



(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 I729472 B

(45)公告日：中華民國 110 (2021) 年 06 月 01 日

(21)申請案號：108130620

(22)申請日：中華民國 108 (2019) 年 08 月 27 日

(51)Int. Cl. : **G06F17/00 (2019.01)**

(30)優先權：2018/11/26 中國大陸 201811416994.X

(71)申請人：開曼群島商創新先進技術有限公司(開曼群島) ADVANCED NEW TECHNOLOGIES CO., LTD. (KY)

開曼群島

(72)發明人：李懷松 (CN)；潘健民 (CN)；周緒剛 (CN)

(74)代理人：林志剛

(56)參考文獻：

TW 201333735A1 CN 106095736A

Shaosheng Cao et al., "cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information", The Thirty-Second AAAI Conference on Artificial Intelligence, 2018/02/02, pages 5053 to 5061.

審查人員：鄭書季

申請專利範圍項數：18 項 圖式數：7 共 57 頁

(54)名稱

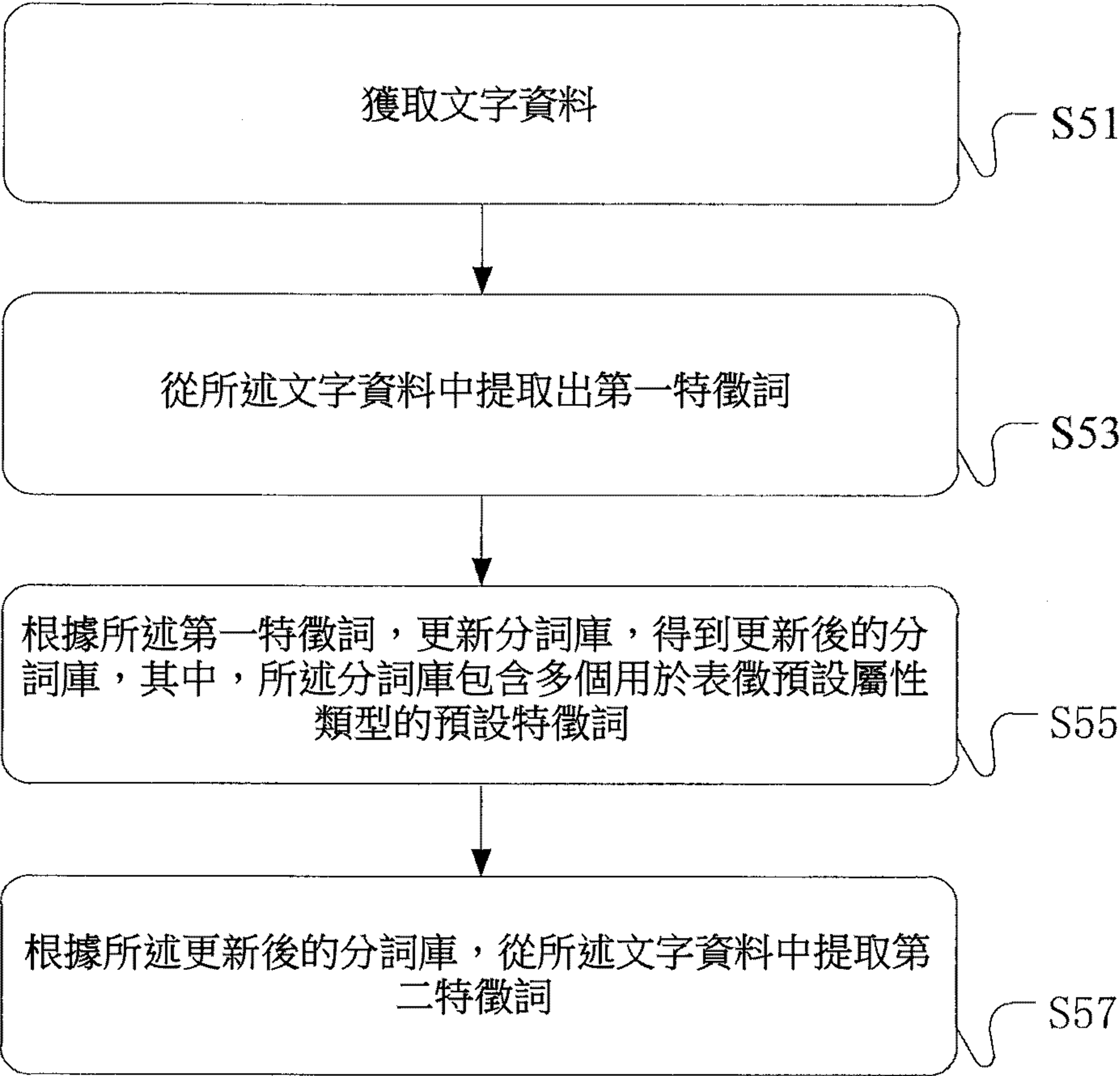
特徵詞的確定方法、裝置和伺服器

(57)摘要

本說明書提供了一種特徵詞的確定方法、裝置和伺服器。其中，該方法包括：獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據更新後的分詞庫和預設特徵詞，從文字資料中提取第二特徵詞。在本說明書實施例中，透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而可以利用更新後的分詞庫和預設特徵詞從文字資料中提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞的過程中，由於分詞錯誤導致的特徵詞提取不準確，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

指定代表圖：

符號簡單說明：  
無



【圖 5】



I729472

公告本

## 【發明摘要】

### 【中文發明名稱】

特徵詞的確定方法、裝置和伺服器

### 【中文】

本說明書提供了一種特徵詞的確定方法、裝置和伺服器。其中，該方法包括：獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據更新後的分詞庫和預設特徵詞，從文字資料中提取第二特徵詞。在本說明書實施例中，透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而可以利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞的過程中，由於分詞錯誤導致的特徵詞提取不準確，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

【指定代表圖】第( 5 )圖。  
【代表圖之符號簡單說明】無

【特徵化學式】無

# 【發明說明書】

## 【中文發明名稱】

特徵詞的確定方法、裝置和伺服器

## 【技術領域】

本說明書屬於網際網路技術領域，尤其涉及一種特徵詞的確定方法、裝置和伺服器。

## 【先前技術】

在網路交易監管中，常常會透過檢索與網路交易相關的資料所攜帶的文字資料中的特徵詞(例如能反映某一屬性類型的網路交易共性的詞組)，來確定該交易的屬性類型，從而可以有針對性地對某種屬性類型的網路交易進行監控管理。

例如，可以透過檢索辨識與網路交易相關的文字資料中是否攜帶有“上分”、“下分”等黑詞(一種表徵違法交易的特徵詞)，來判斷該交易是否屬於違法交易，從而可以及時地發現並處理這類違法交易。

基於上述方法來確定網路交易的屬性類型時，所使用的特徵詞是否準確、涵蓋的範圍是否全面，會對網路交易的屬性類型的判斷是否準確產生較大的影響。而且隨著人們語言習慣、用詞方式的改變，對於同一屬性類型的網路交易，能夠有效表徵該屬性類型交易共性的特徵詞也會發生變化。例如，可能會出現一些新的詞組成為表徵該屬性

類型交易的特徵詞。因此，亟需一種能夠準確地確定出用於表徵網路交易的屬性類型的特徵詞的方法。

### 【發明內容】

本說明書目的在於提供一種特徵詞的確定方法、裝置和伺服器，以避免提取第二特徵詞的過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確、高效地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

本說明書提供的一種特徵詞的確定方法、裝置和伺服器是這樣實現的：

一種特徵詞的確定方法，包括：獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

一種特徵詞的確定裝置，包括：獲取模組，用於獲取文字資料；第一提取模組，用於從所述文字資料中提取出第一特徵詞；更新模組，用於根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；第二提取模組，用於根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

一種伺服器，包括處理器以及用於儲存處理器可執行

指令的儲存器，所述處理器執行所述指令時實現獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

一種電腦可讀儲存媒體，其上儲存有電腦指令，所述指令被執行時實現獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

本說明書提供的一種特徵詞的確定方法、裝置和伺服器，透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

### 【圖式簡單說明】

為了更清楚地說明本說明書實施例或現有技術中的技術方案，下面將對實施例或現有技術描述中所需要使用的圖式作簡單地介紹，顯而易見地，下面描述中的圖式僅僅

是本說明書中記載的一些實施例，對於本領域普通技術人員來講，在不付出創造性勞動性的前提下，還可以根據這些圖式獲得其他的圖式。

圖1是在一個場景示例中，應用本說明書實施例提供的特徵的確定方法的一種實施例的示意圖；

圖2是在一個場景示例中，應用本說明書實施例提供的特徵的確定方法的一種實施例的示意圖；

圖3是在一個場景示例中，應用本說明書實施例提供的特徵的確定方法的一種實施例的示意圖；

圖4是在一個場景示例中，應用本說明書實施例提供的特徵的確定方法的一種實施例的示意圖；

圖5是本說明書實施例提供的特徵詞的確定方法的流程的一種實施例的示意圖；

圖6是本說明書實施例提供的伺服器的結構的一種實施例的示意圖；

圖7是本說明書實施例提供的特徵詞的確定裝置的結構的一種實施例的示意圖。

## 【實施方式】

為了使本技術領域的人員更好地理解本說明書中的技術方案，下面將結合本說明書實施例中的圖式，對本說明書實施例中的技術方案進行清楚、完整地描述，顯然，所描述的實施例僅僅是本說明書一部分實施例，而不是全部的實施例。基於本說明書中的實施例，本領域普通技術人

員在沒有作出創造性勞動前提下所獲得的所有其他實施例，都應當屬於本說明書保護的範圍。

考慮到現有的特徵詞的確定方法通常是先直接對所採集的文字資料進行分詞處理；再透過深度學習演算法得到分詞的向量；繼而透過語義向量距離和 PageRank 等方法計算分詞的向量與已有特徵詞之間的向量距離；最後根據向量距離確定出新的特徵詞。

然而，在網際網路領域，使用的詞組、短語的變化速度相對較快。例如，可能每隔一兩天網際網路上就會出現大量之前沒有出現過的新的詞語。但上述方法具體實施時，並沒有考慮到類似新詞的影響，導致在分詞處理過程容易將新詞組合錯誤地進行拆分，以致後續無法得到新詞真正的向量，難以挖掘出新的特徵詞。

針對上述情況，本說明書考慮可以將新詞的發現與特徵詞的確定相結合。具體的，可以先對所採集的文字資料進行新詞挖掘，提取得到文字資料中的新詞作為第一特徵詞；再根據第一特徵詞更新包含有已有的預設特徵詞的分詞詞庫；透過更新後的分詞詞庫對文字資料進行分詞處理，再結合預設特徵詞，根據分詞結果的向量確定出新的特徵詞，即挖掘出了第二特徵詞。從而避免了現有方法在提取特徵詞的過程中，由於分詞錯誤導致的特徵詞提取不準確，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

本說明書實施方式提供一種可以應用本說明書提供的

特徵詞的確定方法的系統架構中。其中，該系統中具體可以包括伺服器，且該伺服器可以與網路平臺的業務系統對接，進而可以採集得到網路平臺上的文字資料；再從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫；透過所述更新後的分詞庫和所述預設特徵詞，從所述文字資料中提取第二特徵詞，即挖掘出新的特徵詞。

在本實施方式中，所述伺服器可以為一種應用於後臺的，能夠實現資料獲取、資料處理等功能的伺服器。具體的，所述伺服器可以為一個具有資料運算、儲存功能以及網路交互功能的電子設備；也可以為運行於該電子設備中，為資料處理、儲存和網路交互提供支持的軟體程式。在本實施方式中，並不具體限定所述伺服器的數量。所述伺服器具體可以為一個伺服器，也可以為幾個伺服器，或者，由若干伺服器形成的伺服器集群。

