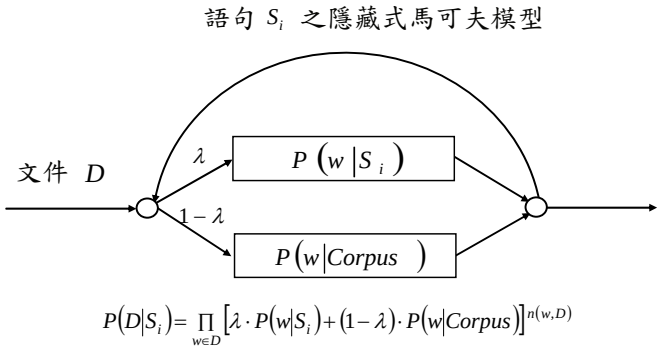

圖 2.10 隱藏式馬可夫模型

相關度就越高，其公式如下：

$$P(D\mid S_i)=\prod_{w\in D}\left[\lambda\cdot P(w\mid S_i)+(1-\lambda)\cdot P(w\mid C)\right]^{n(w,D)} \tag{2-23}$$

$n(w,D)$為索引 w 在 D 出現的次數，λ 是權重參數(Weighting Parameter)。其中參數 λ 與每一文句 S_i 產生各索引 w 的機率值 $P(w_i\mid S_i)$，可以透過期望值最大化(Expectation-Maximization, EM)訓練[Dempster et al. 1997]自動調整。

當有訓練文件集(文件及其對應的摘要資訊)時，隱藏式馬可夫摘要模型可以透過監督式(Supervised)方式進行模型的參數訓練，但是在一般實用的情況下，這些訓練所需的相關資訊不易取得，因此可以採用非監督式(Unsupervised)模型訓練方式。文件中的每一文句可視為一連串觀測值，用來訓練文句本身的隱藏式馬可夫模型，模型參數 λ 與每一文句 S_i 產生索引 w 的機率值 $P(w\mid S_i)$ 可以藉由期望值最大化法自動訓練與調整，其數學式為：

$$\hat{\lambda}=\frac{\sum_{w\in S_i} n(w,S_i)P(S_i\mid w,D)}{\sum_{w_j\in S_i} n(w_j,S_i)} \tag{2-24}$$

$$\hat{P}(w\mid S_i)=\frac{n(w,S_i)P(S_i\mid w,D)}{\sum_{w_s\in S_i} n(w_s,S_i)P(S_i\mid w_s,D)} \tag{2-25}$$

$$P(S_i \mid w, D) = \frac{\lambda P(w \mid S_i)}{\lambda P(w \mid S_i) + (1-\lambda) P(w \mid C)} \tag{2-26}$$

$n(w, S_i)$ 表示索引 w 於文句 S_i 中出現的次數。

B. 主題混合模型

主題混合模型最早由[Chen et al. 2004; Chen 2006]所提出並且使用於語音資訊檢索，[Chen et al. 2006; 黃耀民 2005; 陳怡婷 et al. 2005]將它延伸用於語音文件自動摘要。當給定一篇欲摘要的文件 D 時，主題混合模型將文件 D 中每一文句 S_i 視為一個包含了 K 個潛藏主題的混合模型(Mixture Model)，其中每一潛藏主題 T_k 可分別由一個單連語言模型 $P(w \mid T_k)$ 所表示，而且每一文句 S_i 對於每一潛藏主題 T_k 有不同的權重 $P(T_k \mid S_i)$，如圖 2.11 所示。文件 D 與每一文句 S_i 的相關程度可被表示為：

$$P(D|S_i) = \prod_{w \in D} \left[\sum_{k=1}^{K} P(w|T_k) P(T_k|S_i) \right]^{n(w,D)} \tag{2-27}$$

由上式可看出，主題混合模型不同於隱藏式馬可夫模型所採用的逐字比對方式(Literal Term Matching)。隱藏式馬可夫模型中文件 D 與每一文句 S_i 的相關程度取決於 $P(w \mid S_i)$，也就是文件 D 中的索引出現在文句 S_i 的機率；在主題混合模型

