(19)中華民國智慧財產局
(12)發明說明書公開本 (11)公開編號：TW 201214167 A1
(43)公開日： 中華民國 101 (2012) 年 04 月 01 日

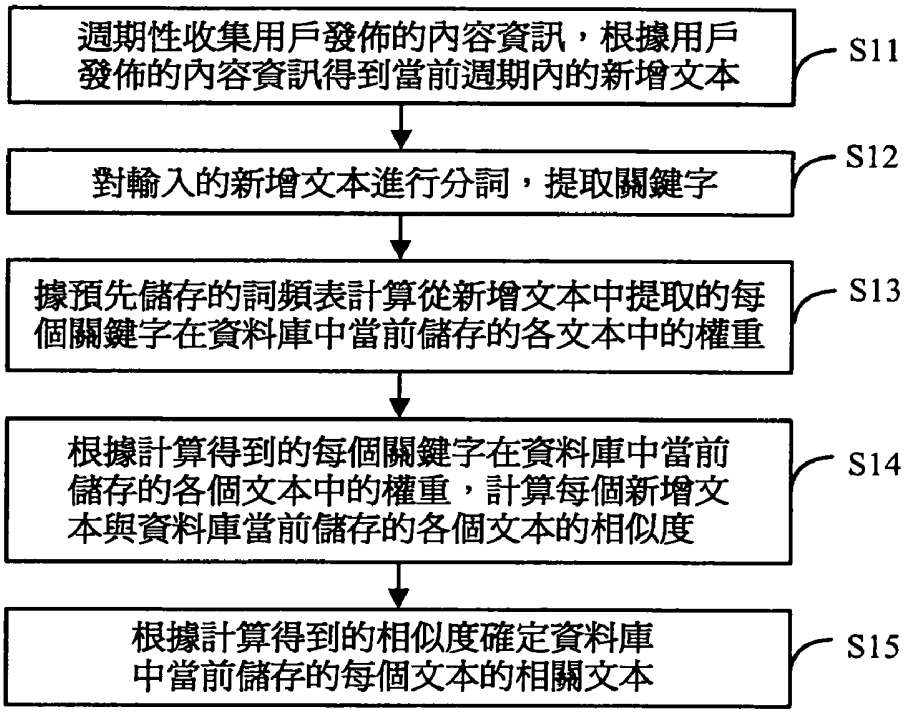
(21)申請案號：099140210 (22)申請日： 中華民國 99 (2010) 年 11 月 22 日
(51)Int. Cl.： *G06F17/30 (2006.01)* *G06F17/27 (2006.01)*
(30)優先權：2010/09/20 中國大陸 201010290693.4
(71)申請人：阿里巴巴集團控股有限公司 (開曼群島) ALIBABA GROUP HOLDING LIMITED (KY)
香港
(72)發明人：張旭 (CN) ；蘇寧軍 (CN) ；顧海傑 (CN) ；祁建程 (CN)
(74)代理人：林志剛
申請實體審查：無 申請專利範圍項數：15 項 圖式數：8 共 44 頁

(54)名稱

文本匹配方法及裝置

(57)摘要

本申請公開了一種文本匹配方法及裝置，該方法包括：根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；對輸入的新增文本進行分詞並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。通過建立和更新詞頻表的方式避免了現有技術中每次匹配都需要對所有文本進行計算的問題，減少了匹配運算工作量，提高了系統性能。



(19)中華民國智慧財產局
(12)發明說明書公開本 (11)公開編號：TW 201214167 A1
(43)公開日： 中華民國 101 (2012) 年 04 月 01 日

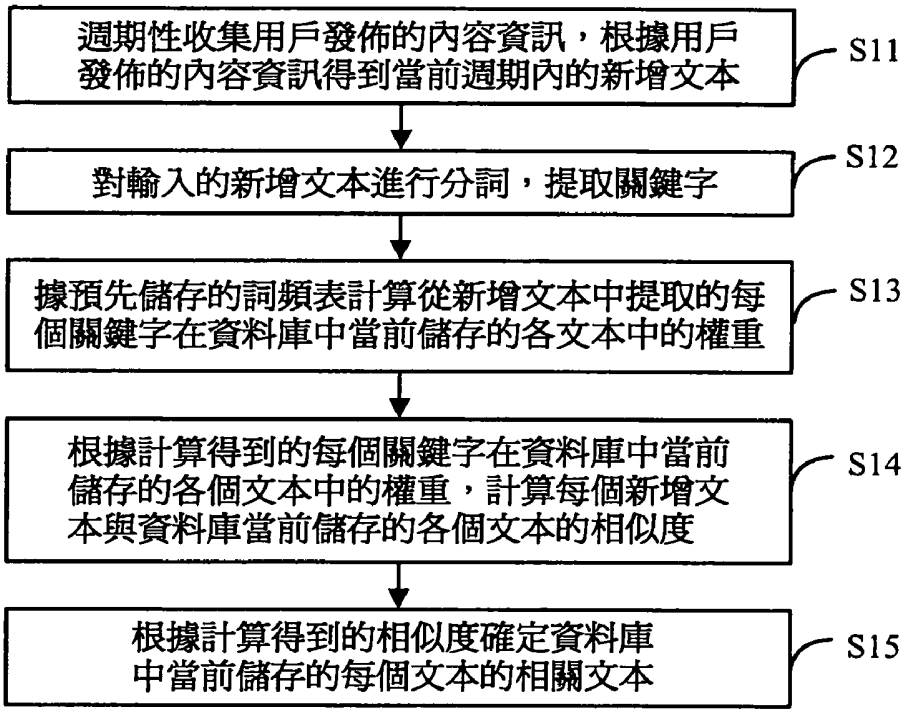
(21)申請案號：099140210 (22)申請日： 中華民國 99 (2010) 年 11 月 22 日
(51)Int. Cl.： *G06F17/30 (2006.01)* *G06F17/27 (2006.01)*
(30)優先權：2010/09/20 中國大陸 201010290693.4
(71)申請人：阿里巴巴集團控股有限公司 (開曼群島) ALIBABA GROUP HOLDING LIMITED (KY)
香港
(72)發明人：張旭 (CN) ；蘇寧軍 (CN) ；顧海傑 (CN) ；祁建程 (CN)
(74)代理人：林志剛
申請實體審查：無 申請專利範圍項數：15 項 圖式數：8 共 44 頁

