I443530

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I443530 B**

(45)公告日： 中華民國 103 (2014) 年 07 月 01 日

(21)申請案號：098134734 (22)申請日： 中華民國 98 (2009) 年 10 月 14 日

(51)Int. Cl.： *G06F17/21 (2006.01)*

(71)申請人：國立交通大學 (中華民國) NATIONAL CHIAO TUNG UNIVERSITY (TW)
新竹市大學路 1001 號

(72)發明人：李嘉晃 LEE, CHA HOANG (TW) ；廖贊瑋 LIAO, ZAN WEI (TW)

(74)代理人：黃于真；李國光

(56)參考文獻：

TW	I289261	TW	200612264A
US	5721897	US	2002/0078090A1
US	2005/0216443A1	US	2009/0063473A1

審查人員：吳偉賢

申請專利範圍項數：18 項 圖式數：15 共 29 頁

(54)名稱

文件處理系統及方法

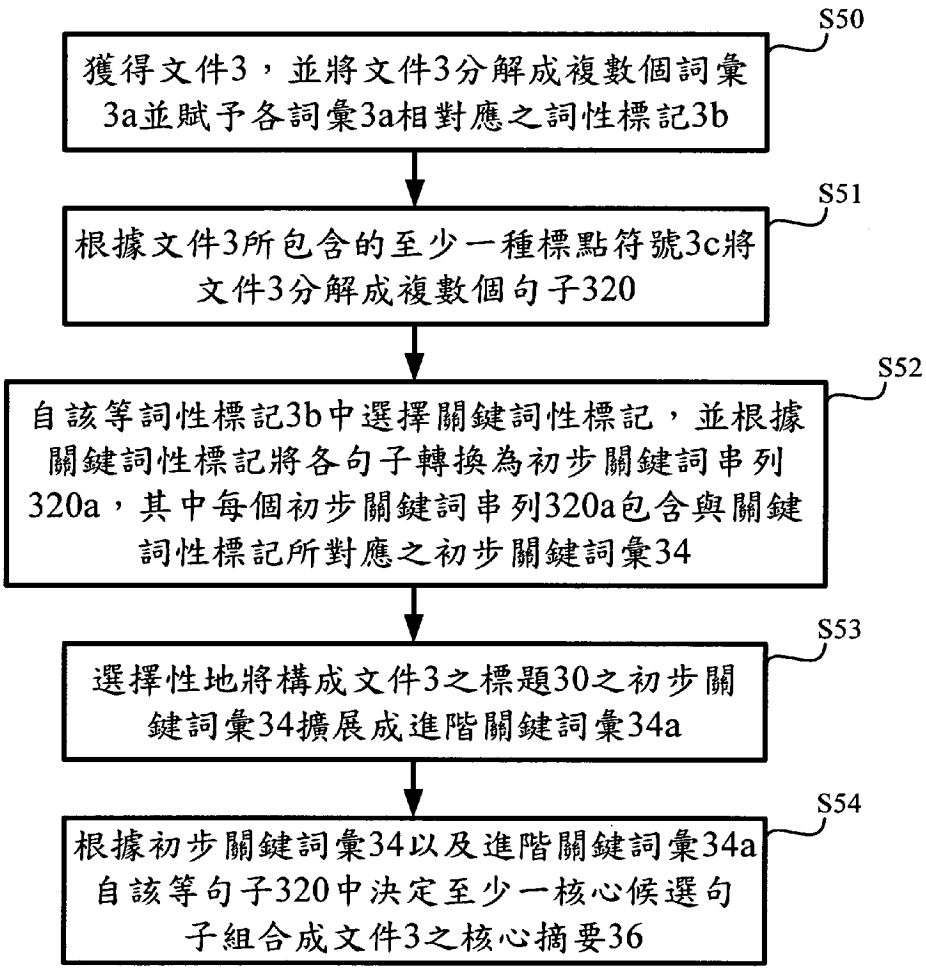
DOCUMENT PROCESSING SYSTEM AND METHOD

(57)摘要

本發明提供一種文件處理系統及方法。根據本發明之文件處理方法包含下列步驟：首先，獲得一文件，並將文件分解成多個詞彙並賦予各詞彙相對應之詞性標記。隨後，將該文件分解成多個句子。接著，自該等詞性標記中選擇關鍵詞性標記，並根據關鍵詞性標記將各句子轉換為初步關鍵詞串列。每個初步關鍵詞串列包含與該等關鍵詞性標記所對應之初步關鍵詞彙。進一步，選擇性地將構成該文件之代表句子之初步關鍵詞彙擴展成進階關鍵詞彙。最後，根據初步關鍵詞彙以及進階關鍵詞彙自多個句子中決定至少一核心候選句子組合成該文件之核心摘要。

The invention provides a document processing system and method. The document processing method includes the following steps: obtaining a document and dividing the document into a plurality of terms and giving a syntactic index to each of the terms. Afterward, divides the document into a plurality of sentence. Then, selects key syntactic indexes from the syntactic indexes, and transforms each of the sentences to be a primary key terms sequence, which includes primary key terms corresponding to the key syntactic indexes, according to the key syntactic indexes. Furthermore, selectively expands the primary key terms to include secondary key terms. Finally, determines at least one candidate sentence from the sentences to construct a core abstract of the document according to the primary key terms and the secondary key terms.

S50~S54・・・流程步驟



圖三

I443530

專利案號：098134734

發明摘要

102年07月12日 修正替換頁

申請日： 98.10.14

IPC分類： G06F 17/21 (2006.01)

公告本

【發明摘要】

【中文發明名稱】 文件處理系統及方法

【英文發明名稱】 DOCUMENT PROCESSING SYSTEM AND METHOD

【中文】

本發明提供一種文件處理系統及方法。根據本發明之文件處理方法包含下列步驟：首先，獲得一文件，並將文件分解成多個詞彙並賦予各詞彙相對應之詞性標記。隨後，將該文件分解成多個句子。接著，自該等詞性標記中選擇關鍵詞性標記，並根據關鍵詞性標記將各句子轉換爲初步關鍵詞串列。每個初步關鍵詞串列包含與該等關鍵詞性標記所對應之初步關鍵詞彙。進一步，選擇性地將構成該文件之代表句子之初步關鍵詞彙擴展成進階關鍵詞彙。最後，根據初步關鍵詞彙以及進階關鍵詞彙自多個句子中決定至少一核心候選句子組合成該文件之核心摘要。

【英文】

The invention provides a document processing system and method. The document processing method includes the following steps: obtaining a document and dividing the document into a plurality of terms and giving a syntactic index to each of the terms. Afterward, divides the document into a plurality of sentence.Then, selects key syntactic indexes from the syntactic indexes, and transforms each of the sentences to be a primary key terms sequence, which includes primary key terms corresponding to the key syntactic indexes, according to the key syntactic indexes. Furthermore, selectively expands the primary key terms to include secondary key terms. Finally, determines at least one candidate sentence from the sentences to construct a core abstract of the document according to the primary key terms and the secondary key terms.

102年07月12日 修正替換頁

【指定代表圖】 第（ 三 ）圖。

【代表圖之符號簡單說明】

S50~S54：流程步驟

【特徵化學式】

102年07月12日 修正替換頁

發明專利說明書

【發明說明書】

【中文發明名稱】 文件處理系統及方法

【英文發明名稱】 DOCUMENT PROCESSING SYSTEM AND METHOD

【技術領域】

【0001】 本發明關於一種文件處理系統及方法，並且特別地，本發明關於一種數位文件自動摘要系統及方法。

【先前技術】

【0002】 由於數位科技的進步以及網際網路的快速發展，在網路上公開的資訊量迅速攀升，人們可以輕易地透過網路瀏覽或查詢各種即時資訊，不再僅限於透過書籍、電視或談話獲取資料，省卻了許多詢問或查找資料的困擾。

【0003】 然而，由於網路上的資料量過於龐大，使用者有時候必須花費更多精神才能找到所需要的文件或內容，某種程度上反而增加了找到合適資料的困難度。此外，過多的資料容易讓使用者不知不覺間花了很多時間在瀏覽重複的或無實質幫助的數位內容。

【0004】 以網路新聞爲例，目前許多入口網站均有提供即時新聞，讓使用者可在短時間內瀏覽各媒體機構所發布的新聞文件。這些新聞文件的數量少則數百，多則上千，並且每數分鐘便更新一次，很容易對使用者造成資訊過載的問題。特別是，這些新聞文件有許多是各媒體機構針對同一新聞事件所作的報導，其內容可能大同小異，甚至一模一樣。因此，如何快速處理數位文件或資訊，讓使用者在最短時間內獲取所需或有用的資訊有其必要性。

【0005】 有鑒於搜尋引擎的廣泛使用，若能搭配自動摘要之方法，讓使用者在搜尋的過程中，可以根據摘要的內容去判斷是否要讀取這篇文章。如此一來，

可以加快使用者搜尋的速度，更可減少使用者瀏覽文件的時間。

【發明內容】

【0006】 因此，本發明之一範疇在於提供一種文件處理方法，以解決先前技術的問題。

【0007】 根據一具體實施例，該方法包含下列步驟：首先，獲得一文件，並將該文件分解成複數個詞彙並賦予各詞彙相對應之詞性標記。隨後，根據該文件所包含的標點符號將該文件分解成複數個句子。

【0008】 接著，自該等詞性標記中選擇複數個關鍵詞性標記，並根據該等關鍵詞性標記將各句子轉換爲初步關鍵詞串列。每個初步關鍵詞串列包含與該等關鍵詞性標記所對應之初步關鍵詞彙。進一步，選擇性地將構成該文件之代表句子之初步關鍵詞彙擴展成進階關鍵詞彙。最後，根據初步關鍵詞彙以及進階關鍵詞彙自複數個句子中決定至少一核心候選句子組合成該文件之核心摘要。