在一個場景示例中，可以參閱圖1所示，可以應用本說明書實施例提供的特徵詞的確定方法對在某交易網路平臺上採集得到的相關文字資料進行具體分析處理，從中提取得到能夠用於表徵非法交易的新的特徵詞，以便後續可以根據上述新的特徵詞，及時地發現交易網路平臺上的非法交易，並及時進行監控處理。

在本場景示例中，具體實施時，參閱圖2所示，伺服器可以先搜集該交易網路平臺上最近一周的網路交易資料(例如轉帳資料)中所攜帶的交易附言等作為文字資料。其

中，上述交易網路平臺具體可以是一種以中文為主要語言的購物網站、理財網站、支付平臺等，相應的，所涉及的文字資料通常會包含有大量的中文漢字。

上述文字資料具體可以理解為一種交易網路平臺上包含有漢字、英文字母、拼音符號或數字等資訊的資料。具體的，上述文字資料可以包括交易附言，也可以包括文字標籤等。當然，上述所列舉的交易附言、文字標籤只是為了更好地說明本說明書實施方式。具體實施時，根據具體的應用場景和使用需求，上述文字資料還可以包括有其他類型或形式的資料。對此，本說明書不作限定。

其中，上述交易附言具體可以理解為一種由交易資料發起者自定義編輯的用以描述交易資料的相關特徵的資訊資料。具體的，交易資料發起者在發送交易資料時，可以透過預設的文字資料輸入介面，輸入或設置用以描述交易資料的例如交易目的、交易時間或交易對象等特徵資訊，作為交易附言。例如，用戶A在透過交易網路平臺向用戶B支付XX貨款10元時，在確認發送該筆交易資料前，可以在預設的留言介面中輸入“XX貨款”作為交易附言，以表徵該筆交易資料的交易目的是用於支付XX貨款，進而交易網路平臺會將上述交易附言連同交易資料一同發送給用戶B，以使用戶B可以知曉所收到的該筆交易資料的所對應的交易目的。

上述文字標籤具體可以理解為一種由交易網路平臺採集並根據交易資料發起者的所實施的與該筆交易資料相關

操作，自動生成的描述該筆交易資料的類型、目的或內容等的資料標籤。例如，用戶C在操作介面中，點擊還貸選項，在所展示的下一級介面中輸入“5000”元，發起一筆交易資料。交易網路平臺可以根據用戶C的上述操作確定該筆資料的目的是還貸，因此，可以自動生成“還貸”作為文字標籤添加在該筆交易資料上，以便進行記錄。

伺服器在得到了上述文字資料後，可以先對上述文字資料進行新詞挖掘。具體的，可以透過新詞發現演算法，對上述文字資料進行分析、處理，從中提取得到新詞，作為第一特徵詞。

其中，上述第一特徵詞具體可以理解為一種新詞，即之前未被伺服器發現，也沒有被記錄在伺服器的資料庫或分詞庫中的特定的漢字組合。

需要說明的是，網際網路領域區別於其他技術領域，網際網路領域中的知識資訊更新相對較為快速、頻繁。例如，可能每隔一兩天，網路中可能就會出現大量之前沒有出現過的，能夠表徵一些特定語義的網路詞彙。根據現有的特徵詞的確定方法，往往忽略上述新詞的影響，即沒有及時地利用新詞對分詞庫進行更新，導致後續進行分詞處理時，容易出現錯誤，例如將一些新出現的漢字組合錯誤地進行拆，影響後續語義地確定，造成最終確定的特徵詞存在誤差或者不夠全面。

例如，漢字組合“狗帶”是一種網路新詞。之前的分詞庫如果沒有儲存該詞，後續在分詞處理時，基於原有的分

詞庫很有可能將該詞錯誤拆分為“狗”和“帶”，導致對該詞的語義的確定產生誤差，進而影響後續特徵詞確定。

本說明書正是考慮到網際網路領域中存在的上述特點，在具體確定特徵詞前，先對文字資料進行新詞確定，得到第一特徵詞；進而後續可以結合第一特徵詞，更新分詞庫，再根據更新後的分詞庫，進行更加準確的分詞處理。

具體實施時，伺服器可以從所述文字資料中篩選出字元長度小於預設長度閾值(例如5個字元長度)的字串作為候選字串；分別計算所述候選字串中的各個字串的指標參數；根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為新詞，即第一特徵詞。

具體的，考慮到漢語的組詞習慣，通常一個詞組的字元長度相對較短，因此，可以先根據字元長度，從文字資料中篩選掉明顯不符合組詞習慣的，具有較大機率不能構成詞組的字串，得到候選字串。在本場景示例中，可以將預設長度閾值設為5個字元長度，這樣就可以先過濾掉字元長度大於5個字元的字串，以減少後續的工作量，提高處理效率。需要說明的是，上述所列舉的預設長度閾值只是一種示意性說明。具體實施時，根據具體情況還可以設置其他數值作為上述預設長度閾值。對此，本說明書不作限定。

在本場景示例中，在從所述文字資料中篩選出字元長度小於預設長度閾值的字串後，進一步可以根據已有的資

料庫(例如現有的分詞庫)，從篩選出的字串中過濾掉與已有的資料庫中的已有字串相同的字串，將過濾後得到的字串作為候選字串。這樣可以使得得到的候選字串不會包含已有詞詞組所對應的字串。

上述指標參數具體可以理解為一種用於表徵字串能否構造詞組的特徵參數。通常一個候選字串的指標參數滿足一定的要求，則可以認為該字串可以構成一個詞組，例如構成新詞。

在本場景示例中，上述指標參數具體可以包括以下至少之一：凝固度、資訊熵和頻數等。當然，需要說明的是，上述所列舉的指標參數只是為了更好地說明本說明書實施方式。具體實施時，根據具體情況還可以引入其他類型的特徵參數作為上述指標參數。對此，本說明書不作限定。

其中，上述凝固度具體可以理解為文字資料中字串中所包含的字元結合在一起的機率。通常一個字串的凝固度數值越高，該字串具有越高的機率成為一個詞組。例如，透過計算發現字串“電影院”的凝固度比字串“的電影”的凝固度更高，則可以判斷字串“電影院”相對於“的電影”具有更高的機率、更容易成為一個詞。

具體計算字串的凝固度時，可以先分別計算該字串整體在文字資料中出現的機率、該字串的各種拆分組合在文字資料中出現的機率；再計算該字串整體在文字資料出現的機率與該字串的各種拆分組合在文字資料出現的機率的

比值；最後將數值最小的比值確定為該字串的凝固度。

例如，在計算字串“電影院”的凝固度時，可以先計算字串“電影院”在文字資料中出現的機率為  $p_0 = p(\text{電影院})$ 。字串“電影院”可以拆分出兩種拆分組合，即第一拆分組合：“電”和“影院”，以及第二拆分組合：“電影”和“院”。分別計算兩種拆分組合在文字資料中出現的機率。以確定第一拆分組合的機率為例，可以先分別計算字元“電”的在文字資料中的出現機率和字元“影院”在文字資料中的出現機率；再將上述兩個機率的乘積作為第一拆分組合在文字資料中出現的機率，可以表示為： $p_1 = p(\text{電}) * p(\text{影院})$ 。按照類似的方式，可以計算出第二拆分組合在文字資料中出現的機率  $p_2 = p(\text{電影}) * p(\text{院})$ 。再分別計算  $p_0$  與  $p_1$ 、 $p_2$  的比值為  $p_0/p_1$ 、 $p_0/p_2$ 。比較  $p_0/p_1$ 、 $p_0/p_2$  的數值大小，將數值較小的  $p_0/p_2$  確定為字串“電影院”的凝固度。

上述資訊熵，也可稱為自由度，具體可以理解為字串表徵的某種特定資訊的出現機率。通常當字串所表徵的某種特定資訊的出現機率較高的時候，表明它被傳播得較為廣泛，或者說，被引用的程度相對較高。在說明書中，具體可以透過字串的資訊熵來衡量該字元段的左鄰字集合和右鄰字集合的隨機性，從而可以利用該指標參數反映出該字串作為詞組所攜帶的資訊量 and 使用的頻繁程度。

具體計算資訊熵時，可以先分別搜索文字資料確定字串的左鄰字和右鄰字，再分別統計各個左鄰字的出現機率和各個右鄰字的出現機率；根據各個左鄰字的出現機率按

照預設公式計算左鄰字的資訊熵，根據各個右鄰字的出現機率計算右鄰字的資訊熵。

例如，文字資料為“吃葡萄不吐葡萄皮不吃葡萄倒吐葡萄皮”，在計算字串“葡萄”的資訊熵時，可以先搜索該文字資料找到字串“葡萄”的左鄰字分別為：吃、吐、吃、吐，右鄰字分別為：不、皮、倒、皮。再分別計算各個左鄰字的出現機率為  $p(\text{吃})$  為  $1/2$ ， $p(\text{吐})$  為  $1/2$ ；各個右鄰字的出現機率為  $p(\text{不})$  為  $1/4$ ， $p(\text{倒})$  為  $1/4$ ， $p(\text{皮})$  為  $1/2$ 。再根據各個左鄰字的出現機率按照以下預設公式  $H(x) = -\sum p(x_i) \log(p(x_i))$  計算左鄰字的資訊熵。具體可以表示為  $-(1/2) \cdot \log(1/2) - (1/2) \cdot \log(1/2) \approx 0.693$ 。類似的，根據各個右鄰字的出現機率可以得到右鄰字的資訊熵為  $-(1/2) \cdot \log(1/2) - (1/4) \cdot \log(1/4) - (1/4) \cdot \log(1/4) \approx 1.04$ 。透過比較左鄰字的資訊熵和右鄰字的資訊熵可以發現，對於字串“葡萄”，右鄰字的資訊熵數值更大，因此，該字串所對接的右鄰字相對更豐富一些。

上述頻數具體可以理解為字串在文字資料中出現的次數。通常一個字串的頻數越高，相對的具有較大機率成為一個詞組。具體的，例如統計文字資料，字串“機器”出現了180次，則可以將字串“機器”的頻數確定為180。

在按照上述方式計算得到了各個候選字串的凝固度、資訊熵和頻數等指標參數，進一步可以將各個候選字串的凝固度、資訊熵、頻數分別與預設的凝固度閾值、預設的資訊熵閾值、預設的頻數閾值進行比較，將凝固度小於等於預設的凝固度閾值、資訊熵小於等於預設的資訊熵閾

值、頻數小於等於預設的頻數閾值的候選字串確定為符合第一預設要求的第一特徵字串，即新詞。

在本場景示例中，上述預設的凝固度閾值、預設的資訊熵閾值、預設的頻數閾值的具體數值可以根據具體情況和處理要求靈活設置。對此，本說明書不作限定。

伺服器在根據文字資料確定出了上述第一特徵詞後，進一步，可以根據上述第一特徵詞對分詞庫進行更新，得到更新後的分詞庫。

上述分詞庫具體可以理解為一種基於歷史資料建立的，包含有多個用於表徵預設屬性類型的預設特徵詞的詞庫。其中，上述預設屬性類型具體可以包括：合法和非法等。其中，合法類型又可以進一步細化為：還款、消費、借出等類型。類似的，非法類型又可以進一步細化為多種不同的非法類型等。上述預設特徵詞具體可以理解為一種能夠用於表徵預設屬性類型的已有詞組(或已有詞)。

在本場景示例中，考慮到所應用的交易網路平臺為使用中文的交易網路平臺，主要分詞處理的文字資料為中文文字資料，因此，具體可以使用jieba(一種基於隱馬爾可夫模型(HMM)的分詞庫)作為上述分詞庫。由於jieba針對中文漢字詞組具有較好的匹配性，所以在本場景示例中利用jieba作為分詞庫可以使得後續的分詞處理更加準確。當然，需要說明的是，上述所列舉的jieba只是一種示意性說明。具體實施時，根據具體的應用場景和所涉及的語言類型還可以選擇使用其他類型的分詞庫。對此，本說明書不