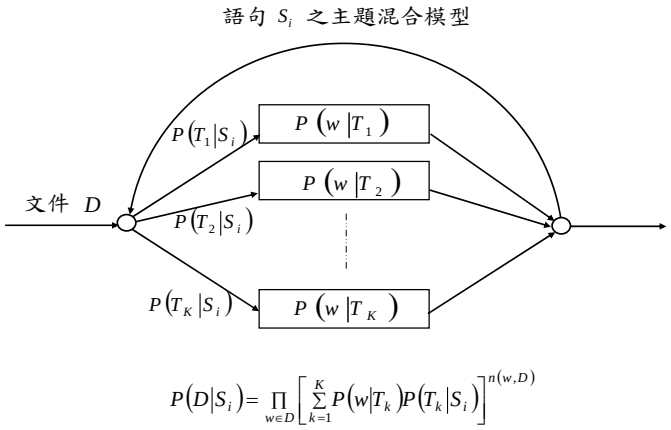

圖 2.11 主題混合模型

中，文件D與每一文句S_i的相關程度取決於D中的索引出現於某一個主題的機率$P(w|T_k)$，以及文句S_i對於此潛藏主題T_k的權重值$P(T_k|S_i)$，當兩者的乘積越大代表文件D與文句S_i愈相關，即使文件D中的大部分索引並未出現於文句S_i中。因此，使用主題混合模型可以達到概念比對(Concept Matching)的目的。主題混合模型中的機率值$P(w|T_k)$、$P(T_k|S_i)$同樣可以使用期望值最大化法(EM)來訓練。

此外，由於我們可以從網絡上得到與被摘要語音文件(例如廣播新聞)同一時期(Contemporary)或同一領域(In-Domain)的文字新聞文件集$\{D_C\}$，這些新聞文件通常都包含有一句人工產生的標題，它可視為一個非常簡潔的文件摘要資訊，於是上述文件集中每篇文件D_C與其標題H_C的對應關係$\{D_C, H_C\}$，可作為主題混合模型的訓練使用。我們將每篇文字新聞文件的標題視為一個主題混合模型來產生文件本身：

$$P(D_C|H_C)=\prod_{w\in D_C}\left[\sum_{k=1}^{K}P(w|T_k)P(T_k|H_C)\right]^{n(w,D_C)} \tag{2-28}$$

其中$n(w,D_C)$是索引w在文件D_C出現的次數。我們假設文字新聞文件的標題模型間共用相同的主題機率分佈$P(w|T_k)$，而各自對於每一潛藏主題T_k有不同的權重$P(T_k|H_C)$。於是，模型參數的期望值最大化訓練公式可用下列數學式表示：

$$\hat{P}(w|T_k)=\frac{\sum_{D_C\in\{D_C\}}n(w,D_C)P(T_k|w,H_C)}{\sum_{D_C\in\{D_C\}}\sum_{w'\in D_C}n(w',D_C)P(T_k|w',H_C)} \tag{2-29}$$

$$\hat{P}(T_k|H_C)=\frac{\sum_{w'\in D_C}n(w,D_C)P(T_k|w,H_C)}{\sum_{w'\in D_C}n(w',D_C)} \tag{2-30}$$

$$P(T_k|w,H_C)=\frac{P(w|T_k)P(T_k|H_C)}{\sum_{l=1}^{K}P(w|T_l)P(T_l|H_C)} \tag{2-31}$$

$P(T_k|w,H_C)$為在給定標題H_C的主題混合型與詞w的條件下，潛藏主題T_k發生的機率。

當在從事語音文件摘要時，若我們可以取得與被摘要語音文件同一時期或同

一領域的文件集$\{D_C\}$，則文句S_i主題混合模型中機率值$P(w|T_k)$，可以直接使用式(2-29)所訓練的$P(w|T_k)$機率，而機率值$P(T_k|S_i)$可以假設為一致性分佈，給予相同的機率值，或是同樣將每一文句視為與自己相關的資訊，以期望值最大化法來進行非監督式訓練：

$$P(T_k|S_i)=\frac{\sum_{w\in S_i} n(w,S_i)P(T_k|w,S_i)}{\sum_{w\in S_i} n(w,S_i)} \tag{2-32}$$

$$P(T_k|w,S_i)=\frac{P(w|T_k)P(T_k|S_i)}{\sum_{l=1}^{K} P(w|T_l)P(T_l|S_i)} \tag{2-33}$$