【0009】 本發明之另一範疇在於提供一種文件處理系統，以解決先前技術的問題。

【0010】 根據一具體實施例，該文件處理系統包含接收模組、第一分解模組、第二分解模組、轉換模組、處理模組以及核心摘要模組。

【0011】 接收模組用以獲得文件，而第一分解模組連接接收模組，用以將文件分解成複數個詞彙並賦予各詞彙相對應之詞性標記。此外，第二分解模組同樣連接接收模組，用以根據文件所包含的至少一種標點符號將文件分解成複數個句子。

【0012】 進一步，轉換模組連接第一分解模組以及第二分解模組，用以自該等詞性標記中選擇複數個關鍵詞性標記，並根據關鍵詞性標記將各句子轉換爲初步關鍵詞串列。其中，每個初步關鍵詞串列包含與關鍵詞性標記所對應之

至少一初步關鍵詞彙。

【0013】 此外，處理模組連接轉換模組，用以選擇性地將構成該文件之代表句子之至少一初步關鍵詞彙擴展成進階關鍵詞彙。另外，核心摘要模組連接處理模組，用以根據初步關鍵詞彙以及進階關鍵詞彙自複數個句子中決定至少一核心候選句子組合成該文件之核心摘要。

【0014】 綜上所述，根據本發明之文件處理系統及方法可有效率地產生文件的摘要。特別地，根據本發明之文件處理系統及方法所產生的摘要可精確地包含文件中的重點，讓使用者可根據摘要的內容判斷是否要讀取該文件，節省使用者瀏覽文件的時間。

【0015】 關於本發明之優點與精神可以藉由以下的發明詳述及所附圖式得到進一步的瞭解。

【圖式簡單說明】

【0016】 圖一顯示一文件之範例。

圖二繪示根據本發明之一具體實施例的文件處理系統之功能方塊圖。

圖三繪示根據本發明之一具體實施例的文件處理方法之流程圖。

圖四顯示圖一之文件經第一分解模組處理後之範例。

圖五顯示圖一之文件經第二分解模組處理後之範例。

圖六A顯示轉換模組轉換文件之標題的範例。

圖六B顯示轉換模組轉換文件之文章段落的範例。

圖七A顯示處理模組將初步關鍵詞彙擴展成進階關鍵詞彙的範例。

圖七B繪示處理模組根據初步關鍵詞彙的關鍵詞性標記進行擴展之流程圖。

圖八A顯示本發明之核心摘要模組所產生的核心摘要。

圖八B繪示核心摘要模組產生核心摘要之流程圖。

圖九A顯示本發明之輔助摘要模組所產生的輔助摘要。

圖九B繪示輔助摘要模組產生輔助摘要之流程圖。

圖十A顯示本發明之顯示摘要模組所產生的顯示摘要。

圖十B繪示顯示摘要模組產生顯示摘要之流程圖。

【實施方式】

【0017】 本發明提供一種文件處理系統及方法，用以自動產生文件的摘要。關於本發明之文件處理系統及方法的若干具體實施例係揭露如下。

【0018】 請一併參見圖一、圖二以及圖三，圖一顯示一文件之範例；圖二繪示根據本發明之一具體實施例的文件處理系統之功能方塊圖；圖三則繪示根據本發明之一具體實施例的文件處理方法之流程圖。

【0019】 如圖二所示，根據本發明之文件處理系統1包含接收模組10、第一分解模組11、第二分解模組12、轉換模組13、處理模組14、核心摘要模組15、輔助摘要模組16以及顯示摘要模組17。

【0020】 接收模組10可自網路上接收、由儲存裝置中獲得或由使用者輸入如圖一所示之文件3。於一實施例中，文件3可包含標題30以及多個文章段落32；於另一實施例中文件3亦可以僅包含多個文章段落32。

【0021】 請一併參見圖四，圖四顯示圖一之文件經第一分解模組11處理後之範例。於本具體實施例中，第一分解模組11連接接收模組10，用以將文件3分解成複數個詞彙3a並賦予各詞彙3a相對應之詞性標記3b（圖三步驟S50）。舉例來說，詞彙「社會」、「人員」、「辦法」、「路線」…等之詞性標記爲「普通名詞(Na)」；「提出」、「放」、「辭去」…等之詞性標記爲「動作及物動詞(VC)」。

【0022】 請一併參見圖五，圖五顯示圖一之文件經第二分解模組12處理後之範例。於本具體實施例中，第二分解模組12連接接收模組10，用以根據文件3所包

含的至少一種標點符號3c將文件3分解成複數個句子320（圖三步驟S51）。於本具體實施例中，第二分解模組12根據逗號、分號、句號、問號、驚嘆號等標點符號3c分解文件3。然而，於實務中，第二分解模組12也可根據其他合適的標點符號進行斷句。

【0023】 轉換模組13連接第一分解模組11以及第二分解模組12，用以自該等詞性標記中選擇複數個關鍵詞性標記。請參見下表一，其列示本具體實施例中所選擇的關鍵詞性標記之一實施例。請注意，於實務中，吾人也可視情況選擇其他詞性標記作爲關鍵詞性標記。請參見圖六A，圖六A顯示轉換模組13轉換文件之標題的範例。如圖所示，轉換模組13根據表一所示之關鍵詞性標記將文件3之標題30轉成爲多個初步關鍵詞彙34，而刪除了其他與表一所示之關鍵詞性標記無關之詞彙(圖三步驟S52)。進一步，前述步驟應用了壓縮的概念，即一段文字中僅僅保留與關鍵詞性標記有關之詞彙而刪除了與關鍵詞性標記無關之詞彙；關鍵詞性標記越多表示壓縮越少，關鍵詞性標記越少表示壓縮越多。

【0024】 請再參見圖六B，圖六B顯示轉換模組13轉換文件之文章段落的範例。如圖所示，轉換模組13還根據表一所示之關鍵詞性標記將文件3之文章段落32所包含的各句子320（詳圖五)轉換爲初步關鍵詞串列320a(詳圖六B)，且每個初步關鍵詞串列320a包含與關鍵詞性標記所對應之初步關鍵詞彙34。此轉換可將原本句子中一些停用字(Stop Word)去除，僅保留重要的關鍵詞部分。

【0025】 表一

關鍵詞性標記	詞性標記意義
Na	普通名詞
Nb	專有名稱
Nc	地名
VA	動作不及物動詞
VB	動作類及物動詞
VC	動作及物動詞
VE	動作句賓動詞
VH	狀態不及物動詞
VHC	狀態使動動詞
VK	狀態句賓動詞

【0026】 基於中文之結構特性與寫作手法，作者於標題中常使用縮詞或代名詞等，以期用較精簡的字數規納文件要點。爲了讓標題32的初步關鍵詞彙34更完整，本發明之處理模組14被用來將初步關鍵詞彙34擴展成進階關鍵詞彙。

【0027】 請參見圖七A，圖七A顯示處理模組將初步關鍵詞彙擴展成進階關鍵詞彙的範例。處理模組14連接轉換模組13，用以選擇性地將構成文件3之代表句子(本具體實施例係指標題30)之初步關鍵詞彙34，擴展成進階關鍵詞彙34a(步驟S53)，此擴展主要是尋找文章中與代表句子之初步關鍵詞彙34相似之其他詞彙，以及依據一同義詞資料庫進一步找出同義詞，進而擴展成進

階關鍵詞彙34a。

【0028】 於實務中，當文件中沒有標題時，代表句子可改爲各文章段落之首句及/或末句，或者改爲其他方式所挑選的句子。

【0029】 於實務中，前述步驟圖三之S53可進一步分解爲下列步驟。請參見圖七B，圖七B繪示處理模組14根據初步關鍵詞彙的關鍵詞性標記進行擴展之流程圖。如圖所示，處理模組14先判斷初步關鍵詞彙之關鍵詞性標記(步驟S530)，當初步關鍵詞彙34的關鍵詞性標記爲表一所示之動詞詞性標記時，處理模組14擷取初步關鍵詞彙(如，詞彙「入黨」)的最後一個字(即，「黨」)作爲基礎字(步驟S531)。接著，以基礎字比對圖六B所示之該等關鍵詞串列320a中之初步關鍵詞彙34，找出相符合者成爲候選詞彙(例如有包含「黨」之初步關鍵詞彙，即，「民進黨」)(步驟S532)。並且，計算候選詞彙於文件3中的出現頻率，當其出現頻率大於預設值時，保留此候選詞彙。

【0030】 另外，當初步關鍵詞彙34的關鍵詞性標記爲表一所示之名詞詞性標記時，處理模組14擷取初步關鍵詞彙(如，詞彙「謝籲」)的第一個字作爲第一基礎字(即，「謝」)，且擷取該初步關鍵詞彙的最後一個字作爲第二基礎字(即，「籲」)(步驟S533)。接著，分別以第一基礎字以及第二基礎字比對圖六B所示之該等關鍵詞串列320a中之初步關鍵詞彙34，找出相符合者成爲候選詞彙(例如有包含「謝」之初步關鍵詞彙，即，「謝長廷」)(步驟S534)。

【0031】 如圖二所示，處理模組14還連接同義詞資料庫2，並以前述之候選詞彙查詢同義詞資料庫2，找出候選詞彙之至少一同義詞(步驟S535)。例如，以「改革」查詢同義詞資料庫2將會查到「改進」、「改良」、「改善」、「改造」、「更新」、「維新」…等同義詞。處理模組14進一步將這些同義詞與文章段落32的關鍵詞串列320a作比對，用以判斷這些同義詞是否出現於