作限定。

在本場景示例中，目的是為找到針對非法交易的新的特徵詞，因此，預設屬性類型為非法類型。相應的預設特徵詞為能夠表徵非法類型的已有詞組。具體的，例如，“上分”、“下分”是一種較為常見的與非法交易行為關聯的關鍵詞，因此，上述兩個詞組可以是分詞庫所包含的一種預設特徵詞。

具體實施時，可以將新確定的第一特徵詞，即新詞添加至已有分詞庫中，從而達到對已有分詞庫的更新、擴展，得到更加完整、準確的，包含有當前最新發現的新詞的更新後的分詞庫。這樣得到的更新後的分詞庫考慮到了當前剛出現的、已有分詞庫中沒有記載的新詞。因此，後續利用上述更新後的分詞庫取代已有分詞庫進行分詞處理時，可以避免由於使用已有分詞庫，導致將一些已有分詞庫中未記載的新詞進行錯誤拆分。

在得到了上述更加準確、完整的分詞庫後，伺服器可以根據上述更新後的分詞庫，從所述文字資料中提取第二特徵詞，即用於表徵預設屬性類型的新的特徵詞。

上述第二特徵詞具體可以理解為一種區別於預設特徵詞，伺服器之前沒有發現、確定的，能表徵預設屬性類型的新的詞組。

具體實施時，可以根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；再對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向

量；根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞。

在本場景示例中，由於是利用基於第一特徵詞更新後的分詞庫對文字資料進行分詞處理，從而可以有效避免了對一些新詞的錯誤拆分，使得的分詞處理得到的分詞單元更加準確。

在得到分詞單元後，具體實施時，可以對分詞單元分別進行向量化處理，得到能夠表徵分詞單元語義的詞向量，進而可以後續可以根據上述分詞單元的詞向量與預設特徵詞的詞向量，從所述分詞單元中尋找出符合第二預設要求(即語義與預設特徵詞相對較接近，所表徵的屬性類型較為相似)的分詞單元，作為第二特徵詞。

在本場景示例中，考慮到所應用的是使用的中文的交易網路平臺。而中文中漢字不同於英文中字母。英文字母是一種表音的文字，即單個的英文字母只表徵發音，而不表徵語義。而中文漢字是一種既表音又表音的文字(例如形聲字、象形字等)，即單個漢字除了表徵發音外，漢字本身的結構(例如漢字內部的筆劃組成)也能在一定程度上反映出語義。

現有的對分詞單元進行向量化處理大多是透過PageRank方法(一種詞向量化演算法)對漢字字元單元進行向量化處理。而PageRank方法本身是針對英文字母的特點設計，透過該方法得到的詞向量只能反映出詞的上下文外

部資訊，而無法反應出漢字內部結構所攜帶的語義資訊。因此，根據現有的特徵詞的確定方法利用 PageRank 方法對漢字組成的分詞單元進行向量處理得到的詞向量所包含的資訊往往會遺漏掉分詞單元中漢字內部結構所表徵的語義資訊，即資訊不夠完整、準確，會影響後續第二特徵詞的提取。

正是考慮到上述問題，為了使得所獲取的分詞單元的詞向量所攜帶的資訊更加豐富、完整，以進一步提高確定第二特徵詞的準確度，在本場景示例中，可以按照以下方式對多個分詞單元中的各個分詞單元分別進行詞向量化處理。具體的，伺服器可以先將分詞單元中的漢字拆分為多個筆劃；將不同的筆劃映射為不同的數值，根據所述分詞單元的多個筆劃，建立分詞單元的筆劃向量；同時，搜索並獲取文字資料中與所述分詞單元相連的詞語（例如左鄰詞和右鄰詞），作為上下文詞語；並按照常用方法獲取所述上下文詞語的詞向量；再根據所述分詞單元的筆劃向量和所述上下文詞語的詞向量，確定所述分詞單元的詞向量。這樣得到的分詞單元的詞向量即攜帶有分詞單元的上下文外部資訊，又包含了分詞單元中漢字的內部結構所反映的語義資訊，從而使得獲取的分詞單元的詞向量更加的準確、完整。

當然，上述所列舉的獲取分詞單元的詞向量的方式只是一種示意性說明。具體實施時，還可以透過 Cw2vec 演算法 (Learning Chinese Word Embeddings with Stroke n-

gram Information，一種詞向量化演算法)對所述多個分詞單元分別進行詞向量化處理，以得到分詞單元的詞向量。對此，本說明書不作限定。

在獲取得到了上述分詞單元的詞向量後，進一步，伺服器可以根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求(即與預設特徵詞語義或者所表徵的屬性類型相近)的分詞單元，作為所述第二特徵詞，即得到了能夠表徵預設屬性類型的新的特徵詞。

在本場景示例中，具體實施時，可以先從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞(也可稱為spy)；根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集(也可稱為黑樣本，positive，記為P)，並將該樣本集中的分詞單元的詞向量標記為1；根據所述測試詞的詞向量和所述分詞單元的詞向量，建立非標記樣本集(也可以稱為白樣本，定義為unlabeled，記為U)，並將該樣本集中的分詞單元的詞向量標記為0。需要說明的是，上述標記樣本集中的樣本數量小於非標記樣本集中的樣本數量。

在建立了上述標記樣本集和非標記樣本集後，可以根據所述標記樣本集和非標記樣本集，透過多次疊代擬合，確定擬合分數閾值。

具體的，可以按照以下步驟處理確定出擬合分數閾值：對標記樣本集和非標記樣本集中的分詞單元的詞向量

分別進行GBDT(Gradient Boosting Decision Tree，梯度提升決策樹)擬合，根據擬合結果對每個分詞單元的詞向量進行打分(記為score)；對擬合結果的分值，進一步作如下處理：將屬於P的分詞單元的詞向量的score置為1，其餘保持原來具體score，然後進行歸一化處理。按照上述步驟進行多次處理(例如重複兩次)，直到找出可以使得閾值比例(例如90%)的被歸入U中的預設特徵詞的詞向量(即spy)被辨識出來閾值(記為t)作為上述擬合分數閾值。

在確定出擬合分數閾值後，進一步可以根據上述擬合分數閾值，從分詞單元向量中確定出符合第二預設要求的分詞單元向量所對應的分詞單元作為第二特徵詞。

具體的，可以按照以下步驟處理確定出第二特徵詞：將被歸入U中的預設特徵詞的詞向量(即spy)重新歸屬回P中；再將剩下的U中所有擬合分數值小於擬合分數閾值(即t)的分詞單元的詞向量的score賦為0，P中所有預設特徵詞的詞向量的score賦為1，其餘保持當前的score然後進行歸一化處理；再對所有的詞向量進行GBDT擬合，並根據擬合結果對每個詞的詞向量進行重新打分得到擬合分數，記為score'；根據擬合分數，將屬於P的詞向量的擬合分數score'置為1，其餘保持原有的score'，然後進行歸一化處理。按照上述步驟進行多次處理(例如重複5次)，得到每個詞向量最終的擬合分數，記為score''，將score''大於特徵詞分數閾值(記為v)的詞向量所對應的分詞單元確定為第二特徵詞，即符合第二預設要求的新的能夠表徵該屬性

類型的詞組。其中，上述特徵詞分數閾值具體可以根據具體情況和精度要求設置。對於特徵詞分數閾值的具體取值，本說明書不作限定。

當然，上述所列舉的確定第二特徵詞的方式只是一種示意性說明。具體實施時，還可以透過PU\_learning演算法(Learning from Positive and Unlabeled Example，正例和無標記樣本學習演算法)對所述多個分詞單元的詞向量進行分析處理，以確定出第二特徵詞。對此，本說明書不作限定。

在本場景示例中，具體實施時，可以將特徵詞分數閾值為0.5，進而可以按照上述處理步驟逐步從多個分詞單元中確定出符合第二預設要求的能夠表徵違法的分詞單元，作為第二特徵詞，即新的特徵詞。

在按照上述方式確定出能夠表徵非法交易的新的特徵詞，即第二特徵詞後，伺服器進一步可以將第二特徵詞與預設特徵詞合併；根據合併後特徵詞對交易網路平臺上交易附言、文字標籤等文字資料進行檢測、辨識，以便及時發現交易網路平臺上的非法交易，並對非法交易進行有針對性的監控和處理。

由上述場景示例可見，本說明書提供的特徵詞的確定方法，由於透過先從文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞

錯誤導致的特徵詞提取不準確，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

在另一個場景示例中，為了避免一些無意義的字串的干擾，減少後續工作量，提高處理效率，具體實施時，可以參閱圖3所示，在從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串之前，所述方法還可以包括以下內容：過濾所述文字資料中的無效字串，得到過濾後的文字資料，再從過濾後的文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字元。

在本場景示例中，上述無效字串具體可以理解為不含漢字，或明顯不會構成詞組的字串。具體的，上述無效字串可以是全由字母、數字或表情符號組成的字串，也可以是網頁鏈接，還可以用於表徵繁體轉簡體的文字資料等等。當然，需要說明的是，上述所列舉的無效字串只是一種示意性說明。具體實施時，根據具體的應用場景，上述無效字串具體還可以包括其他類型的字串。對此，本說明書不作限定。

在另一個場景示例中，為了避免無意義的分詞單元的干擾，進一步減少後續工作量，提高處理效率，具體實施時，可以參閱圖4所示，在根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞前，所述方法還可以包括以下內容：過濾所述分詞單元的詞向量中的停用詞的詞向量，得到過濾後的分詞單元的詞向量；再從

過濾後的分詞單元的詞向量中確定出符合第二預設要求的分詞單元。

在本場景示例中，上述停用詞具體可以理解為所表徵的內容沒有實際意義或與交易資料的屬性類型無關的詞組。具體的，上述停用詞可以是一些連接詞或助詞，例如“的”、“是”、“了”等，也可以是一些與交易資料的無關，寬泛的代詞，例如“我”、“這”、“那”，還可以是數字、字母或單個字的詞等等。當然，需要說明的是上述所列舉的停用詞只是一種示意性說明。具體實施時，根據具體的應用場景，上述停用詞還可以包括其他的詞，例如“在”、“有”、“人”、“一”等。對於上述停用詞的具體內容，本說明書不作限定。

由上述場景示例可見，本說明書提供的特徵詞的確定方法，由於透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果；又透過先將分詞單元中的漢字拆分為多個筆劃，得到分詞單元的筆劃向量；再根據分詞單元的筆劃向量和上下文詞語的詞向量，確定分詞單元的詞向量，從而使得得到的分詞單元的詞向量同時包含上下文外部資訊和漢字內部構造資訊，能夠反映出更加豐富、準確的語義資訊，再基於上述

分詞單元的詞向量進行第二特徵詞的提取，提高了確定特徵詞的準確度；還透過先根據從多個預設特徵詞中提取出的預設個數的預設特徵詞和分詞單元建立非標記樣本集，根據剩餘的預設特徵詞建立標記樣本集；再基於上述非標記樣本集和標記樣本透過多次疊代擬合，確定出較為準確的擬合分數閾值，以便根據上述擬合分數閾值從分詞單元中確定出第二特徵詞，進一步提高了確定特徵詞的準確度。

參閱圖5所示，本說明書實施例提供了一種特徵詞的確定方法，其中，該方法具體應用於後臺伺服器一側。具體實施時，該方法可以包括以下內容：

S51：獲取文字資料。

在本實施方式中，上述文字資料具體可以理解為一種包含有漢字、英文字母、拼音符號或數字等資訊的資料。具體的，上述文字資料可以包括交易附言，也可以包括文字標籤等。當然，上述所列舉的交易附言、文字標籤只是為了更好地說明本說明書實施方式。具體實施時，根據具體的應用場景和使用需求，上述文字資料還可以包括有其他類型或形式的資料。對此，本說明書不作限定。