2.2.4 重要文句的精簡與壓縮(Sentence Compaction)

如上面所述，我們已經將常用的許多摘錄式摘要方法作概略的介紹，經由重要文句的選取，可簡單地產生文件的摘要內容。但是，對於文句中多餘的字詞(例如綴詞、形容詞)，及語音文件中多餘的語助詞、辨識錯誤等無意義的內容，並沒有被考慮與處理，使得被產生出的摘要內容往往包含有雜訊，不夠精簡。因此，有學者進一步提出重要文句化簡與壓縮的方法[Knight and Marcu 2002; Kikuchi et al 2003; Furui et al. 2004]，來產生更為精確的摘要內容。

[Knight and Marcu 2002] 提出二種文句壓縮的方法，噪音管道模型(Noisy-Channel Model)及以判斷為基礎的模型(Decision-Based Model)，用來找出最精簡的文句，並且同時保有文法結構與重要資訊。前者是將文句壓縮的問題視為：包含一長串詞的文句原來是一句簡短的詞串文句，然後被加上了一些額外詞或內容(例如綴詞)，形成最後文句。因此，文句壓縮便是要找出原來的精簡的文句。其作法是先以文法樹(Parse Tree)的方式來表示一個文句，再由其中找出一棵可有效表達文句概念的子樹(Subtree)。後者將文句壓縮看成是一個文句重寫(Rewrite)的問題。其中，重寫的操作可被分解成一連串的動作與步驟，可透過訓練語料(訓練文件與其所對應的摘要)將這些重寫的動作與步驟，建立出一套規則(Rules)用來進行文句的重寫。

日本學者[Kikuchi et al 2003; Furui et al. 2004]，提出使用多種摘要特徵，如語言學分數、重要性分數、信心度分數及詞連續分數(Word Concatenation Score)，以動態規劃法(Dynamic Programming Technique)依據不同的壓縮比例(Compression Ratio)，來找出最佳與最適當的壓縮文句。其中，語言學分數與重要性分數是希望遵循著語言學特性及保留文句中重要的詞，信心度分數的使用是用來選擇出辨識正確的內容，詞連續性的分數是希望所產生的摘要，內容順暢並具連貫性。對於重要文句 S_i ($S_i = w_1, w_2, ..., w_N$)，每一個可能的壓縮文句 $V = v_1, v_2, ..., v_M$ 可表示為：

$$S(V) = \sum_{m=1}^{M} \{ L(v_m \mid ....v_{m-1}) + \lambda_I I(v_m) + \lambda_C C(v_m) + \lambda_T Tr(v_{m-1}, v_m) \} \qquad (2\text{-}34)$$

其中，M 為壓縮文句的長度($M < N$)；$L(v_m \mid ....v_{m-1})$ 為語言學分數，$I(v_m)$ 為重要性分數，$C(v_m)$ 代表辨識信心度，$Tr(v_{m-1}, v_m)$ 代表詞連續分數，λ_I 、λ_C 及 λ_T 分別表示所對應的權重值。動態規劃法為從所有可能的壓縮文句中 $S(V)$ 找到分數最大的那一個。

2.2.5 本節小結

上述的摘錄式方法被應用於文字文件摘要及語音文件摘要中，各有其優劣；同時，可結合文句壓縮的方法來進一步精簡文句的內容。許多針對於文字文件所探討的摘要方法，一般皆可被應用於語音文件摘要，只是語音文件的特性與文字文件不盡相同，像是語音文件中的辨識錯誤，或是文句邊界定義的問題，因此可能無法達到很好的摘要結果。論文中，便是希望針對二者之間的差異性，探討與研究適用於語音文件的摘錄式摘要方法。我們提出以機率生成模型為基礎的摘要方法，並證明對於文字文件與語音文件摘要皆有較顯著的結果，並較其他摘要方法比較有較好的摘要結果。