初步關鍵詞串列320a中，若是，則以此同義詞作爲初步關鍵詞彙34之進階關鍵詞彙34a（步驟S536）。

【0032】 請一併參見圖八A以及圖八B，圖八A顯示根據本發明之核心摘要模組15所產生的核心摘要36，圖八B則繪示核心摘要模組產生核心摘要之流程圖。核心摘要模組15連接處理模組14，用以根據初步關鍵詞彙34以及進階關鍵詞彙34a產生文件3之核心摘要。於本具體實施例中，核心摘要模組15根據初步關鍵詞彙34以及進階關鍵詞彙34a自複數個句子320中決定至少一核心候選句子組合成文件3之核心摘要36（步驟S54）。

【0033】 進一步，如圖八B所示，於核心摘要36的產生流程中核心摘要模組15先賦予初步關鍵詞彙34第一權重，並賦予進階關鍵詞彙34a第二權重，且第一權重不等於第二權重(步驟S540)。例如，第一權重：第二權重＝2：1。接著，核心摘要模組15依照初步關鍵詞彙34的第一權重計算包含初步關鍵詞彙34之句子之分數(步驟S541)。於實務中，由於每個句子都已被轉換成關鍵詞串列，因此可將每個關鍵詞串列所含的初步關鍵詞彙數量乘上第一權重作爲句子的分數(例如，某個句子的關鍵詞串列包含5個關鍵詞彙，其中有3個爲初步關鍵詞彙，則該句子的分數即3 × 2 = 6)；當該句子之分數大於一預設值時即將該句子選入核心摘要。於另一實施例中，除了計算上述句子之分數外，接著並進一步計算初步關鍵詞彙34在各句子所佔之比例(步驟S542)（例如，某個句子的關鍵詞串列包含5個關鍵詞彙，其中有1個爲初步關鍵詞彙，則初步關鍵詞彙佔該句子的比例即爲1/5)。若比例大於一定值，如1/4，則將該句子視爲含較少資訊量之句子，將其剔除。最後，將未被剔除的句子組成如圖八所示之核心摘要36（步驟S543)。

【0034】 也就是說，於實務中核心摘要的產生流程可以僅包含前述步驟S541，並且核心摘要模組根據各句子的分數高低，挑選分數較高的句子(例如分數最高

的前10句)組成核心摘要。

【0035】 請一併參見圖九A以及圖九B，圖九A顯示根據本發明之輔助摘要模組16所產生的輔助摘要37，圖九B則繪示輔助摘要模組產生輔助摘要之流程圖。輔助摘要模組16連接轉換模組13以及處理模組14，輔助摘要模組16以初步關鍵詞彙34以及進階關鍵詞彙34a所形成之關鍵詞彙集合搭配初步關鍵詞串列320形成矩陣(步驟S55)。對該矩陣進行奇異值分解(Singular Value Decomposition)，以獲得各初步關鍵詞串列所對應之句子之權重(步驟S56)，並根據權重選擇該等句子中之至少一成爲輔助候選句子，並將輔助候選句子組合成文件之輔助摘要37 (步驟S57)。請注意，於實務中，輔助摘要模組16以及輔助摘要37並非必要的，本發明之方法可視情況決定是否加入輔助摘要37。

【0036】 關於前述之矩陣的形成以及奇異值分解等方法請參見Y. Gong等人於2001年所發表之標題爲「應用相關性衡量以及潛在語意分析進行關鍵詞摘要」的論文(Y. Gong, and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in Proc. ACM SIGIR Conference on R&D in Information Retrieval, pp. 19-25, 2001)，該論文係以全文引用方式納入本文中。

【0037】 請一併參見圖十A以及圖十B，圖十A顯示根據本發明之顯示摘要模組17所產生的顯示摘要38，圖十B則繪示顯示摘要模組產生顯示摘要之流程圖。顯示摘要模組17連接核心摘要模組15 (於另一實施例中亦可以同時連接核心摘要模組15以及輔助摘要模組16)，摘要模組17主要是從核心摘要中選出所需之摘要句數，或從核心摘要與輔助摘要中選出所需之摘要句數。所需之摘要句數可以預先設定，例如5~10句之間。摘要句數也可以由摘要模組17接收文件3之壓縮比例(例如，15%)，並根據該壓縮比例計算摘要句數(步

驟S58)。於實務中，壓縮比例可由系統預設、由文件3的提供者設定或由欲觀看摘要的使用者設定。

【0038】 如果爲了增加摘要句數，於另一實施中，顯示摘要模組17判斷核心摘要所包含的句子數量是否小於摘要句數(步驟S59)。若否，則從核心摘要36中挑選摘要句數之句子作爲文件之顯示摘要38（步驟S60)。反之，若核心摘要36所包含的句子數量小於摘要句數，則從輔助摘要37中挑選不與核心摘要36重覆之輔助候選句子併入核心摘要36中，成爲文件之顯示摘要38（步驟S61)。請注意，於實務中，步驟S59至步驟S61並非必要的。

【0039】 最後，爲了增加摘要的可讀性，顯示摘要模組17還可對上述整合過後的顯示摘要做句尾之處理。舉例來說，顯示摘要模組17進一步判斷顯示摘要中的最末句子結尾的標點符號是否爲句號、分號、問號或驚嘆號(步驟S62)。若否，則自該文件中，挑選最末句子的下一句子加入顯示摘要中，並重新進行判斷(步驟S63)。

【0040】 於實務中，本發明之文件處理系統1還包含顯示模組，連接該顯示摘要模組17，用以接收並顯示該顯示摘要。此外，顯示模組也可顯示操作介面，供使用者輸入欲進行自動摘要的文章、壓縮比例或其他參數。

【0041】 綜上所述，根據本發明之文件處理系統及方法可有效率地產生文件的摘要。特別地，根據本發明之文件處理系統及方法所產生的摘要可精確地包含文件中的重點，讓使用者可根據摘要的內容判斷是否要讀取該文件，節省使用者瀏覽文件的時間。

【0042】 此外，本發明之文件處理系統及方法可與搜尋引擎結合，當使用者透過搜尋引擎查詢一關鍵字時，搜尋引擎可先透過本發明之文件處理方法對結果網頁的內容進行摘要，並在顯示搜尋結果時一併顯示摘要給使用者。

【0043】 藉由以上較佳具體實施例之詳述，係希望能更加清楚描述本發明之特徵與精神，而並非以上述所揭露的較佳具體實施例來對本發明之範疇加以限制。相反地，其目的是希望能涵蓋各種改變及具相等性的安排於本發明所欲申請之專利範圍的範疇內。因此，本發明所申請之專利範圍的範疇應該根據上述的說明作最寬廣的解釋，以致使其涵蓋所有可能的改變以及具相等性的安排。

【符號說明】

【0044】 1：文件處理系統

10：接收模組

11：第一分解模組

12：第二分解模組

13：轉換模組

14：處理模組

15：核心摘要模組

16：輔助摘要模組

17：顯示摘要模組

2：同義詞資料庫

3：文件

3a：詞彙

3b：詞性標記

3c：標點符號

30：標題

32：文章段落

320：句子

320a：初步關鍵詞串列

34：初步關鍵詞彙

34a：進階關鍵詞彙

36：核心摘要

37：輔助摘要

38：顯示摘要

S50~S63、S530~S536、S540~S543：流程步驟

【主張利用生物材料】

【0045】

102年07月12日 修正替換頁

申請專利範圍

【發明申請專利範圍】

【第1項】 一種文件處理方法，包含下列步驟：

(a) 獲得一文件，並將該文件分解成複數個詞彙並賦予各該等詞彙相對應之一詞性標記；

(b) 根據該文件所包含的至少一種標點符號將該文件分解成複數個句子；

(c) 自該等詞性標記中選擇複數個關鍵詞性標記，並根據該等關鍵詞性標記將各該等句子轉換爲一初步關鍵詞串列，其中每個初步關鍵詞串列包含與該等關鍵詞性標記所對應之至少一初步關鍵詞彙；

(d) 選擇性地將構成該文件之一代表句子之該至少一初步關鍵詞彙擴展成一進階關鍵詞彙；

(e) 根據該初步關鍵詞彙以及該進階關鍵詞彙自該複數個句子中決定至少一核心候選句子組合成該文件之一核心摘要；

(f) 以該初步關鍵詞彙以及該進階關鍵詞彙所形成之一關鍵詞彙集合搭配步驟(c)之該等初步關鍵詞串列形成一矩陣；

(g) 對該矩陣進行奇異值分解(Singular Value Decomposition)，以獲得各該初步關鍵詞串列所對應之該句子之一權重；

(h) 根據該權重選擇該等句子中之至少一成爲一輔助候選句子，並將該至少一輔助候選句子組合成該文件之一輔助摘要；

(i) 接收一壓縮比例，根據該壓縮比例計算一摘要句數；

(j) 判斷該核心摘要所包含的句子數量是否小於該摘要句數；

(k) 若步驟(j)之判斷爲否，則從該核心摘要中挑選該摘要句數之句子作爲該文件之一顯示摘要；以及

(l) 若步驟(j)之判斷爲是，則從該輔助摘要中挑選不與核心摘要重覆之該輔助候選句子併入該核心摘要中，成爲該文件之該顯示摘要。

【第2項】 如申請專利範圍第1項所述之文件處理方法，其中該詞性標記選自由一普通名詞詞性標記、一專有名稱詞性標記、一地名詞性標記、一動作不及物動詞詞性標記、一動作類及物動詞詞性標記、一動作及物動詞詞性標記、一動作句賓動詞詞性標記、一狀態不及物動詞詞性標記、一狀態使動動詞詞性標記以及一狀態句賓動詞詞性標記所組成之群組。