其中，上述交易附言具體可以理解為一種由交易資料發起者自定義編輯的用以描述交易資料的相關特徵的資訊資料。具體的，交易資料發起者在發送交易資料時，可以透過預設的文字資料輸入介面，輸入或設置用以描述交易資料的例如交易目的、交易時間或交易對象等特徵資訊，

作為交易附言。上述文字標籤具體可以理解為一種由交易網路平臺採集並根據交易資料發起者的所實施的與該筆交易資料相關操作，自動生成的用以描述該筆交易資料的類型、目的或內容等的資料標籤。

在本實施方式中，上述獲取文字資料具體可以包括：伺服器搜集得到預設時間段(例如近一個月)內交易網路平臺上的交易附言及/或文字標籤等資料作為文字資料。

S53：從所述文字資料中提取出第一特徵詞。

在本實施方式中，上述第一特徵詞具體可以理解為一種新詞，即之前未被伺服器發現，也沒有被記錄在伺服器的資料庫或分詞庫中的特定的漢字組合。

在本實施方式中，具體實施時，伺服器可以透過新詞發現演算法對上述文字資料進行分析處理，從上述文字資料中提取得到第一特徵詞。

在本實施方式中，從所述文字資料中提取出第一特徵詞，具體可以包括以下內容：從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串；計算所述候選字串的指標參數；根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞。

在本實施方式中，透過從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串，可以先將明顯不符合組詞習慣的，具有較大機率不能構成詞組的字串事先過濾掉，從而可以避免上述類型字串的干擾，減少工

作量，提高處理效率。

在本實施方式中，上述預設長度閾值具體可以設置為5個字元長度。當然，需要說明的是，上述所列舉的預設長度閾值只是一種示意性說明。具體實施時，根據具體情況還可以設置其他數值作為上述預設長度閾值。對此，本說明書不作限定。

在本實施方式中，上述指標參數具體可以理解為一種用於表徵字串能否構造詞組的特徵參數。通常一個候選字串的指標參數滿足一定的要求，則可以認為該字串可以構成一個詞組，例如構成新詞。

在本實施方式中，在從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串後，在計算所述候選字串的指標參數前，所述方法還包括：根據已有的資料庫，過濾上述候選字串中與已有的資料庫中的字串相同的字串，保留已有的資料庫中沒有的候選字串，以便進行指標參數的計算。這樣可以先過濾掉已有的資料庫中所記錄的已有詞組所對應的字串，減少了後續處理的工作量，提高了處理效率。

在本實施方式中，上述指標參數具體可以包括以下至少之一：凝固度、資訊熵和頻數等。當然，需要說明的是，上述所列舉的指標參數只是為了更好地說明本說明書實施方式。具體實施時，根據具體情況還可以引入其他類型的特徵參數作為上述指標參數。例如，互資訊、tf-idf等詞頻資訊。對此，本說明書不作限定。

其中，上述凝固度具體可以理解為文字資料中字串中所包含的字元結合在一起的機率。通常一個字串的凝固度數值越高，該字串具有越高的機率成為一個詞組。上述資訊熵，也可稱為自由度，具體可以理解為字串表徵的某種特定資訊的出現機率。通常當字串所表徵的某種特定資訊的出現機率較高的時候，表明它被傳播得較為廣泛，或者說，被引用的程度相對較高。上述頻數具體可以理解為字串在文字資料中出現的次數。通常一個字串的頻數越高，相對的具有較大機率成為一個詞組。

在本實施方式中，上述根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞，具體實施時，可以包括以下內容：將各個候選字串的凝固度、資訊熵、頻數分別與預設的凝固度閾值、預設的資訊熵閾值、預設的頻數閾值進行比較；根據比較結果，將凝固度小於等於預設的凝固度閾值、資訊熵小於等於預設的資訊熵閾值、頻數小於等於預設的頻數閾值的候選字串確定為符合第一預設要求的第一特徵字串，即新詞。

在本實施方式中，上述預設的凝固度閾值、預設的資訊熵閾值、預設的頻數閾值的具體數值可以根據具體情況和處理要求靈活設置。對此，本說明書不作限定。

S55：根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞。

在本實施方式中，上述分詞庫具體可以理解為一種基於歷史資料建立的，包含有多個用於表徵預設屬性類型的預設特徵詞的詞庫。

在本實施方式中，上述預設屬性類型具體可以包括：合法和非法等。其中，合法類型又可以進一步細化為：還款、消費、借出等類型。類似的，非法類型又可以進一步細化為多種不同的非法類型等。

在本實施方式中，上述預設特徵詞具體可以理解為一種能夠用於表徵預設屬性類型的已有詞組(或已有詞)。其中，預設特徵詞與所表徵的預設屬性類型對應。例如，用於表徵違法類型的預設特徵詞也可以稱為黑詞等。

在本實施方式中，進一步考慮到所應用的交易網路平臺為使用中文的交易網路平臺，主要分詞處理的文字資料為中文文字資料，因此，具體可以使用jieba(一種基於隱馬爾可夫模型(HMM)的分詞庫)作為上述分詞庫。由於jieba針對中文漢字詞組具有較好的匹配性，所以在本實施方式中利用jieba作為分詞庫可以使得後續的分詞處理更加準確。當然，需要說明的是，上述所列舉的jieba只是一種示意性說明。具體實施時，根據具體的應用場景和所涉及的語言類型還可以選擇使用其他類型的分詞庫。對此，本說明書不作限定。

在本實施方式中，上述更新後的分詞庫是在基於歷史資料得到的已有分詞庫的基礎上，增添了新確定的新詞(即第一特徵詞)，擴展後的分詞庫。

在本實施方式中，根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，具體實施時，可以包括：可以將所確定的第一特徵詞，添加至已有分詞庫，從而達到對已有分詞庫的更新、擴展，得到更加完整、準確的，包含有當前最新發現的新詞的更新後的分詞庫。這樣後續利用上述更新後的分詞庫取代已有分詞庫進行分詞處理時，可以避免由於使用已有分詞庫，導致將一些已有分詞庫中未記載的新詞進行錯誤拆分。

S57：根據更新後的分詞庫，從所述文字資料中提取第二特徵詞根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

在本實施方式中，上述第二特徵詞具體可以理解為一種區別於預設特徵詞，之前沒有發現、確定的，能表徵預設屬性類型的新的詞組。

在本實施方式中，上述根據更新後的分詞庫，從所述文字資料中提取第二特徵詞根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞，具體實施時，可以包括以下內容：根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量；根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞。

由上可見，本說明書實施例提供的特徵詞的確定方

法，由於透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

在一個實施方式中，所述文字資料具體可以包括：交易附言，及/或，文字標籤等。當然，需要說明的是上述所列舉的交易附言、文字標籤只是為了更好地說明本說明書實施方式。具體實施時，根據具體的應用場景和使用需求，上述文字資料還可以包括有其他類型或形式的資料。對此，本說明書不作限定。

在一個實施方式中，從所述文字資料中提取出第一特徵詞，具體實施時，可以包括以下內容：從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串；計算所述候選字串的指標參數；根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞。

在一個實施方式中，所述指標參數具體可以包括以下至少之一：凝固度、資訊熵和頻數等。當然，需要說明的是，上述所列舉的指標參數只是為了更好地說明本說明書實施方式。具體實施時，根據具體情況還可以引入其他類型的特徵參數作為上述指標參數。對此，本說明書不作限定。

在一個實施方式中，在從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串之前，所述方法具體實施時還可以包括以下內容：過濾所述文字資料中的無效字串。

在本實施方式中，考慮到所獲取的文字資料中可能包含有大量無意義的字串，這類字串通常不含有漢字，或明顯不會構成詞組，在本實施方式中，這類字串可以稱為無效字串。具體的，可以是全由字母、數字或表情符號組成的字串，也可以是網頁鏈接，還可以用於表徵繁體轉簡體的文字資料等等。

在本實施方式中，為了避免後續的分析處理受到上述無效字串的干擾，提高處理效率，具體實施時，可以在從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串之前，先將文字資料中無效字串過濾掉(或稱清洗掉)，得到過濾後的字串，後續再基於過濾後的字串進行候選字串的確定。

在一個實施方式中，上述根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞，具體實施時，可以包括以下內容：根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量；根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞。

在一個實施方式中，上述對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量，具體可以包括以下內容：將分詞單元中的漢字拆分為多個筆劃；根據所述分詞單元的多個筆劃，建立分詞單元的筆劃向量；獲取文字資料中與所述分詞單元相連的詞語，作為上下文詞語；獲取所述上下文詞語的詞向量；根據所述分詞單元的筆劃向量和所述上下文詞語的詞向量，確定所述分詞單元的詞向量。

在本實施方式中，考慮到現有的詞向量化演算法(例如 **PageRank** 方法)大多是針對英文字母的特點設計的，而單獨的英文字母本身只能表音而無法表意。而在本實施方式中，所針對的分詞單元主要是由漢字組成的。漢字不同於英文字母除了表音還能表意。即漢字本身的內部結構也能反映相應的語義資訊。如果直接利用例如 **PageRank** 方法對由漢字構成的分詞單元進行向量化處理得到的詞向量所包含的資訊往往會遺漏掉分詞單元中漢字內部結構所表徵的語義資訊，即資訊不夠完整、準確，會影響後續第二特徵詞的提取。

正是考慮到上述問題，為了能夠得到資訊更為完整、準確的詞向量，在本實施方式不是直接利用已有的 **PageRank** 等演算法進行向量化處理，而是針對漢字的結構特點，根據分詞單元中漢字的筆劃，建立分詞單元的筆劃向量，以表徵漢字的內部結構所攜帶的語義資訊；並根據基於文字資料中與分詞單元相連的詞語(例如左鄰詞和右鄰詞)，得

到上下文詞語的詞向量，以表徵文字資料中分詞單元的上下文外部資訊；綜合上述兩種不同的詞向量，得到更為準確、完整的分詞單元的詞向量，有助於後續更準確地確定出第二特徵詞。

在一個實施方式中，具體實施時，還可以透過 **Cw2vec** 演算法對所述多個分詞單元分別進行詞向量化處理，以得到分詞單元的詞向量。

在一個實施方式中，根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞，具體實施時，可以包括以下內容：從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞；根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集；根據所述測試詞的詞向量和所述分詞單元的詞向量，建立非標記樣本集；根據所述標記樣本集和非標記樣本集，透過疊代擬合，確定擬合分數閾值；根據所述擬合分數閾值，從所述分詞單元中確定出符合第二預設要求的分詞單元，作為所述第二特徵詞。

在本實施方式中，具體實施時，可以先從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞（也可稱為 **spy**）；根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集（也可稱為黑樣本，**postive**，記為 **P**），並將該樣本集中的分詞單元的詞向量標記為 1；根據所述測試詞的詞向量和所述分詞單元的詞向

量，建立非標記樣本集（也可以稱為白樣本，定義為 **unlabeled**，記為 **U**），並將該樣本集中的分詞單元的詞向量標記為 0。需要說明的是，上述標記樣本集中的樣本數量小於非標記樣本集中的樣本數量。

在建立了上述標記樣本集和非標記樣本集後，可以根據所述標記樣本集和非標記樣本集，透過多次疊代擬合，確定擬合分數閾值。

具體的，可以按照以下步驟處理確定出擬合分數閾值：對標記樣本集和非標記樣本集中的分詞單元的詞向量分別進行 **GBDT**(**Gradient Boosting Decision Tree**，梯度提升決策樹)擬合，根據擬合結果對每個分詞單元的詞向量進行打分（記為 **score**）；對擬合結果的分值，進一步作如下處理：將屬於 **P** 的分詞單元的詞向量的 **score** 置為 1，其餘保持原來具體 **score**，然後進行歸一化處理。按照上述步驟進行多次處理（例如重複兩次），直到找出可以使得閾值比例（例如 90%）的被歸入 **U** 中的預設特徵詞的詞向量（即 **spy**）被辨識出來閾值（記為 **t**）作為上述擬合分數閾值。