除了針對文字文件及語音文件上所探討的摘要方法，目前亦有學者提出使用隱藏式馬可夫模型，直接對語音訊號內容作摘要的處理[Maskey et al. 2006]。方

法是將摘要處裡視為一個二分法的問題，語音訊號依某些規則切段後，根據所抽取出一些的語音聲韻特徵(Prosodic Features)來判斷每段語音訊號是否屬於摘要內容，其中不需要經過語音辨識的過程將訊號內容自動轉寫，亦不使用詞彙特徵(Lexical Feature)。此種摘要方法為語音、影音檔案開啟了另一個研究的方向與新的思維。如此，語音檔案的摘要系統(無論是針對自動轉寫文件或是語音訊號內容)將還有很大的空間值得研究與探討。

2.3 自動摘要的評估方法

近幾十年來，自動摘要的方法不斷的被廣泛探討與研究，許多學者提出不同的摘要方法，來提高摘要對於表達文件內容的正確性及代表性，但是摘要結果的優劣該如何評斷？不同摘要方法間的好壞又該如何比較？文件摘要結果需要一個公平且客觀的方式來評估，如此才能不斷地提昇文件摘要的研究成果。直至目前為止，自動摘要的評估仍沒有一套完整且統一的方法，許多文獻所使用的評估方式不一，評估結果亦不相同[Zhu and Penn 2005]。目前被大家所廣泛使用的評估方式，大略可分為主觀評估與客觀評估二類。主觀評估是指以人的主觀判斷與評估作為自動摘要的評估結果，而客觀評估通常以計算摘要正確率作為評估的結果，系統採用某種計算方式，自動地計算自動摘要與參考摘要之間的相似或相關程度，作為摘要的正確率。所謂的參考摘要，通常指經由人工標註或重寫的文件摘要，作為計算摘要正確率的參照。因此，參考摘要的產生亦可能包含著某種程度上的主觀因素，因為每一位標註者對文件摘要的定義並不相同。為了避免此問題，計算摘要正確率時可採用多份參考摘要來進行評估，藉由與不同的參考摘要正確率的平均結果，作為摘要系統的評估結果。

參考摘要依其形成的方式可分為摘錄式摘要(或稱為句排名式摘要)及非摘錄式摘要(或稱為重寫摘要)。摘錄式摘要是標註者由原文件中，標註或選出重要的文句、段落或章節所形成的摘要，可依不同的摘要比例而產生不同長度的摘要

內容，例如 10%、20%、30%、50% 等摘要比例。本論文中所使用的摘錄式參考摘要，即是標註者根據文件中文句的重要性將文句作排名。非摘錄式摘要為標註者根據特定的摘要比例(例如 20%~30%)，針對文件的內容所重寫的摘要內容，其長度、內容固定無法依摘要比例作任意的更改，同時，摘要中所使用的字、詞可能與文件內容不相同。本論文所使用的重寫摘要，是標註者依據摘要比例約25%所重寫，而且被要求盡量使用文件中所使用的字、詞來表示。

2.3.1 主觀評估

主觀評估是由評估者對自動摘要結果進行評估，評估的依據包括：1. 摘要內容是否符合文件中重要的資訊內容；2. 摘要內容是否涵蓋文件中所有重要主題；3. 摘要內容是否表達出整篇文件的重點；4. 摘要是否通順流暢且易懂等。評估方式通常是將自動摘要結果作不同等級的評分，[Hirohata et al. 2005]將評估結果分為五種等級，依序為很好、好、普通、差、很差，評估者依照自己的想法決定摘要結果所屬的等級。另外，[Wu et al. 2005]是採用分數評分的方式，依照摘要結果給定 0 到 10 之間的分數，最好摘要內容的給 10 分，最差的給 0 分。為了避免單一評估者的評估結果之可信度及客觀性不足，一般會由多位評估者同時進行摘要結果的評估，然後以大多數評估者的結果作為自動摘要的評估結果。

2.3.2 客觀評估

常見的客觀評估方法有：摘要準確率(Summarization Accuracy)[Hori et al. 2004; Hirohata et al. 2005]、文句的召回率(Recall)／精確率(Precision)[Hirohata et al. 2005]、ROUGE 評估方法 [Lin et al. 2003; Lin 2003; Hirohata et al. 2005]及餘弦評估(Cosine Measure)[Saggion et al. 2002; Ho 2003; Steinberger et al. 2004]等四種方法。下面將略述四種方法的概念與評估方式：