【第3項】 如申請專利範圍第1項所述之文件處理方法，其中該代表句子係該文件之一標題。

【第4項】 如申請專利範圍第1項所述之文件處理方法，其中該文件包含至少一段落，且該代表句子係該至少一段落之一首句及/或末句。

【第5項】 如申請專利範圍第1項所述之文件處理方法，其中步驟(d)進一步包含下列步驟：

(d1) 判斷該初步關鍵詞彙之該關鍵詞性標記；

(d2) 當該關鍵詞性標記爲一動詞詞性標記時，擷取該初步關鍵詞彙的最後一個字作爲一基礎字；

(d3) 以該基礎字比對該等初步關鍵詞串列中之該等初步關鍵詞彙，找出相符合者成爲一候選詞彙；

(d4) 計算該候選詞彙於該文件中的出現頻率，當其出現頻率大於一預設值時，以該候選詞彙查詢一同義詞資料庫，找出該候選詞彙之至少一同義詞；以及

(d5) 判斷該至少一同義詞是否出現於該等初步關鍵詞串列，若是，則以該至少一同義詞作爲該初步關鍵詞彙之該進階關鍵詞彙。

【第6項】 如申請專利範圍第5項所述之文件處理方法，進一步包含下列步驟：

(d2') 當該初步關鍵詞彙之該關鍵詞性標記爲一名詞詞性標記時，擷取

該初步關鍵詞彙的第一個字作爲一第一基礎字，且擷取該初步關鍵詞彙的最後一個字作爲一第二基礎字；以及

(d3’) 分別以該第一基礎字以及該第二基礎字比對該等初步關鍵詞串列中之該等初步關鍵詞彙，找出相符合者成爲該候選詞彙。

【第7項】 如申請專利範圍第1項所述之文件處理方法，其中步驟(e)進一步包含下列步驟：

(e1) 分別賦予該初步關鍵詞彙以及該進階關鍵詞彙一第一權重以及一第二權重；

(e2) 依照該第一權重計算包含該初步關鍵詞彙之該初步關鍵詞串列之分數，再根據該初步關鍵詞串列之分數計算包含該初步關鍵詞串列之句子之分數，並根據所計算之各句子之該分數挑選出若干句子；以及

(e3) 將該些被挑選出來的句子中之至少一組成該核心摘要。

【第8項】 如申請專利範圍第7項所述之文件處理方法，進一步包含下列步驟：

(e21) 計算該初步關鍵詞彙在各該句子所佔之一比例；以及

(e31) 自該些被挑選出來的句子中篩選出該初步關鍵詞彙所佔比例較低者成爲該核心候選句子組合成該核心摘要。

【第9項】 如申請專利範圍第1項所述之文件處理方法，進一步包含下列步驟：

(m) 判斷該顯示摘要中的最末句子結尾的標點符號是否爲一句號、一分號、一問號或一驚嘆號；以及

(n) 若步驟(m)之判斷爲否，則自該文件中，挑選該最末句子的下一句子加入該顯示摘要中，並重新進行步驟(m)。

【第10項】 一種文件處理系統，包含：

一接收模組，用以獲得一文件；

一第一分解模組，連接該接收模組，用以將該文件分解成複數個詞彙並賦予各該等詞彙相對應之一詞性標記；

一第二分解模組，連接該接收模組，用以根據該文件所包含的至少一種標點符號將該文件分解成複數個句子；

一轉換模組，連接該第一分解模組以及該第二分解模組，用以自該等詞性標記中選擇複數個關鍵詞性標記，並根據該等關鍵詞性標記將各該等句子轉換爲一初步關鍵詞串列，其中每個初步關鍵詞串列包含與該等關鍵詞性標記所對應之至少一初步關鍵詞彙；

一處理模組，連接該轉換模組，用以選擇性地將構成該文件之一代表句子之該至少一初步關鍵詞彙擴展成一進階關鍵詞彙；

一核心摘要模組，連接該處理模組，用以根據該初步關鍵詞彙以及該進階關鍵詞彙自該複數個句子中決定至少一核心候選句子組合成該文件之一核心摘要；

一輔助摘要模組，連接該轉換模組以及該處理模組，該輔助摘要模組以該初步關鍵詞彙以及該進階關鍵詞彙所形成之一關鍵詞彙集合搭配該等初步關鍵詞串列形成一矩陣，對該矩陣進行奇異值分解(Singular Value Decomposition)，以獲得各該初步關鍵詞串列所對應之該句子之一權重，並根據該權重選擇該等句子中之至少一成爲一輔助候選句子，並將該至少一輔助候選句子組合成該文件之一輔助摘要；以及

一顯示摘要模組，連接該核心摘要模組以及該輔助摘要模組，用以接收一壓縮比例，根據該壓縮比例計算一摘要句數，判斷該核心摘要所包含的句子數量是否小於該摘要句數，若否，則從該核心摘要中挑選該摘要句數之句子作爲該文件之一顯示摘要，反之，則從該輔助摘要中挑選不與核心摘要重覆之該輔助候選句子併入該核心摘要中，成爲該文件之該顯示摘要。

【第11項】 如申請專利範圍第10項所述之文件處理系統，其中該詞性標記選自由一普通名詞詞性標記、一專有名稱詞性標記、一地名詞性標記、一動作不

及物動詞詞性標記、一動作類及物動詞詞性標記、一動作及物動詞詞性標記、一動作句賓動詞詞性標記、一狀態不及物動詞詞性標記、一狀態使動動詞詞性標記以及一狀態句賓動詞詞性標記所組成之群組。

【第12項】 如申請專利範圍第10項所述之文件處理系統，其中該代表句子係該文件之一標題。

【第13項】 如申請專利範圍第10項所述之文件處理系統，其中該文件包含至少一段落，且該代表句子係該至少一段落之一首句及/或末句。

【第14項】 如申請專利範圍第10項所述之文件處理系統，其中該處理模組判斷該初步關鍵詞彙之該關鍵詞性標記，當該關鍵詞性標記爲一動詞詞性標記時，擷取該初步關鍵詞彙的最後一個字作爲一基礎字；以該基礎字比對該等初步關鍵詞串列中之該等初步關鍵詞彙，找出相符合者成爲一候選詞彙；計算該候選詞彙於該文件中的出現頻率，當其出現頻率大於一預設值時，以該候選詞彙查詢一同義詞資料庫，找出該候選詞彙之至少一同義詞；並且判斷該至少一同義詞是否出現於該等初步關鍵詞串列，若是，則以該至少一同義詞作爲該初步關鍵詞彙之該進階關鍵詞彙。

【第15項】 如申請專利範圍第14項所述之文件處理系統，其中當該初步關鍵詞彙之該關鍵詞性標記爲一名詞詞性標記時，該處理模組擷取該初步關鍵詞彙的第一個字作爲一第一基礎字，且擷取該初步關鍵詞彙的最後一個字作爲一第二基礎字，並分別以該第一基礎字以及該第二基礎字比對該等初步關鍵詞串列中之該等初步關鍵詞彙，找出相符合者成爲該候選詞彙。

【第16項】 如申請專利範圍第10項所述之文件處理系統，其中該核心摘要模組分別賦予該初步關鍵詞彙以及該進階關鍵詞彙一第一權重以及一第二權重；依照該第一權重計算包含該初步關鍵詞彙之該初步關鍵詞串列之分數，再根據該初步關鍵詞串列之分數計算包含該初步關鍵詞串列之句子之分數，並根據所計算之各句子之該分數挑選出若干句子；並且將該些被挑

選出來的句子中之至少一組成該核心摘要。

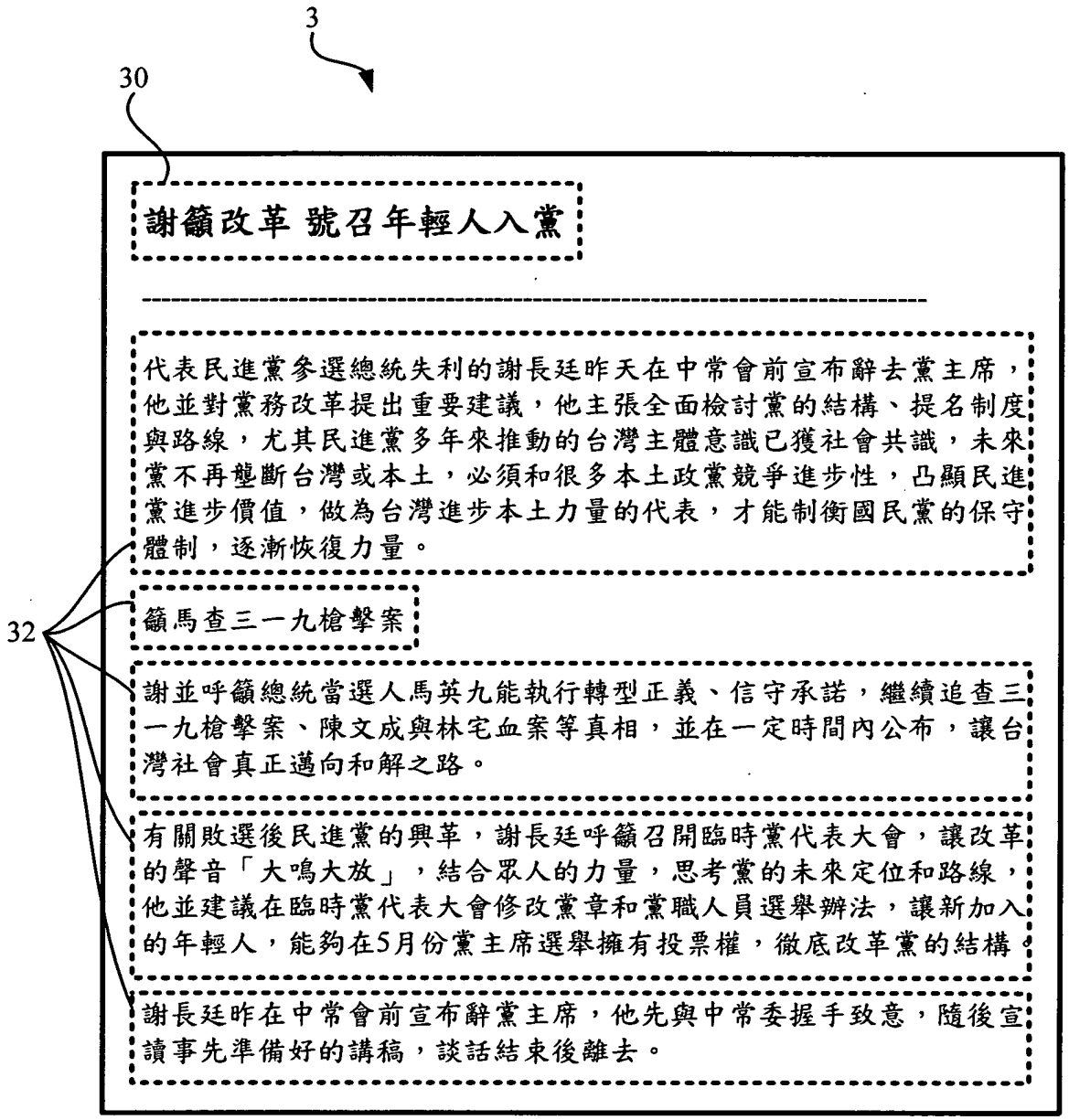
【第17項】 如申請專利範圍第16項所述之文件處理系統，其中該核心摘要模組還計算該初步關鍵詞彙在各該句子所佔之一比例，並自該些被挑選出來的句子中篩選出該初步關鍵詞彙所佔比例較低者成爲該核心候選句子組合成該核心摘要。