在確定出擬合分數閾值後，進一步可以根據上述擬合分數閾值，從分詞單元向量中確定出符合第二預設要求的分詞單元向量所對應的分詞單元作為第二特徵詞。

具體的，可以按照以下步驟處理確定出第二特徵詞：將被歸入 **U** 中的預設特徵詞的詞向量（即 **spy**）重新歸屬回 **P** 中；再將剩下的 **U** 中所有擬合分數值小於擬合分數閾值（即 **t**）的分詞單元的詞向量的 **score** 賦為 0，**P** 中所有預設特徵詞

的詞向量的 `score` 賦為 1，其餘保持當前的 `score` 然後進行歸一化處理；再對所有的詞向量進行 **GBDT** 擬合，並根據擬合結果對每個詞的詞向量進行重新打分得到擬合分數，記為 `score'`；根據擬合分數，將屬於 **P** 的詞向量的擬合分數 `score'` 置為 1，其餘保持原有的 `score'`，然後進行歸一化處理。按照上述步驟進行多次處理（例如重複 5 次），得到每個詞向量最終的擬合分數，記為 `score''`，將 `score''` 大於特徵詞分數閾值（記為 `v`）的詞向量所對應的分詞單元確定為第二特徵詞，即符合第二預設要求的新的能夠表徵該屬性類型的詞組。其中，上述特徵詞分數閾值具體可以根據具體情況和精度要求設置。對於特徵詞分數閾值的具體取值，本說明書不作限定。

在一個實施方式中，具體實施時，還可以透過 **PU\_learning** 演算法對所述多個分詞單元的詞向量進行分析處理，以確定出第二特徵詞。對此，本說明書不作限定。

在一個實施方式中，在根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞前，所述方法還可以包括以下內容：過濾所述分詞單元的詞向量中的停用詞的詞向量。

在本實施方式中，上述停用詞具體可以理解為所表徵的內容沒有實際意義或與交易資料的屬性類型無關的詞組。具體的，上述停用詞可以是一些連接詞或助詞，例如“的”、“是”、“了”等，也可以是一些與交易資料的

無關，寬泛的代詞，例如“我”、“這”、“那”，還可以是數字、字母或單個字的詞等等。當然，需要說明的是上述所列舉的停用詞只是一種示意性說明。具體實施時，根據具體的應用場景，上述停用詞還可以包括其他的詞，例如“在”、“有”、“人”、“一”等。對於上述停用詞的具體內容，本說明書不作限定。

在本實施方式中，透過在根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞前，過濾分詞單元的詞向量中的停用詞的詞向量，可以避免後續在對停用詞的詞向量進行分析處理時對時間、資源的佔用，從而可以減少工作量，提高處理效率。

由上可見，本說明書實施例提供的特徵詞的確定方法，由於透過先對文字資料進行新詞提取得到第一特徵詞；再利用第一特徵詞對分詞庫進行更新；進而利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果；又透過先將分詞單元中的漢字拆分為多個筆劃，得到分詞單元的筆劃向量；再根據分詞單元的筆劃向量和上下文詞語的詞向量，確定分詞單元的詞向量，從而使得得到的分詞單元的詞向量同時包含上下文外部資訊和漢字內部構造資訊，能夠反映出更加豐富、準確的語義資訊；再基於上述

分詞單元的詞向量進行第二特徵詞的提取，提高了確定特徵詞的準確度；還透過先根據從多個預設特徵詞中提取出的預設個數的預設特徵詞和分詞單元建立非標記樣本集，根據剩餘的預設特徵詞建立標記樣本集；再基於上述非標記樣本集和標記樣本透過多次疊代擬合，確定出較為準確的擬合分數閾值，以便根據上述擬合分數閾值從分詞單元中確定出第二特徵詞，進一步提高了確定特徵詞的準確度。

本說明書實施例還提供了一種伺服器，包括處理器以及用於儲存處理器可執行指令的儲存器，所述處理器具體實施時可以根據指令執行以下步驟：獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

為了能夠更加準確地完成上述指令，參閱圖6所示，本說明書還提供了另一種具體的伺服器，其中，所述伺服器包括網路通訊端口601、處理器602以及儲存器603，上述結構透過內部線纜相連，以便各個結構可以進行具體的資料交互。

其中，所述網路通訊端口601，具體可以用於獲取文字資料。

所述處理器602，具體可以用於從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得

到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

所述儲存器 603，具體可以用於儲存經網路通訊端口 601輸入的文字資料，以及相應的指令程式。

在本實施方式中，所述網路通訊端口 601可以是與不同的通訊協定進行綁定，從而可以發送或接收不同資料的虛擬端口。例如，所述網路通訊端口可以是負責進行 web 資料通訊的 80號端口，也可以是負責進行 FTP資料通訊的 21號端口，還可以是負責進行郵件資料通訊的 25號端口。此外，所述網路通訊端口還可以是實體的通訊介面或者通訊晶片。例如，其可以為無線行動網路通訊晶片，如 GSM、CDMA等；其還可以為 Wifi晶片；其還可以為藍牙晶片。

在本實施方式中，所述處理器 602可以按任何適當的方式實現。例如，處理器可以採取例如微處理器或處理器以及儲存可由該(微)處理器執行的電腦可讀程式代碼(例如軟體或韌體)的電腦可讀媒體、邏輯閘、開關、專用積體電路(Application Specific Integrated Circuit，ASIC)、可程式化邏輯控制器和嵌入微控制器的形式等等。本說明書並不作限定。

在本實施方式中，所述儲存器 603可以包括多個層次，在數位系統中，只要能保存二進制資料的都可以是儲存器；在積體電路中，一個沒有實物形式的具有儲存功能

的電路也叫儲存器，如RAM、FIFO等；在系統中，具有實物形式的儲存設備也叫儲存器，如記憶體條、TF卡等。

本說明書實施例還提供了一種基於上述特徵詞的確定方法的電腦儲存媒體，所述電腦儲存媒體儲存有電腦程式指令，在所述電腦程式指令被執行時實現：獲取文字資料；從所述文字資料中提取出第一特徵詞；根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

在本實施方式中，上述儲存媒體包括但不限於隨機存取儲存器(Random Access Memory, RAM)、唯讀儲存器(Read-Only Memory, ROM)、快取(Cache)、硬碟(Hard Disk Drive, HDD)或者儲存卡(Memory Card)。所述儲存器可以用於儲存電腦程式指令。網路通訊單元可以是依照通訊協定規定的標準設置的，用於進行網路連接通訊的介面。

在本實施方式中，該電腦儲存媒體儲存的程式指令具體實現的功能和效果，可以與其它實施方式對照解釋，在此不再贅述。

參閱圖7所示，在軟體層面上，本說明書實施例還提供了一種特徵詞的確定裝置，該裝置具體可以包括以下的結構模組：

獲取模組71，具體可以用於獲取文字資料；

第一提取模組72，具體可以用於從所述文字資料中提

取出第一特徵詞；

更新模組 73，具體可以用於根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；

第二提取模組 74，具體可以用於根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞。

在一個實施方式中，所述文字資料具體可以包括：交易附言，及/或，文字標籤等。當然，需要說明的是，上述所列舉的文字資料只是一種示意性說明。對於文字資料的具體類型、形式和內容，本說明書不作限定。

在一個實施方式中，所述第一提取模組 72 具體可以包括以下結構單元：

篩選單元，具體可以用於從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串；

計算單元，具體可以用於計算所述候選字串的指標參數；

提取單元，具體可以用於根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞。

在一個實施方式中，所述指標參數具體可以包括以下至少之一：凝固度、資訊熵和頻數等。當然，需要說明的是，上述所列舉的指標參數只是一種示意性說明。具體實施時，根據具體情況和要求，還可以引入其他類型的特徵參數作為上述指標參數。對此，本說明書不作限定。

在一個實施方式中，所述第一提取模組72具體還可以包括：第一過濾單元，具體可以用於過濾所述文字資料中的無效字串。

在一個實施方式中，所述第二提取模組74具體可以包括以下結構單元：

第一處理單元，具體可以用於根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；

第二處理單元，具體可以用於對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量；

確定單元，具體可以用於根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞。

在一個實施方式中，所述第二處理單元具體可以包括以下結構子單元：

拆分子單元，具體可以用於將分詞單元中的漢字拆分為多個筆劃；

第一建立子單元，具體可以用於根據所述分詞單元的多個筆劃，建立分詞單元的筆劃向量；

第一獲取子單元，具體可以用於獲取文字資料中與所述分詞單元相連的詞語，作為上下文詞語；

第二獲取子單元，具體可以用於獲取所述上下文詞語的詞向量；

第三獲取子單元，具體可以用於根據所述分詞單元的筆劃向量和所述上下文詞語的詞向量，確定所述分詞單元

的詞向量。

在一個實施方式中，所述第二處理單元具體可以透過 **Cw2vec** 演算法對所述多個分詞單元分別進行詞向量化處理，以得到分詞單元的詞向量。

在一個實施方式中，所述確定單元具體可以包括以下結構子單元：

提取子單元，具體可以用於從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞；

第二建立子單元，具體可以用於根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集；

第三建立子單元，具體可以用於根據所述測試詞的詞向量和所述分詞單元的詞向量，建立非標記樣本集；

第一確定子單元，具體可以用於根據所述標記樣本集和非標記樣本集，透過疊代擬合，確定擬合分數閾值；

第二確定子單元，具體可以用於根據所述擬合分數閾值，從所述分詞單元中確定出符合第二預設要求的分詞單元，作為所述第二特徵詞。

在一個實施方式中，所述第二提取模組具體還可以包括過第二過濾單元，具體可以用於過濾所述分詞單元的詞向量中的停用詞的詞向量。

在一個實施方式中，所述確定單元具體可以透過 **PU\_learning** 演算法對所述多個分詞單元的詞向量進行分析處理，以確定出第二特徵詞。

需要說明的是，上述實施例闡明的單元、裝置或模組等，具體可以由電腦晶片或實體實現，或者由具有某種功能的產品來實現。為了描述的方便，描述以上裝置時以功能分為各種模組分別描述。當然，在實施本說明書時可以把各模組的功能在同一個或多個軟體及/或硬體中實現，也可以將實現同一功能的模組由多個子模組或子單元的組合實現等。以上所描述的裝置實施例僅僅是示意性的，例如，所述單元的劃分，僅僅為一種邏輯功能劃分，實際實現時可以有另外的劃分方式，例如多個單元或組件可以結合或者可以整合到另一個系統，或一些特徵可以忽略，或不執行。另一點，所顯示或討論的相互之間的耦合或直接耦合或通訊連接可以是透過一些介面，裝置或單元的間接耦合或通訊連接，可以是電性，機械或其它的形式。

由上可見，本說明書實施例提供的特徵詞的確定裝置，由於透過第一提取模組先對文字資料進行新詞提取得到第一特徵詞；再透過更新模組利用第一特徵詞對分詞庫進行更新；進而透過第二提取模組利用更新後的分詞庫和預設特徵詞從文字資料提取出新的特徵詞作為第二特徵詞，從而避免了提取特徵詞過程中，由於分詞錯誤導致的特徵詞提取不準確、不全面，達到能精確地從文字資料中挖掘出符合要求的新的特徵詞的技術效果。