I. 摘要正確率：計算自動摘要與參考摘要之間的準確率。作法是將自動摘

要與參考摘要分別以詞串來表示，然後計算二詞串之間的準確率，計算公式為：

$$ACCY = \frac{Len - (Sub + Ins + Del)}{Len} \times 100 \quad (2\text{-}35)$$

其中 *Len* 表示參考摘要的長度(即詞的個數)，*Sub* 、*Ins* 、*Del* 分別為自動摘要與參考摘要之間調準(Alignment)後的代換(Substitution)個數、插入(Insertion)個數及刪除(Deletion)個數。對於多份可供評估的參考摘要，可以選定其中一份與自動摘要內容最相近的來計算，亦可分別計算自動摘要與每一份參考摘要的準確率，然後以其平均值或是最大值作為評估結果。

另外，[Hori et al. 2004; Hirohata et al. 2005]是先將多份的參考摘要合併成一個詞網(如圖 2.12 所示)，然後由詞網中選出一條與自動摘要最相近的詞串並計算二者之間的詞準確率(Word Accuracy)，如式(2-35)所示。同樣地，除了可選定一條最相近的詞串來計算外，也可選用多條詞串計算詞正確率，然後再取最大值或平均值。詞網的方式可以將多份的參考摘要組合成更多的

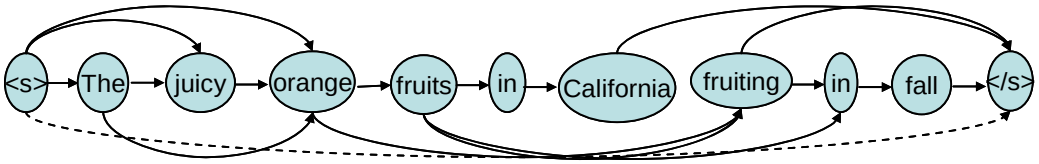

圖 2.12 參考摘要之詞網

詞串選擇，如此，即使自動摘要的內容與個別的參考摘要之間的相似度不高，但是對於涵蓋了多份參考摘要語意的自動摘要內容來說，仍然可以找到一條極為相近的詞串，而得到較高的準確率。

此方法的缺點為每份參考摘要的內容差異甚大且長短不一致，建立的詞網也許適用於高摘要比例(例如摘要比例 50%)的評估，卻可能極不適用於低

摘要比例(例如摘要比例 10%)上的評估。例如，若每一份參考摘要皆由摘要者產生摘要比例 10%~70%等長度不等的文件摘要，當所建立之詞網用於評估摘要比例 10%的自動摘要內容時，卻找到最相近的詞串為 50%的摘要內容，即使此自動摘要內容與文件主題相符，所計算出的準確率仍會因太多的刪除(Deletion)個數而非常的低。此外，由於詞網可能產生的摘要組合有很多種，但是，並不是每一條詞串均是符合文法、語意的摘要內容。其他類似的計算方法可以參考[Zhu and Penn 2005]。

II. 文句的召回率／精確率：計算被正確摘錄出的文句佔參考摘要中文句數的比例(召回率)，以及正確摘錄出的文句佔被摘錄出文句數的比例(精確率)。一般以召回率、精確率或 F-評估(F-Measure)作為評估結果，計算方式如下：

$$R = \frac{\left|S_{man} \cap S_{sum}\right|}{\left|S_{man}\right|} \tag{2-36}$$

$$P = \frac{\left|S_{man} \cap S_{sum}\right|}{\left|S_{sum}\right|} \tag{2-37}$$

$$F = \frac{2R \cdot P}{R + P} \tag{2-38}$$

R 表示召回率式(2-36)，P 表示精確率式(3-37)、F 代表 F-評估值式(2-38)。其中 S_{man} 代表參考摘要，S_{sum} 為自動摘要結果，$\left|S_{man} \cap S_{sum}\right|$ 表示自動摘要中與參考摘要之間的相同文句的個數(以文句為單位)，$\left|S_{man}\right|$ 表示參考摘要的文句數，$\left|S_{sum}\right|$ 表示自動摘要的文句數。