【第18項】 如申請專利範圍第10項所述之文件處理系統，其中該顯示摘要模組進一步判斷該顯示摘要中的最末句子結尾的標點符號是否爲一句號、一分號、一問號或一驚嘆號，若否，則自該文件中，挑選該最末句子的下一句子加入該顯示摘要中，並重新進行判斷。

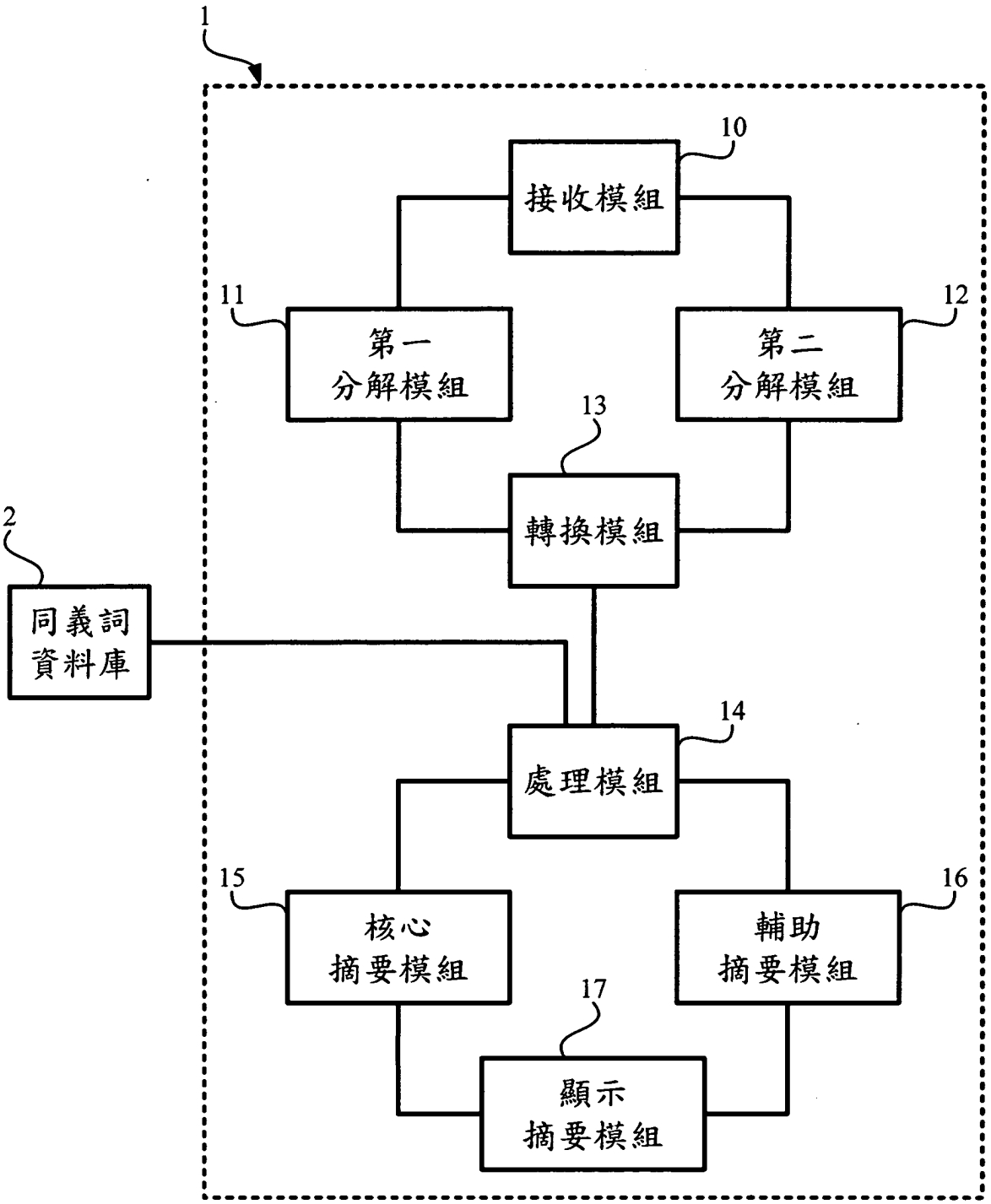
102年07月12日 修正替換頁

【發明圖式】

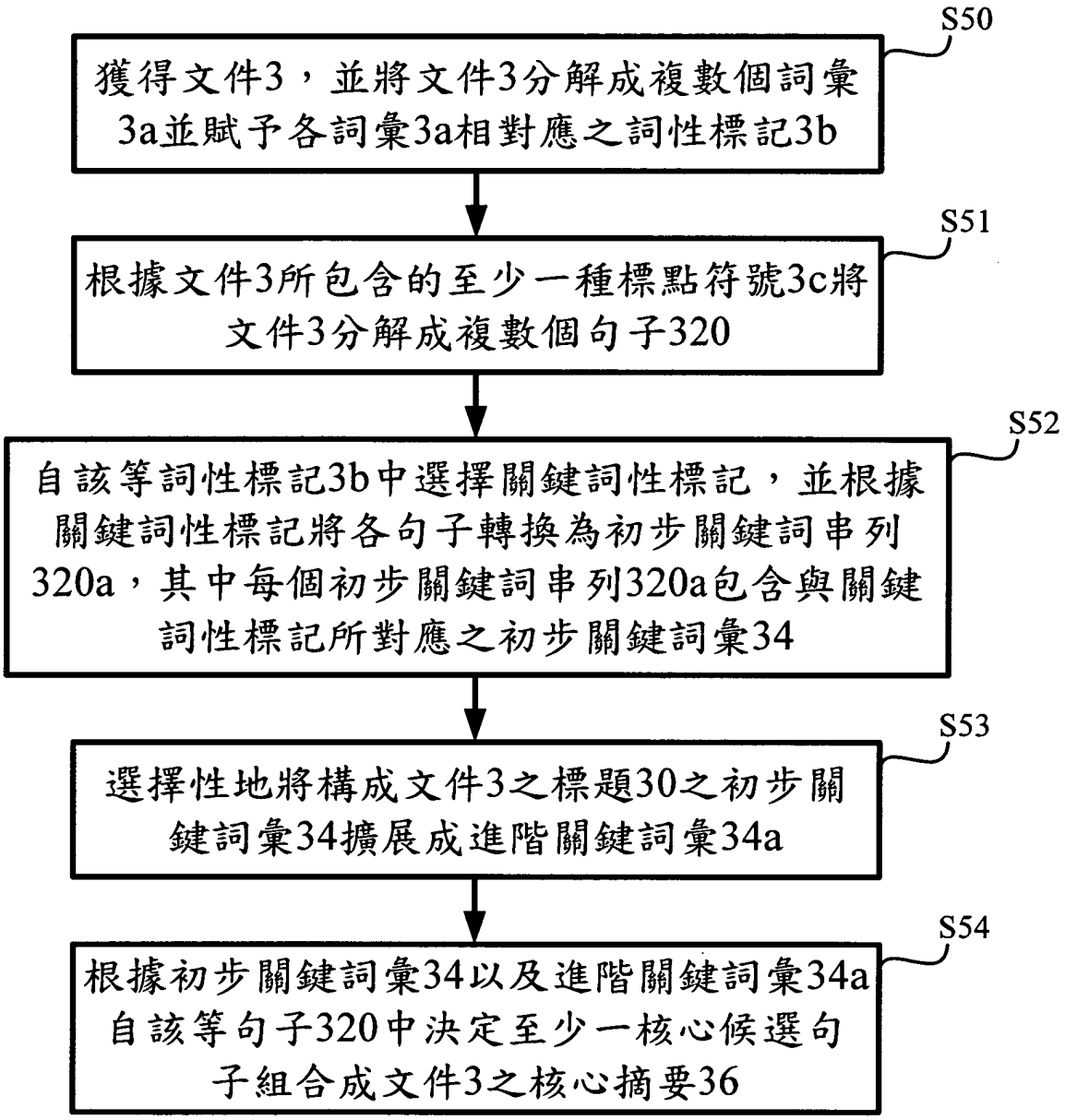
圖式



圖一


1
10
接收模組
11
第一
分解模組
12
第二
分解模組
13
轉換模組
2
同義詞
資料庫
14
處理模組
15
核心
摘要模組
16
輔助
摘要模組
17
顯示
摘要模組