雖然本說明書提供了如實施例或流程圖所述的方法操作步驟，但基於常規或者無創造性的手段可以包括更多或者更少的操作步驟。實施例中列舉的步驟順序僅僅為眾多

步驟執行順序中的一種方式，不代表唯一的執行順序。在實際中的裝置或客戶端產品執行時，可以按照實施例或者圖式所示的方法順序執行或者並行執行(例如並行處理器或者多執行緒處理的環境，甚至為分散式資料處理環境)。術語“包括”、“包含”或者其任何其他變體意在涵蓋非排他性的包含，從而使得包括一系列要素的過程、方法、產品或者設備不僅包括那些要素，而且還包括沒有明確列出的其他要素，或者是還包括為這種過程、方法、產品或者設備所固有的要素。在沒有更多限制的情況下，並不排除在包括所述要素的過程、方法、產品或者設備中還存在另外的相同或等同要素。第一，第二等詞語用來表示名稱，而並不表示任何特定的順序。

本領域技術人員也知道，除了以純電腦可讀程式代碼方式實現控制器以外，完全可以透過將方法步驟進行邏輯程式化來使得控制器以邏輯閘、開關、專用積體電路、可程式化邏輯控制器和嵌入微控制器等的形式來實現相同功能。因此這種控制器可以被認為是一種硬體部件，而對其內部包括的用於實現各種功能的裝置也可以視為硬體部件內的結構。或者甚至，可以將用於實現各種功能的裝置視為既可以是實現方法的軟體模組又可以是硬體部件內的結構。

本說明書可以在由電腦執行的電腦可執行指令的一般上下文中描述，例如程式模組。一般地，程式模組包括執行特定任務或實現特定抽象資料類型的常式、程式、對

象、組件、資料結構、類等等。也可以在分散式計算環境中實踐本說明書，在這些分散式計算環境中，由透過通訊網路而被連接的遠端處理設備來執行任務。在分散式計算環境中，程式模組可以位於包括儲存設備在內的本地和遠端電腦儲存媒體中。

透過以上的實施方式的描述可知，本領域的技術人員可以清楚地瞭解到本說明書可借助軟體加必需的通用硬體平臺的方式來實現。基於這樣的理解，本說明書的技術方案本質上或者說對現有技術做出貢獻的部分可以以軟體產品的形式體現出來，該電腦軟體產品可以儲存在儲存媒體中，如 ROM/RAM、磁碟、光碟等，包括若干指令用以使得一台電腦設備(可以是個人電腦，行動終端，伺服器，或者網路設備等)執行本說明書各個實施例或者實施例的某些部分所述的方法。

本說明書中的各個實施例採用遞進的方式描述，各個實施例之間相同或相似的部分互相參見即可，每個實施例重點說明的都是與其他實施例的不同之處。本說明書可用於眾多通用或專用的電腦系統環境或配置中。例如：個人電腦、伺服器電腦、手持設備或便攜式設備、平板型設備、多處理器系統、基於微處理器的系統、機上盒、程式化的電子設備、網路 PC、小型電腦、大型電腦、包括以上任何系統或設備的分散式計算環境等等。

雖然透過實施例描繪了本說明書，本領域普通技術人員知道，本說明書有許多變形和變化而不脫離本說明書的

精神，希望所附的請求項包括這些變形和變化而不脫離本說明書的精神。

**【符號說明】**

S51：步驟

S53：步驟

S55：步驟

S57：步驟

601：網路通訊介面

602：處理器

603：儲存器

71：獲取模組

72：第一提取模組

73：更新模組

74：第二提取模組

## 【發明申請專利範圍】

### 【第 1 項】

一種特徵詞的確定方法，包括：

獲取文字資料；

從所述文字資料中提取出第一特徵詞；

根據所述第一特徵詞，更新分詞庫，得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；

根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞；

其中，根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞，包括：從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞；根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集；根據所述測試詞的詞向量和所述分詞單元的詞向量，建立非標記樣本集；根據所述標記樣本集和非標記樣本集，透過疊代擬合，確定擬合分數閾值；根據所述擬合分數閾值，從所述分詞單元中確定出符合第二預設要求的分詞單元，作為所述第二特徵詞；所述分詞單元根據所述更新後的分詞庫得到。

### 【第 2 項】

根據請求項 1 所述的方法，所述文字資料包括：交易附言，及/或，文字標籤。

### 【第 3 項】

根據請求項 1 所述的方法，從所述文字資料中提取出第一特徵詞，包括：

從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串；

計算所述候選字串的指標參數；

根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞。

【第 4 項】

根據請求項 3 所述的方法，所述指標參數包括以下至少之一：凝固度、資訊熵和頻數。

【第 5 項】

根據請求項 3 所述的方法，在從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串之前，所述方法還包括：

過濾所述文字資料中的無效字串。

【第 6 項】

根據請求項 1 所述的方法，根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞，包括：

根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；

對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量；

根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單

元，作為所述第二特徵詞。

【第7項】

根據請求項6所述的方法，對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量，包括：

將分詞單元中的漢字拆分為多個筆劃；

根據所述分詞單元的多個筆劃，建立分詞單元的筆劃向量；

獲取文字資料中與所述分詞單元相連的詞語，作為上下文詞語；

獲取所述上下文詞語的詞向量；

根據所述分詞單元的筆劃向量和所述上下文詞語的詞向量，確定所述分詞單元的詞向量。

【第8項】

根據請求項6所述的方法，在根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞前，所述方法還包括：

過濾所述分詞單元的詞向量中的停用詞的詞向量。

【第9項】

一種特徵詞的確定裝置，包括：

獲取模組，用於獲取文字資料；

第一提取模組，用於從所述文字資料中提取出第一特徵詞；

更新模組，用於根據所述第一特徵詞，更新分詞庫，

得到更新後的分詞庫，其中，所述分詞庫包含多個用於表徵預設屬性類型的預設特徵詞；

第二提取模組，用於根據所述更新後的分詞庫，從所述文字資料中提取第二特徵詞；

其中，所述第二提取模組具體用於從多個預設特徵詞中提取預設數量的預設特徵詞作為測試詞；根據所述多個預設特徵詞中除測試詞以外的預設特徵詞的詞向量，建立標記樣本集；根據所述測試詞的詞向量和所述分詞單元的詞向量，建立非標記樣本集；根據所述標記樣本集和非標記樣本集，透過疊代擬合，確定擬合分數閾值；根據所述擬合分數閾值，從所述分詞單元中確定出符合第二預設要求的分詞單元，作為所述第二特徵詞；所述分詞單元根據所述更新後的分詞庫得到。