此方法較適用於以文句摘錄為基礎的文字文件摘要，因語音文件存在著文句邊界定義的問題，無法正確地定義出文句與文句之間的斷點，同時，語音文件中文句的邊界可能與參考摘要中文句邊界不一致，因此一般不常使用於語音文件摘要的評估。針對語音文件的評估，某些文獻[Hori et al. 2004]會以自動摘要結果的文句，與參考摘要中的文句之間詞的重疊(Overlap)

比例大於某種程度(例如 50%的重疊，重疊指文句與文句間相同內容的部分)，來決定自動摘要的文句是否與參考摘要文句相同。

III. ROUGE 評估方法：由亞洲微軟研究院林欽佑博士[Lin et al. 2003; Lin 2003]所提出的召回率導向(Recall-Oriented)自動摘要評估方法，可直接由網路[Lin 2003]下載套裝軟體(Package)使用。作法是計算自動摘要與參考摘要之間的重疊單位元次數佔參考摘要長度(單位元總個數)的比例，單位元可為 *N*-連詞(*N*-gram)、詞序列(Word Sequences)或詞對(Word Pairs)等。以 *N*-連詞單位元(如二連詞、三連詞等)的評估為例，其計算公式如下：

$$ROUGE-N=\frac{\sum_{S\in S_H}\sum_{g_n\in S}C_m(g_n)}{\sum_{S\in S_H}\sum_{g_n\in S}C(g_n)} \tag{2-39}$$

其中 S_H 為參考摘要集合，S 為個別的參考摘要內容，g_n 表示單位元，$C_m(g_n)$ 表示自動摘要與參考摘要的重疊(Overlapping)單位元個數，$C(g_n)$ 表示參考摘要中的單位元個數。此評估方式採用單位元比對的方式，不會有文句邊界定義的問題，適用於文字文件摘要及語音文件摘要，而且適合多份參考摘要的評估。目前許多文獻[Hirohata et al. 2005; Maskey et al. 2005]都使用此種摘要評估方式。關於 ROUGE 套件詳細的介紹可參考 [Lin et al. 2003; Lin 2003]。

IV. 餘弦評估(Cosine Measure)：計算自動摘要結果與參考摘要的相關程度作為摘要正確率[Saggion et al. 2002; Ho 2003; Steinberger et al. 2004]。假設 $A_D(m\%)$ 代表對某篇文件 D 之自動摘要 $m\%$ 摘要比例的結果、$E_{h,D}(m\%)$ 代表第 h 個人對文件 D 以摘錄方式選取 $m\%$ 摘要比例的結果、$R_{h,D}$ 代表第 h 個人對文件 D 重寫摘要結果，則對所有 $|\{D\}|$ 篇文件的自動摘要正確率定義為

[Saggion et al. 2002; Ho 2003]：

$$ACC(m\%)=\frac{1}{\left|\{D\}\right|}\sum_{D\in\{D\}}\frac{\sum_{h=1}^{H}SIM\left(A_D(m\%),E_{h,D}(m\%)\right)+SIM\left(A_D(m\%),R_{h,D}\right)}{H\times 2} \quad (2\text{-}40)$$

其中，H 為參考摘要的份數，$SIM(\cdot,\cdot)$ 為評估自動摘要與參考摘要的餘弦值估測，公式如下：

$$SIM\left(A_D(m\%),E_{h,D}(m\%)\right)=\frac{\vec{V}_{A_D(m\%)}\cdot\vec{V}_{E_{h,D}(m\%)}}{\left|\vec{V}_{A_D(m\%)}\right|\left|\vec{V}_{E_{h,D}(m\%)}\right|} \quad (2\text{-}41)$$

上式中 $\vec{V}_{A_D(m\%)}$ 與 $\vec{V}_{E_{h,D}(m\%)}$ 分別為自動摘要結果 $A_D(m\%)$ 與摘錄式參考摘要結果 $E_{h,D}(m\%)$ 的向量表示式。