圖二


S50
獲得文件3，並將文件3分解成複數個詞彙3a並賦予各詞彙3a相對應之詞性標記3b
S51
根據文件3所包含的至少一種標點符號3c將文件3分解成複數個句子320
S52
自該等詞性標記3b中選擇關鍵詞性標記，並根據關鍵詞性標記將各句子轉換為初步關鍵詞串列320a，其中每個初步關鍵詞串列320a包含與關鍵詞性標記所對應之初步關鍵詞彙34
S53
選擇性地將構成文件3之標題30之初步關鍵詞彙34擴展成進階關鍵詞彙34a
S54
根據初步關鍵詞彙34以及進階關鍵詞彙34a自該等句子320中決定至少一核心候選句子組合成文件3之核心摘要36

圖三

3a 3b

```
謝謝(Nb) 改革(VC) 號召(VF) 年輕人(Na) 入黨(VH)
代表(VK) 民進黨(Nb) 參選(VC) 總統(Na) 失利(VH) 的(DE) 謝長廷(Nb) 昨天(Ndd) 在(P) 中常會(Nc) 前(Ncd) 宣布(VE) 辭去(VC) 黨主席(Na) ，(COMMACATEGORY)
他(Nh) 並(D) 對(P) 黨務(Na) 改革(VC) 提出(VC) 重要(VH) 建議(VE) ，(COMMACATEGORY)
他(Nh) 主張(VE) 全面(A) 檢討黨(Nb) 的(DE) 結構(Na) 、(PAUSECATEGORY) 提名(VC) 制度(Na) 與(Caa) 路線(Na) ，(COMMACATEGORY)
尤其(D) 民進黨(Nb) 多(Neqa) 年(Nf) 來(D) 推動(VC) 的(DE) 台灣(Nc) 主體(Na) 意識(Na) 已(D) 獲(VJ) 社會(Na) 共識(Na) ，(COMMACATEGORY)
未來黨(Na) 不再(D) 壟斷(VC) 台灣(Nc) 或(Caa) 本土(Nc) ，(COMMACATEGORY)
必須(D) 和(P) 很多(Neqa) 本土(Nc) 政黨(Na) 競爭(Na) 進步性(Na) 。(COMMACATEGORY)
凸顯(VJ) 民進黨(Nb) 進步(VH) 價值(Na) ，(COMMACATEGORY)
做為(VG) 台灣(Nc) 進步(VH) 本土(Nc) 力量(Na) 的(DE) 代表(Na) ，(COMMACATEGORY)
才能(Na) 制衡(VA) 國民黨(Nb) 的(DE) 保守(VH) 體制(Na) 。(PERIODCATEGORY)
逐漸(D) 恢復(VHC) 力量(Na) 。(PERIODCATEGORY)
顧馬(Na) 查(VE) 三一九(Neu) 槍擊案(Na)
謝(Nb) 並(D) 呼籲(VE) 總統(Na) 當選人(Na) 馬英九(Nb) 能(D) 執行(VC) 轉型(VH) 正義(Na) 、(PAUSECATEGORY) 信守(VJ) 承諾(Na) ，(COMMACATEGORY)
組織(VP) 追查(VC) 三一九(Neu) 槍擊案(Na) 、(PAUSECATEGORY) 陳文成(Nb) 與(Caa) 林宅(Nc) 血案(Na) 等(Cab) 真相(Na) ，(COMMACATEGORY)
並(Cbb) 在(P) 一定(A) 時間(Na) 內(Ncd) 公布(VE) ，(COMMACATEGORY)
釋(VL) 台灣(Nc) 社會(Na) 真正(D) 邁向(VCL) 和解(VA) 之(DE) 路(Na) 。(PERIODCATEGORY)
有關(VJ) 敗選後(VH) 民進黨(Nb) 的(DE) 興革(VC) ，(COMMACATEGORY)
謝長廷(Nb) 呼籲(VE) 召開(VC) 臨時(A) 黨代表(Na) 大會(Na) ，(COMMACATEGORY)
釋(VL) 改革(VC) 的(DE) 聲音(Na) 「(PARENTHESISCATEGORY) 大鳴(Nb) 大(VH) 放(VC) 」(PARENTHESISCATEGORY) ，(COMMACATEGORY)
結合(VHC) 眾人(Na) 的(DE) 力量(Na) ，(COMMACATEGORY)
思考(VE) 黨(Na) 的(DE) 未來(Nd) 定位(Na) 和(Caa) 路線(Na) ，(COMMACATEGORY)
他(Nh) 並(D) 建議(VE) 在(P) 臨時(A) 黨代表(Na) 大會(Na) 修改(VC) 黨章(Na) 和(Caa) 黨職(Na) 人員(Na) 選舉(Na) 辦法(Na) ，(COMMACATEGORY)
釋(VL) 新(VH) 加入(VC) 的(DE) 年輕人(Na) ，(COMMACATEGORY)
能夠(D) 在(P) 五(Neu) 月份(Na) 黨主席(Na) 選舉(Na) 擁有(VJ) 投票權(Na) ，(COMMACATEGORY)
徹底(VH) 改革(VC) 黨(Na) 的(DE) 結構(Na) 。(PERIODCATEGORY)
謝長廷(Nb) 昨(b) 在(P) 中常會(Nc) 前(Ncd) 宣布(VE) 辭(VC) 黨主席(Na) ，(COMMACATEGORY)
他(Nh) 先(D) 與(P) 中常委(Na) 握手(VA) 致意(VB) ，(COMMACATEGORY)
隨後(Ndd) 宣讀(VC) 事先(D) 準備好(VC) 的(DE) 講稿(Na) ，(COMMACATEGORY)
談話(Na) 結束(VHC) 後(Ng) 離去(VA) 。(PERIODCATEGORY)
```