#### 【第10項】

根據請求項9所述的裝置，所述文字資料包括：交易附言，及/或，文字標籤。

#### 【第11項】

根據請求項9所述的裝置，所述第一提取模組包括：

篩選單元，用於從所述文字資料中篩選出字元長度小於預設長度閾值的字串作為候選字串；

計算單元，用於計算所述候選字串的指標參數；

提取單元，用於根據所述指標參數，從所述候選字串中提取符合第一預設要求的候選字串，作為第一特徵詞。

#### 【第12項】

根據請求項 11 所述的裝置，所述指標參數包括以下至少之一：凝固度、資訊熵和頻數。

**【第 13 項】**

根據請求項 11 所述的裝置，所述第一提取模組還包括：第一過濾單元，用於過濾所述文字資料中的無效字串。

**【第 14 項】**

根據請求項 9 所述的裝置，所述第二提取模組包括：

第一處理單元，用於根據所述更新後的分詞庫，對所述文字資料進行分詞處理，得到多個分詞單元；

第二處理單元，用於對所述多個分詞單元分別進行詞向量化處理，得到分詞單元的詞向量；

確定單元，用於根據所述分詞單元的詞向量和預設特徵詞的詞向量，從所述多個分詞單元中確定符合第二預設要求的分詞單元，作為所述第二特徵詞。

**【第 15 項】**

根據請求項 14 所述的裝置，所述第二處理單元包括：

拆分子單元，用於將分詞單元中的漢字拆分為多個筆劃；

第一建立子單元，用於根據所述分詞單元的多個筆劃，建立分詞單元的筆劃向量；

第一獲取子單元，用於獲取文字資料中與所述分詞單元相連的詞語，作為上下文詞語；