除了上述的評估方法外，其他還有許多應用於摘要評估上的方法。例如[Steinberger et al. 2004]提出二種利用奇異值分解來評估自動摘要與參考摘要之間的相似度。其中，自動摘要與參考摘要內容皆被表示為矩陣的形式，透過奇異值分解成左奇異矩陣、奇異矩陣及右奇異矩陣，然後計算自動摘要所產生的左奇異矩陣與參考摘要所產生的左奇異矩陣之間的相似度，作為摘要的評估結果。由於評估方法很多，適用的對象、目的也不盡相同，因此可以選擇適用或是較為常用的方法來進行摘要評估，不然可以選用一個以上的評估方法，來驗證摘要方法的結果，並增加評估結果的可靠性。本論文採用 ROUGE-2 評估(使用雙連詞為單位元)與餘弦評估二種評估方法。

2.4 摘要的呈現方式

過去，文字為資訊內容呈現的主要方式，文字文件的摘要內容多半以文字來表示；然而，多媒體影音技術的發展，使得資訊的呈現日趨多樣化，文字不再是唯一的方式。摘要的呈現通常可分為文字及語音訊號二種方式[Furui et al. 2004]。

文字的呈現方式較易於使用者瀏覽，同時方便查詢檢索與其他方面的應用，例如搜尋引擎及文件分群等。但是，對於語音文件摘要內容而言，文字無法確切的表達語者當時說話的語氣、情緒及情感，摘要中也可能包含有辨識錯誤的字或詞，可能會造成使用者對內容中語意或文句的誤解；不過根據觀察，除非語音內容的雜訊過多而造成極差的辨識結果，否則使用者大略可依內容的上下文，以及近似音的字或詞來推論出正確的文字。若摘要以語音訊號來呈現，可利於存檔保存，並且方便使用者以語音播放的方式了解摘要內容；此外，許多語音及聲學特徵亦會被保留及呈現出來；然而，語音內容的順暢及悅耳程度將會是主要考量。

對於文字文件而言，摘要以文字來呈現一個是很好的方式，因為文件內容沒有包含辨識錯誤，摘要內容的可讀性與流暢度佳。若以語音訊號來呈現，無論是透過語音合成的技術，或是由相關的語音檔案經訊號切割與串接的技術，所形成的語音檔案，都可能會因這些訊號處理的相關技術的不夠成熟，而讓聽者感到不舒服或是不順暢，甚至會無法清楚地了解摘要內容。

語音文件的摘要內容同樣地也可以文字或語音訊號來呈現。但是，如同上面所說，語音辨識所產生的自動轉寫文件包含了辨識錯誤、語助詞及無意義的內容，可能無法透過文字的作很好表達，而若以語音檔案來呈現，又有不同的實作方式及優缺點。以下，我們針對[Furui et al. 2004]中所討論的以語音訊號來呈現語音文件摘要的方法，作簡單的概述。

- 語音合成：將語音文件的摘要內容，透過語音合成的技術產生語音檔案。缺點是合成的語音內容不夠人性化(機械式說話)，亦無情緒的高低起伏，同時會將轉寫文件中存在的辨識錯誤，合成語音訊號，帶來了錯誤的訊息。
- 語音訊號切割與串接：針對語音文件中所摘錄出的摘要內容，直接由原語音檔案中切割出相對應的訊號內容所串接而成的檔案。優點是保留了原語音檔案中的資訊，語音內容較自然與人性化，且不會包含有辨識錯誤的部份；缺點是所切割出來的語音片斷(單位元)間的串接部分，會產

生不流暢的情形。[Furui et al. 2004]探討了不同切割單位元(例如文句或詞等)的串接結果、訊號品質及摘要內容的理解程度。

作法主要分為二步驟，訊號單位元的選取及訊號單位元的串接。首先決定要選取的單位元，文句、片詞、詞或字，然後訊號單位元的選取可由摘要中單位元所對應的語音訊號檔案的位置切割出來，或者是由語音檔案中切割出與摘要內容相同且訊號較清楚的單位元。最後，串接所選取的訊號，同時考慮單位元與單位元串接的流暢度，以及文句跟文句間的連續性，例如詞與詞之間的連接處，可以訊號作平滑化(Smoothing)或插入極短暫停頓(Short pause)，而對於文句與文句之間的連接處，可插入較長的停頓時間，在文句的起始與結束時，也可用訊號的漸強或漸弱來表示。

上述摘要的語音訊號呈現方式的詳細探討與分析可以參考[Furui et al. 2004]。對於相關的語音訊號處理與語音合成進一步的研究及探討，可參照[Huang et al. 2001]。