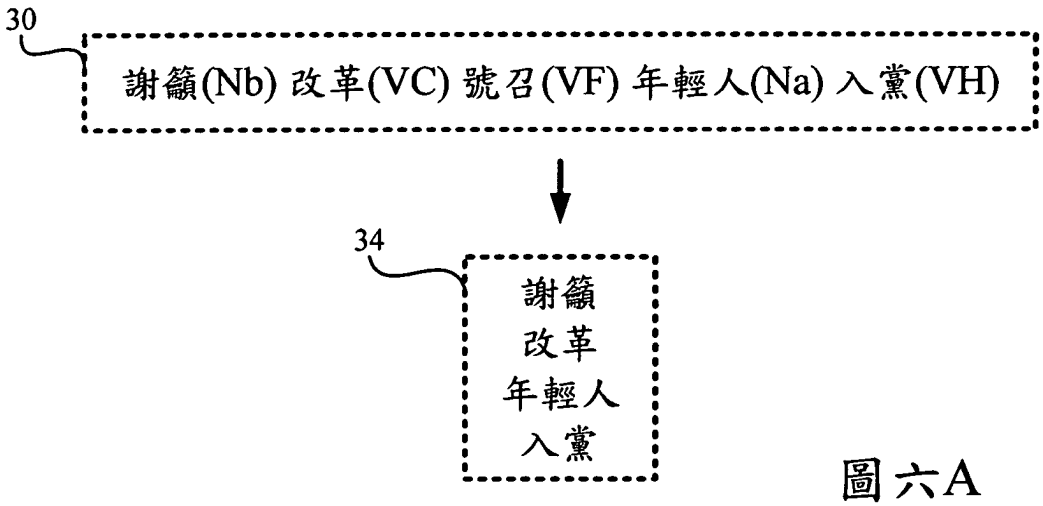
圖四

320

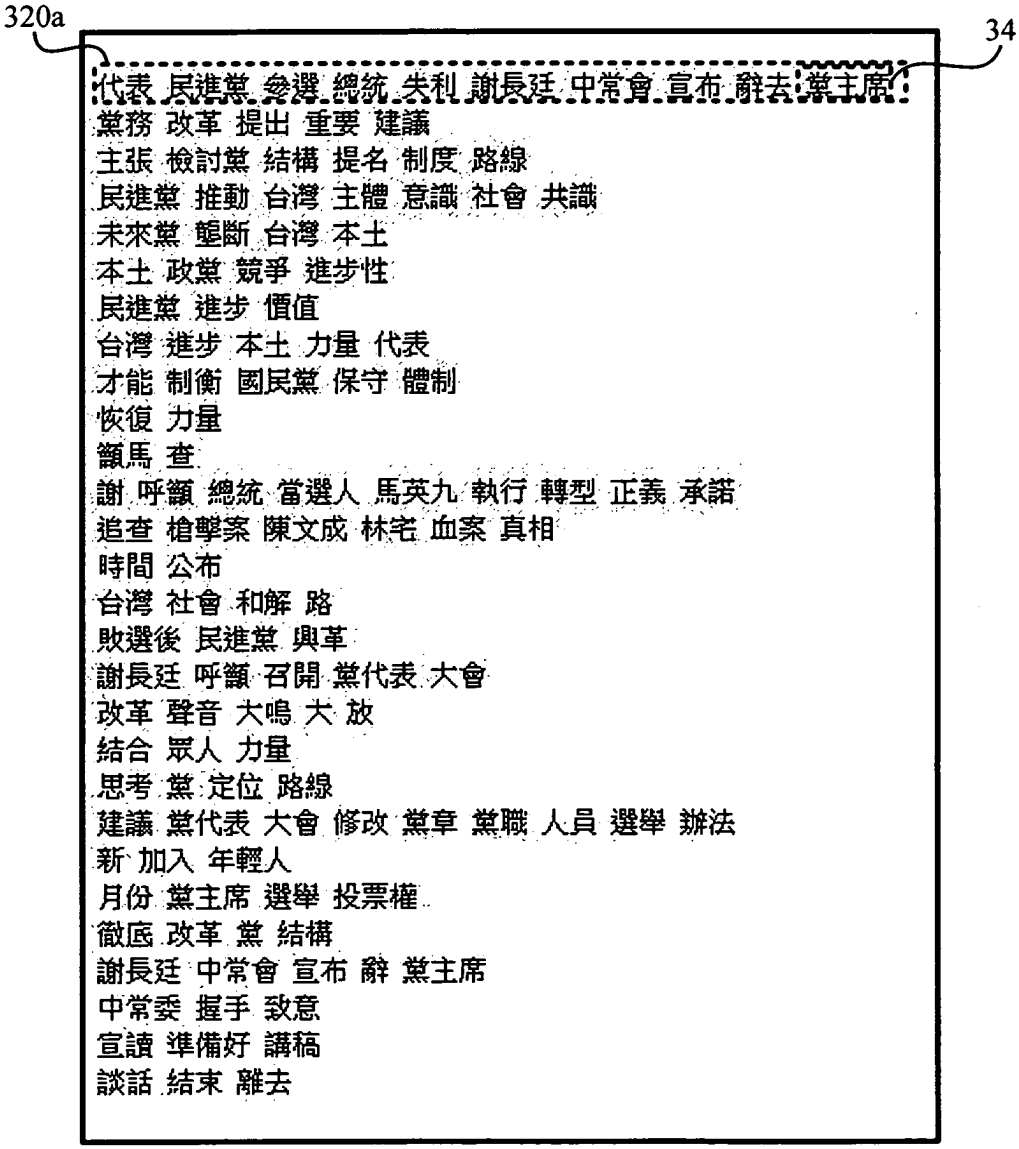
3c

代表民進黨參選總統失利的謝長廷昨天在中常會前宣布辭去黨主席，
他並對黨務改革提出重要建議，
他主張全面檢討黨的結構、提名制度與路線，
尤其民進黨多年來推動的台灣主體意識已獲社會共識，
未來黨不再壟斷台灣或本土，
必須和很多本土政黨競爭進步性，
凸顯民進黨進步價值，
做為台灣進步本土力量的代表，
才能制衡國民黨的保守體制，
逐漸恢復力量。
籲馬查三一九槍擊案
謝並呼籲總統當選人馬英九能執行轉型正義、信守承諾，
繼續追查三一九槍擊案、陳文成與林宅血案等真相，
並在一定時間內公布，
讓台灣社會真正邁向和解之路。
有關敗選後民進黨的興革，
謝長廷呼籲召開臨時黨代表大會，
讓改革的聲音「大鳴大放」，
結合眾人的力量，
思考黨的未來定位和路線，
他並建議在臨時黨代表大會修改黨章和黨職人員選舉辦法，
讓新加入的年輕人，
能夠在5月份黨主席選舉擁有投票權，
徹底改革黨的結構。
謝長廷昨在中常會前宣布辭黨主席，
他先與中常委握手致意，
隨後宣讀事先準備好的講稿，
談話結束後離去。

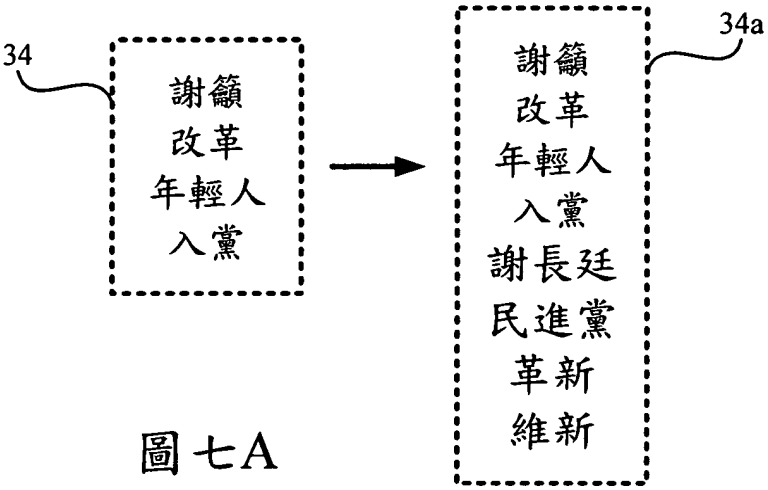
圖五


30
謝籲(Nb) 改革(VC) 號召(VF) 年輕人(Na) 入黨(VH)
34
謝籲
改革
年輕人
入黨

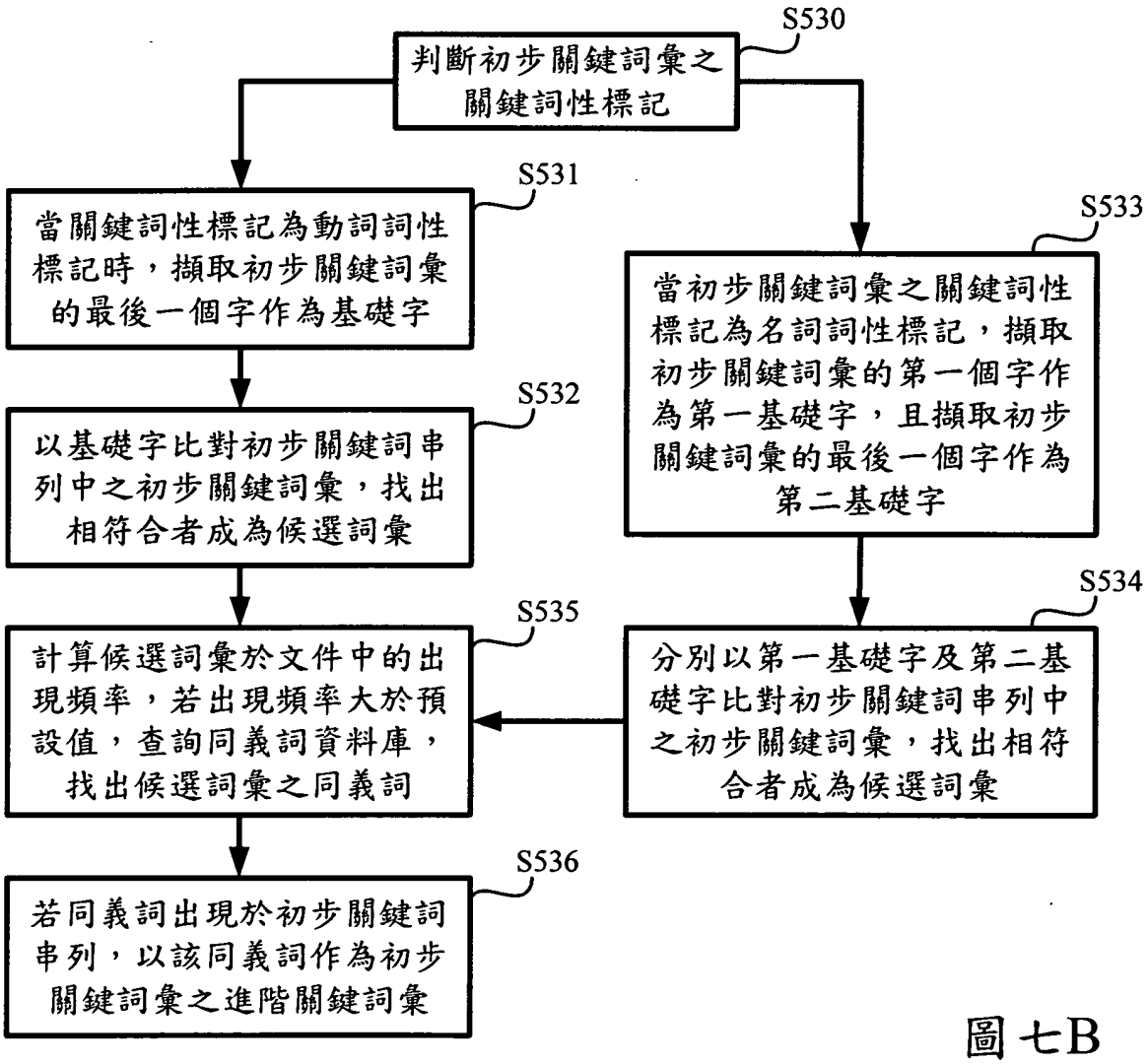
圖六A


320a
34
代表 民進黨 參選 總統 失利 謝長廷 中常會 宣布 辭去 黨主席
黨務 改革 提出 重要 建議
主張 檢討黨 結構 提名 制度 路線
民進黨 推動 台灣 主體 意識 社會 共識
未來黨 壟斷 台灣 本土
本土 政黨 競爭 進步性
民進黨 進步 價值
台灣 進步 本土 力量 代表
才能 制衡 國民黨 保守 體制
恢復 力量
籲馬 查
謝 呼籲 總統 當選人 馬英九 執行 轉型 正義 承諾
追查 槍擊案 陳文成 林宅 血案 真相
時間 公布
台灣 社會 和解 路
敗選後 民進黨 興革
謝長廷 呼籲 召開 黨代表 大會
改革 聲音 大鳴 大 放
結合 眾人 力量
思考 黨 定位 路線
建議 黨代表 大會 修改 黨章 黨職 人員 選舉 辦法
新 加入 年輕人
月份 黨主席 選舉 投票權
徹底 改革 黨 結構
謝長廷 中常會 宣布 辭 黨主席
中常委 握手 致意
宣讀 準備好 講稿
談話 結束 離去