第二獲取子單元，用於獲取所述上下文詞語的詞向

量；

第三獲取子單元，用於根據所述分詞單元的筆劃向量和所述上下文詞語的詞向量，確定所述分詞單元的詞向量。

**【第16項】**

根據請求項14所述的裝置，所述第二提取模組還包括第二過濾單元，用於過濾所述分詞單元的詞向量中的停用詞的詞向量。

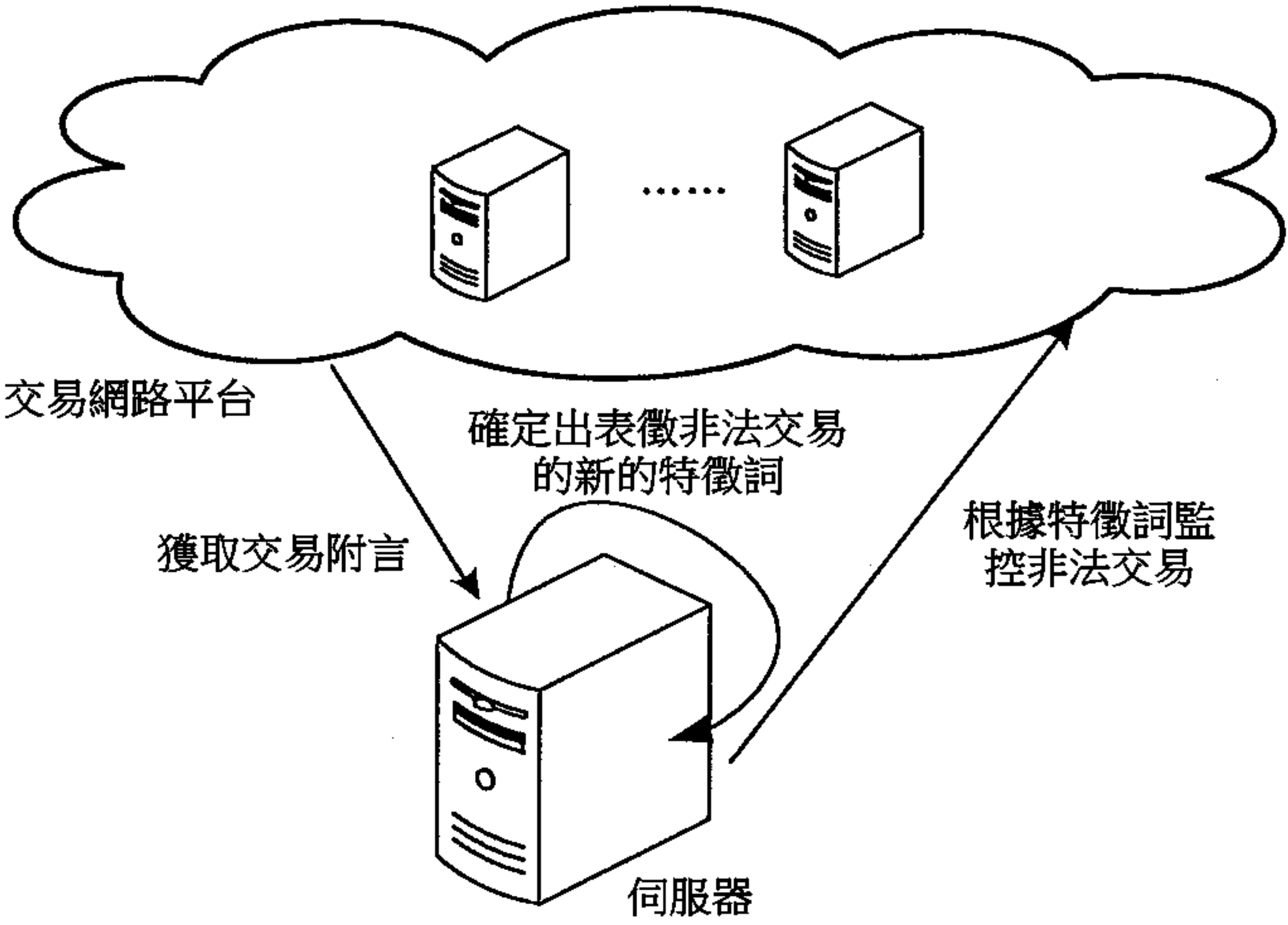
**【第17項】**

一種伺服器，包括處理器以及用於儲存處理器可執行指令的儲存器，所述處理器執行所述指令時實現請求項1至8中任一項所述方法的步驟。

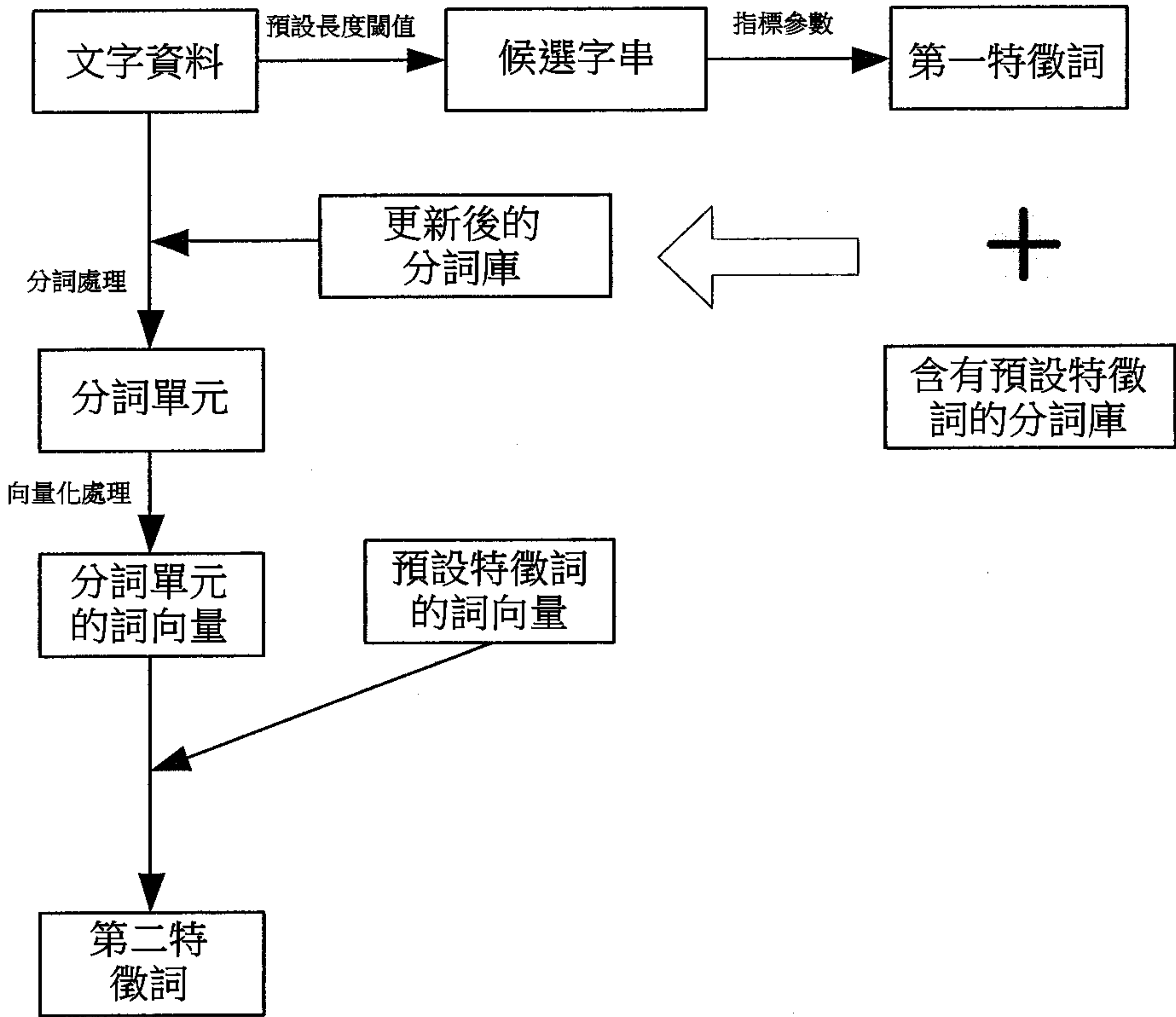
**【第18項】**

一種電腦可讀儲存媒體，其上儲存有電腦指令，所述指令被執行時實現請求項1至8中任一項所述方法的步驟。

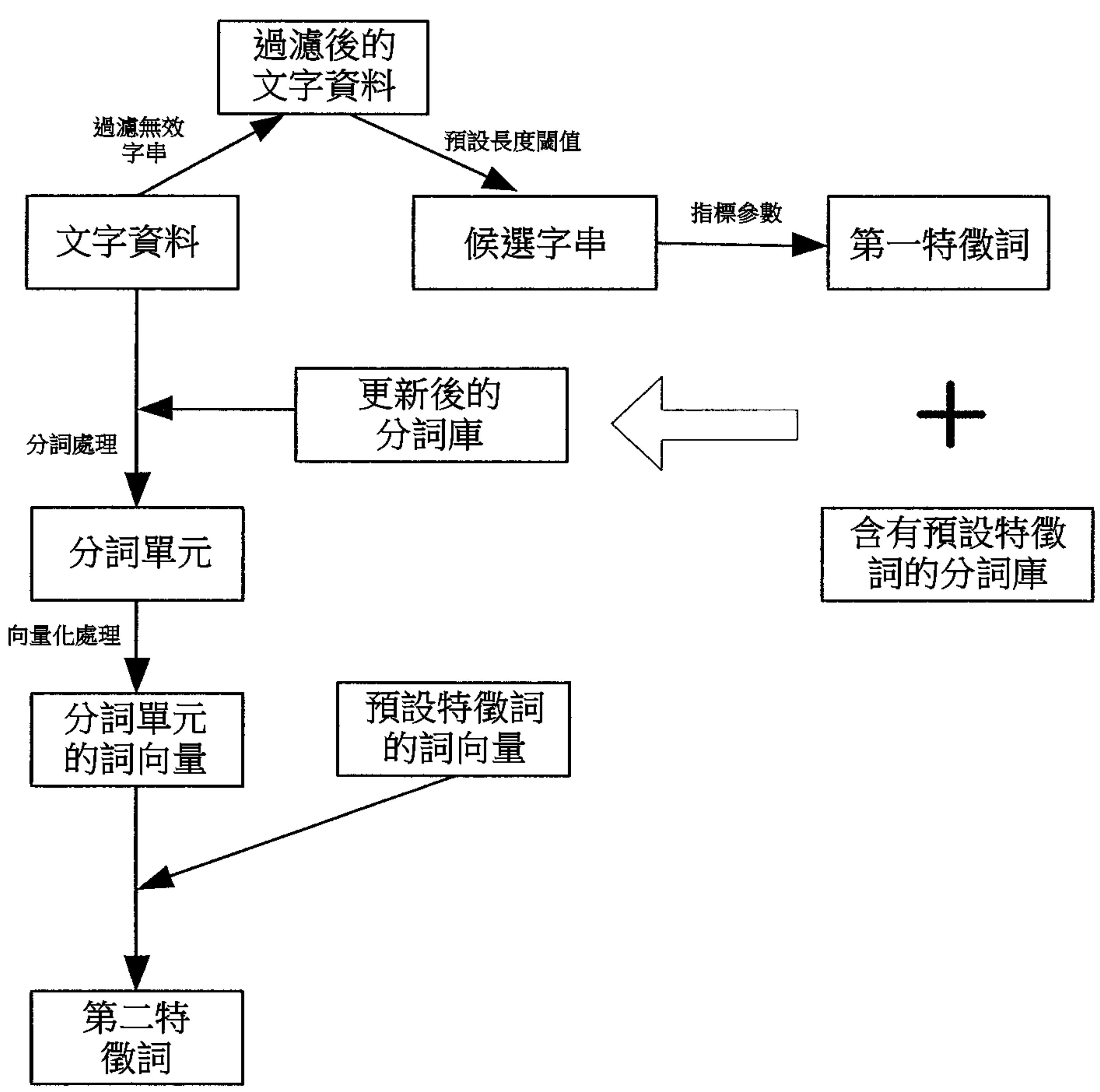
【發明圖式】



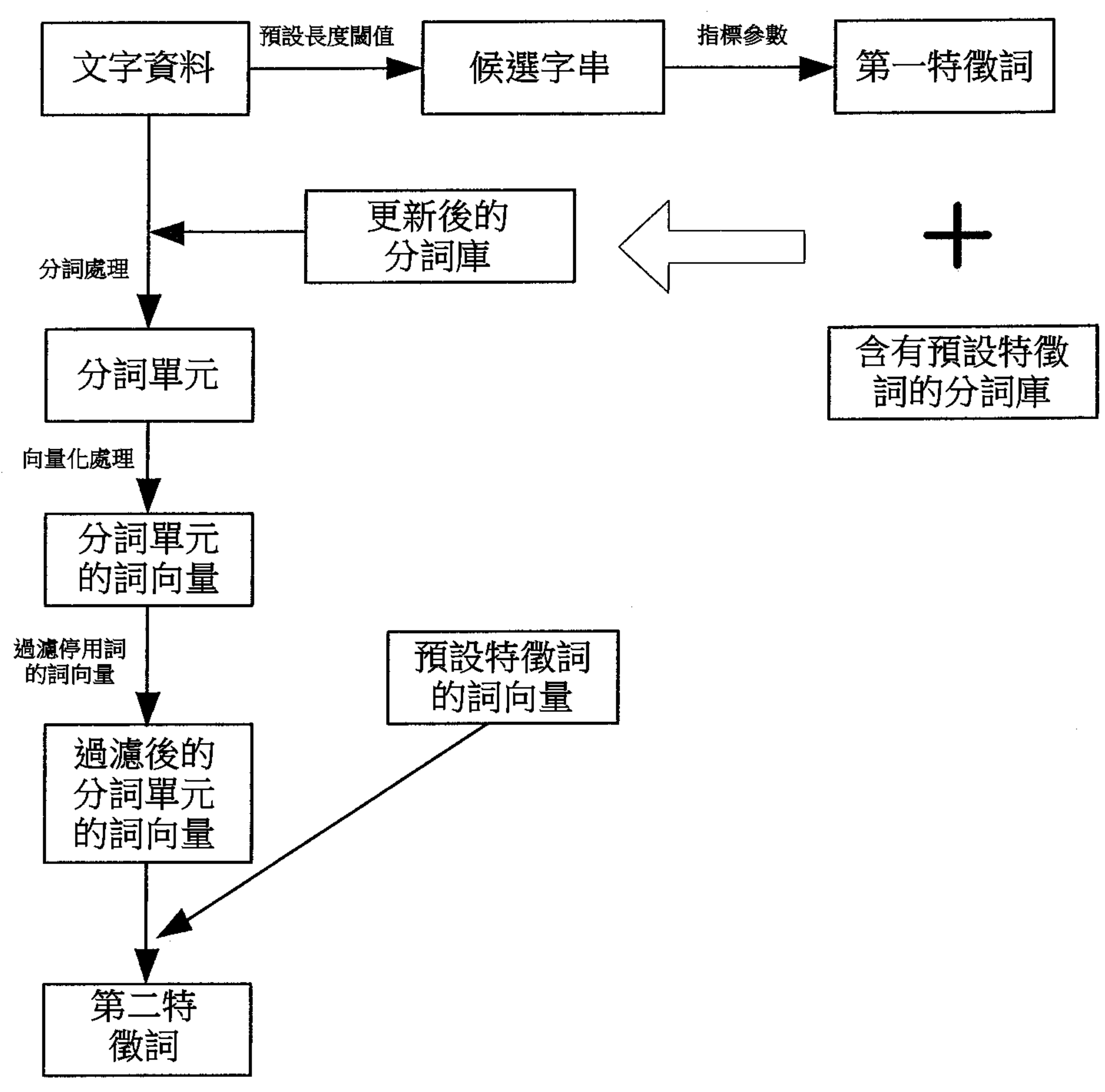
【圖 1】



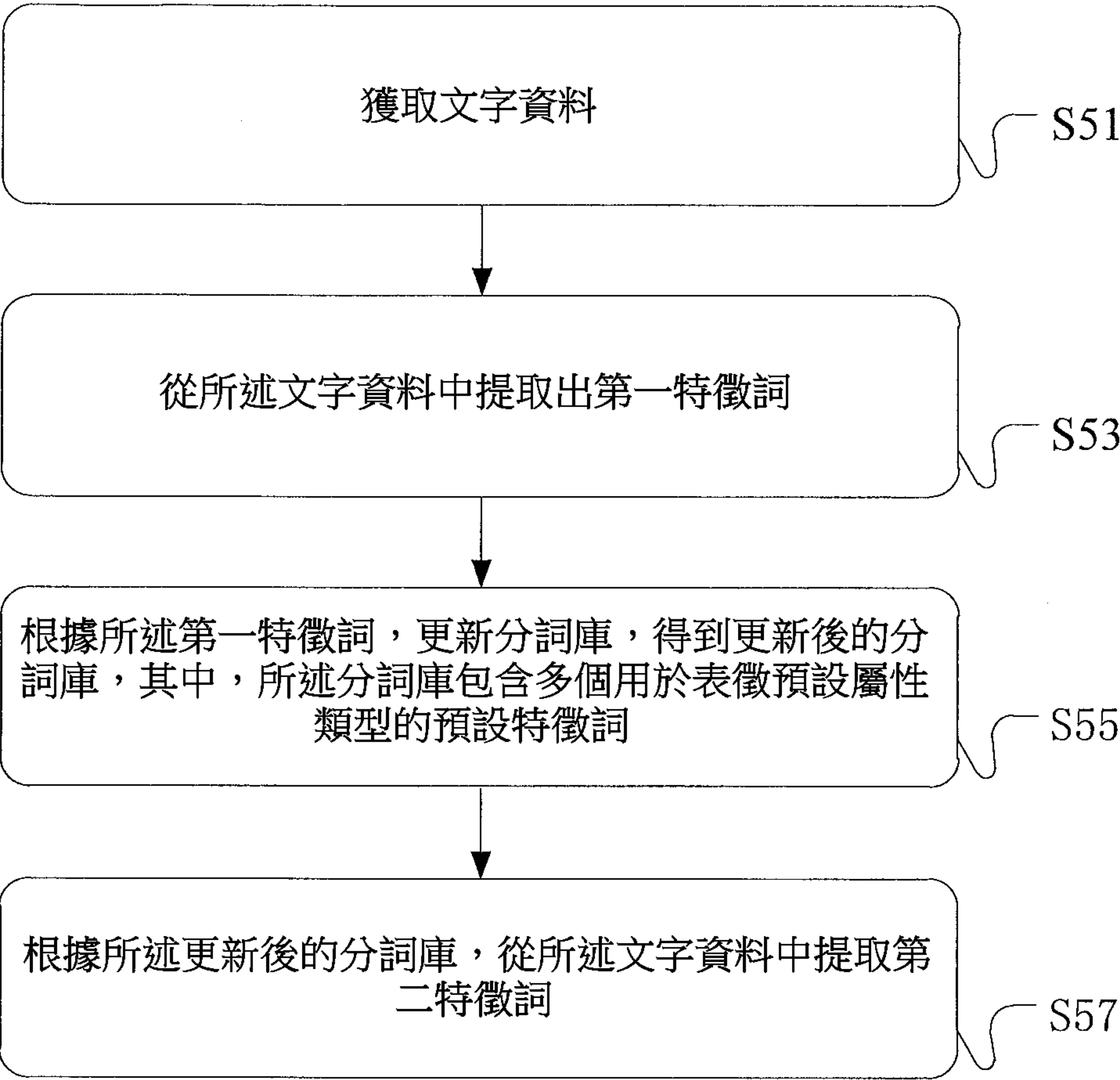
【圖 2】



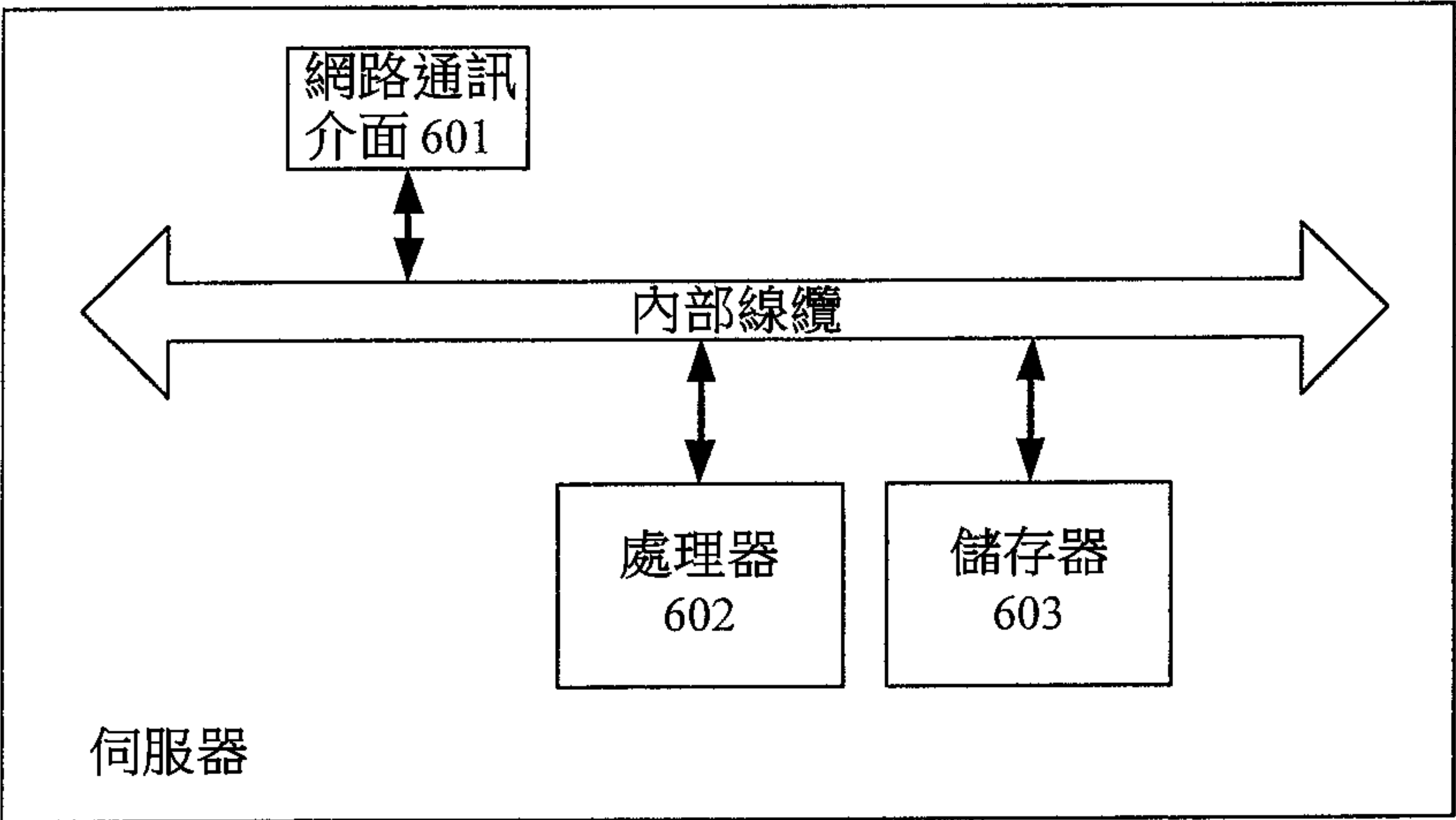
【圖 3】



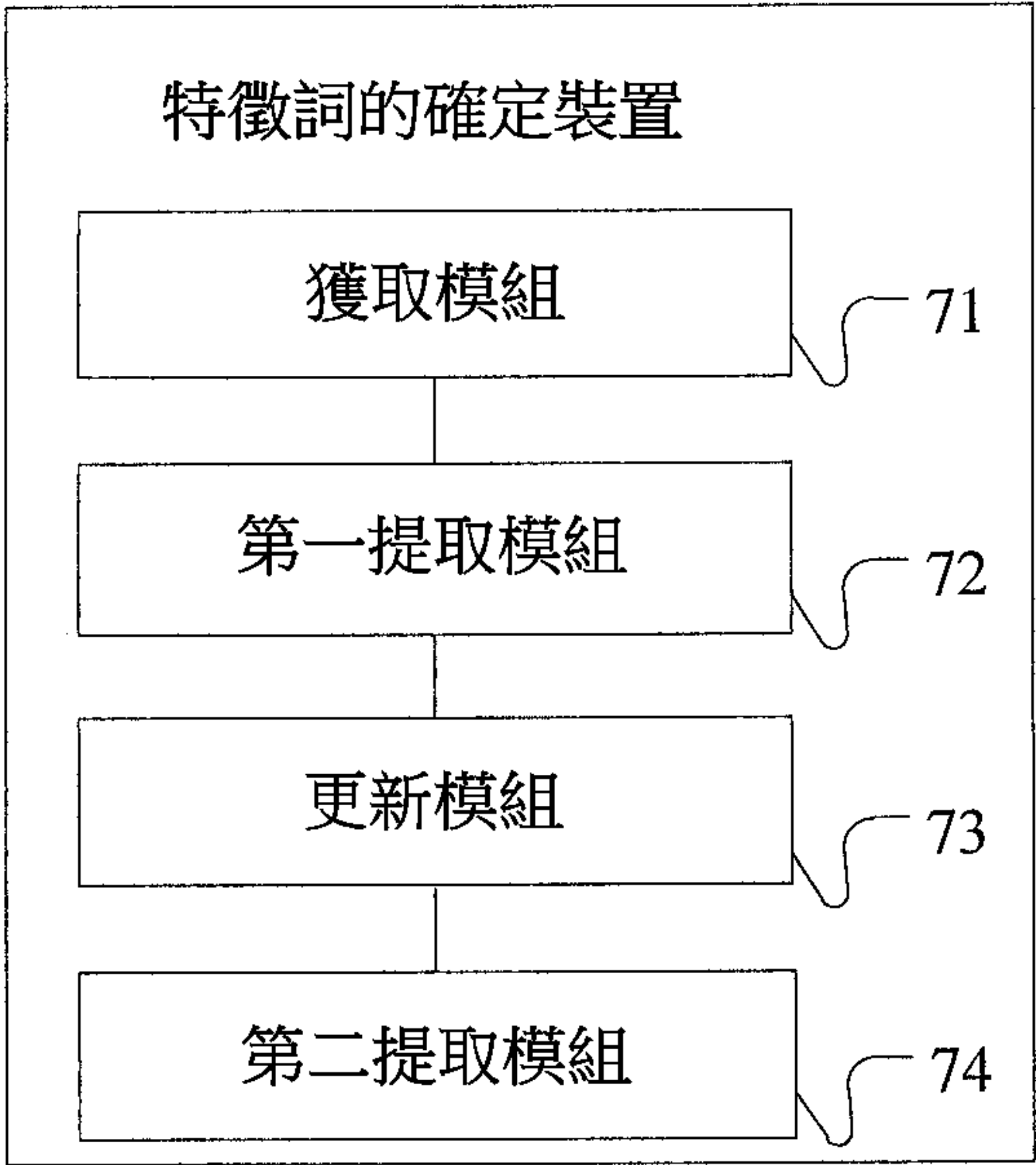
【圖 4】



【圖 5】



【圖 6】



【圖 7】