圖六B


34
謝籲
改革
年輕人
入黨
34a
謝籲
改革
年輕人
入黨
謝長廷
民進黨
革新
維新

圖七A

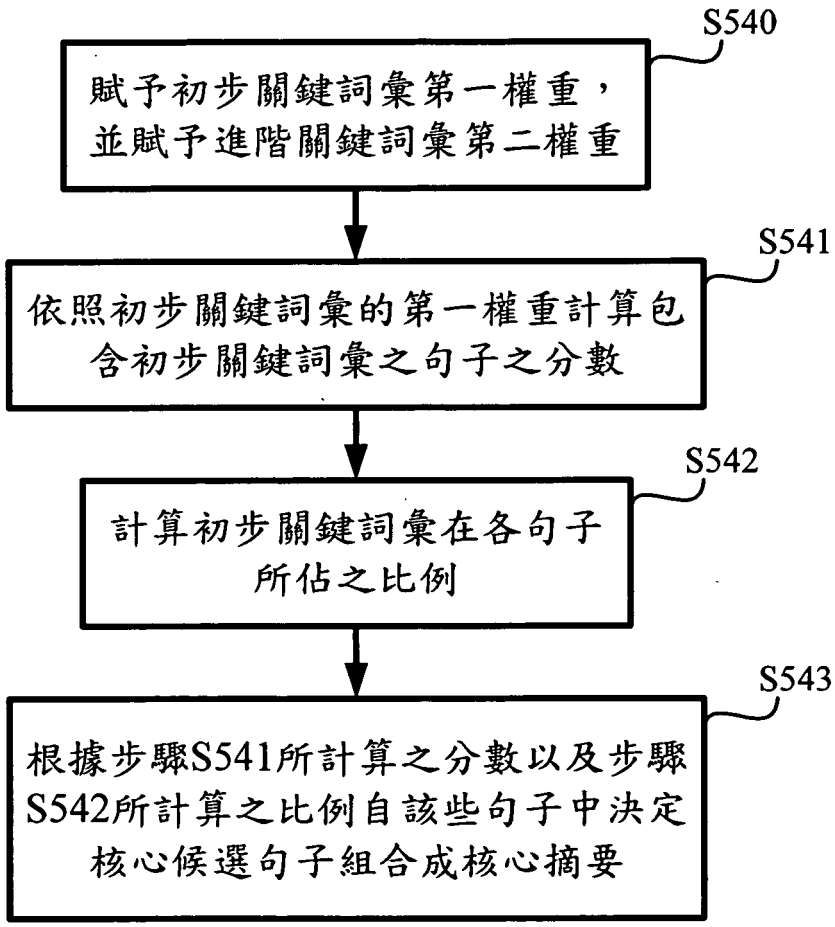

S530
判斷初步關鍵詞彙之關鍵詞性標記
S531
當關鍵詞性標記為動詞詞性標記時，擷取初步關鍵詞彙的最後一個字作為基礎字
S532
以基礎字比對初步關鍵詞串列中之初步關鍵詞彙，找出相符合者成為候選詞彙
S533
當初步關鍵詞彙之關鍵詞性標記為名詞詞性標記，擷取初步關鍵詞彙的第一個字作為第一基礎字，且擷取初步關鍵詞彙的最後一個字作為第二基礎字
S534
分別以第一基礎字及第二基礎字比對初步關鍵詞串列中之初步關鍵詞彙，找出相符合者成為候選詞彙
S535
計算候選詞彙於文件中的出現頻率，若出現頻率大於預設值，查詢同義詞資料庫，找出候選詞彙之同義詞
S536
若同義詞出現於初步關鍵詞串列，以該同義詞作為初步關鍵詞彙之進階關鍵詞彙

圖七B

36

代表民進黨參選總統失利的謝長廷昨天在中常會前宣布辭去黨主席，
他並對黨務改革提出重要建議，
讓改革的聲音「大鳴大放」，
徹底改革黨的結構。

圖八A

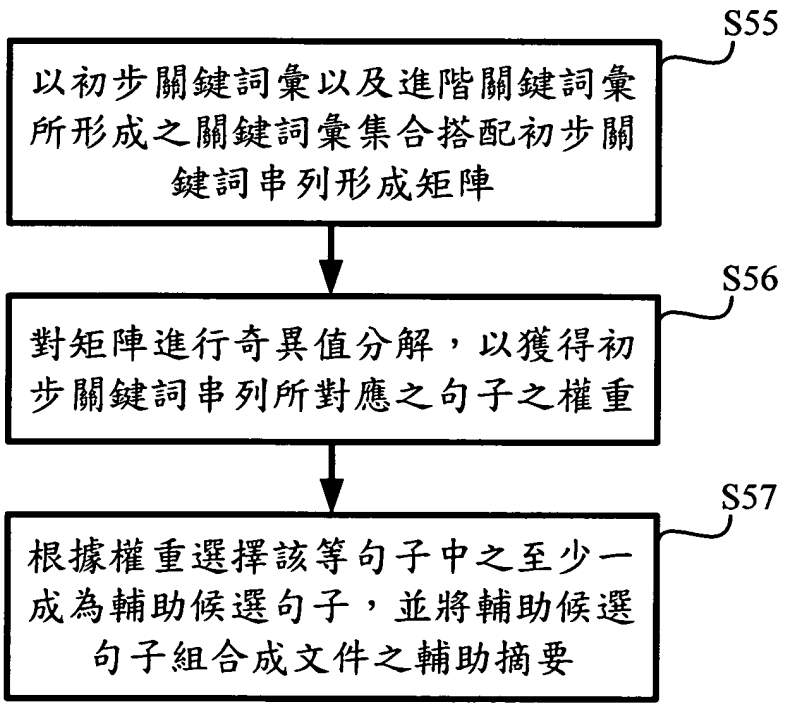


圖八B

37

代表民進黨參選總統失利的謝長廷昨天在中常會前宣布辭去黨主席，
他主張全面檢討黨的結構、提名制度與路線，
凸顯民進黨進步價值，
做為台灣進步本土力量的代表，
才能制衡國民黨的保守體制，
逐漸恢復力量。
籲馬查三一九槍擊案
謝長廷呼籲召開臨時黨代表大會，
讓新加入的年輕人，

圖九A

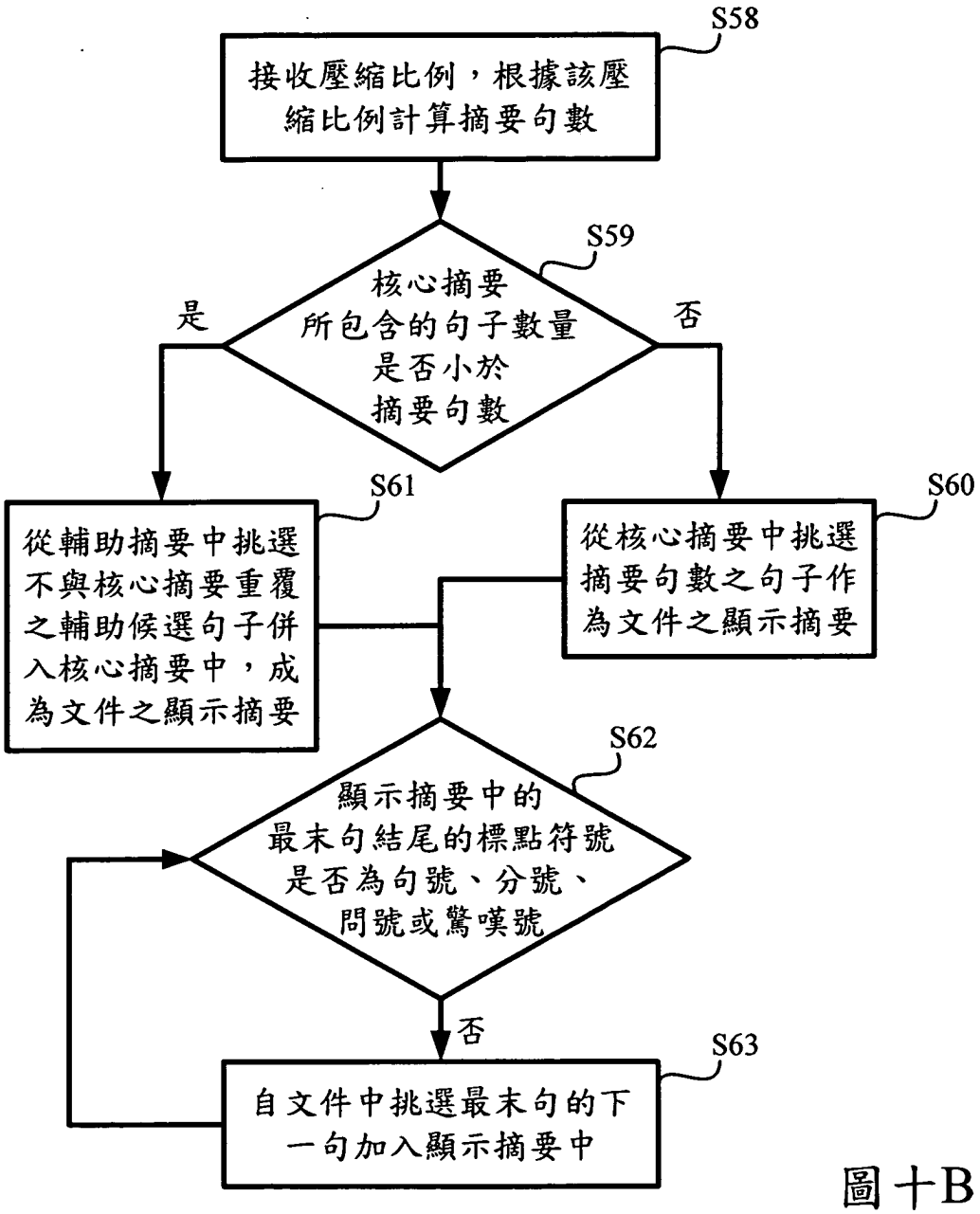


圖九B

38

代表民進黨參選總統失利的謝長廷昨天在中常會前宣布辭去黨主席，
他並對黨務改革提出重要建議，
他主張全面檢討黨的結構、提名制度與路線，
凸顯民進黨進步價值，
讓改革的聲音「大鳴大放」，
徹底改革黨的結構。

圖十A



圖十B