(54)名稱

文本匹配方法及裝置

(57)摘要

本申請公開了一種文本匹配方法及裝置，該方法包括：根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；對輸入的新增文本進行分詞並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。通過建立和更新詞頻表的方式避免了現有技術中每次匹配都需要對所有文本進行計算的問題，減少了匹配運算工作量，提高了系統性能。



六、發明說明：

【發明所屬之技術領域】

本申請涉及資料處理領域，尤指一種大資料量的文本匹配方法及裝置。

【先前技術】

現有的文本比較，一般採用全量運算匹配的方式，當需要計算文本之間的相關程度的時候，需要針對獲取的所有文本進行計算，最終得到兩兩之間的相似度，這樣每計算一次相似度都要針對所有的文本資料進行計算，其計算量將是非常巨大的，其運行時間爲O（N^2）量級的，隨著文本數量N的增大，運算的時間也會很長。

這種大資料量的運算比較對設備的系統性能帶來了很大的影響，使系統的I/O通訊、資料儲存、資料的網路傳輸都面臨很大的壓力，導致設備的資料處理速度緩慢，甚至出現資料傳輸的阻塞或擁塞。

這種全量運算的文本匹配所存在的大資料運算量對系統性能的影響，隨著需要匹配的文本數量的增大，變的越來越嚴重。如何實現對大資料量匹配的高效處理，成爲亟待解決的難題。

由於現有技術中基本上都對基於內容的文本匹配進行全量資料運算，對於基於內容的文本匹配的優化，已有技術可以包括下列方式：

（1）針對單機版的基於內容的文本匹配，通過建索

引的方式提高文本匹配的速度和效率。

（2）針對分散式的基於內容的文本匹配，主要是增加硬體支援，比如增加並行度，執行並行運算。

但是無論是建立索引還是增加並行度都不能很好的解決文本匹配過程中，全量資料運算操作所存在的資料計算量大，運行時間長，需要對所有資料進行運算和一一比對，需要的儲存空間大等問題，因此，現有的文本匹配方式存在的資料處理速度慢、網路傳輸阻塞等系統性能瓶頸依然比較嚴重。

【發明內容】

本申請實施例提供一種文本匹配方法及裝置，用以解決現有技術中存在的文本匹配資料處理量大導致處理速度慢、影響系統性能、引起傳輸阻塞等問題。

一種文本匹配方法，包括：

週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；

對輸入的新增文本進行分詞，並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本；

根據計算得到的每個關鍵字在資料庫中的各文本中的

權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；

根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。

一種文本匹配裝置，包括：

收集模組，用於週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；

分詞模組，用於對輸入的新增文本進行分詞，並提取關鍵字；

權重確定模組，用於根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；

詞頻更新模組，用於根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本；

相似度確定模組，用於根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；

文本比較模組，用於根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。

本申請有益效果如下：

本申請實施例提供的文本匹配方法及裝置，通過週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容

資訊得到當前週期內的新增文本並儲存到資料庫中；對輸入的新增文本進行分詞，並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本；根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。上述方法通過建立和更新詞頻表的方式避免了現有技術中任意兩個文本的匹配都需要對所有文本進行計算的問題，具體爲關鍵字的權重不再依賴於全局資料運算得到總體變數，而依靠詞頻表即可實現，從而減少了匹配運算工作量，提高了系統性能；且通過使用詞頻表可以僅計算部分文本之間的相似度或計算全部文本之間的相似度，因此即使只針對更新後的新增文本進行計算，也能獲取到準確的匹配運算結果。該方式適用於所有文本的匹配，具有很強的通用性和普遍適用性，其匹配過程實現簡單，很好的解決網路系統瓶頸問題。

【實施方式】

本申請實施例提供的文本匹配方法，週期性的獲取新增文本，並將獲取到的新增文本加入資料庫中；預先建立詞頻表，並根據獲取的新增文本或根據資料庫中增加新增

文本之後的所有文本更新詞頻表，從而可以根據詞頻表方便的計算任意兩個文本（包括新增文本和原始文本）之間的相似度。在本申請中根據需要可以計算資料庫中任意兩個文本之間的相似度、也可以只計算新增文本與新增文本以及新增文本與原始文本之間的相似度。

下面通過具體的實施例分別說明這兩種情況的實現流程。其中，資料庫中儲存的原始文本是指當前週期之前儲存的文本，即上一個週期存入新增文本之後資料庫中的所有文本。

本申請實現文本匹配的系統架構如圖1所示，該系統包括伺服器和若干用戶端，伺服器通過週期性收集用戶端的操作行爲，獲取新增文本，實現對文本的匹配。用戶端和伺服器的具體功能，在下面的實施例中進行詳細介紹。

例如：伺服器可以對用戶通過用戶端發佈的商品資訊進行匹配，確定與用戶發佈的商品資訊具有相關性的商品資訊，從而實現在其他用戶瀏覽到用戶發佈的商品時，能夠爲用戶顯示和推薦類似的或相關的商品。當然本申請的文本匹配方法不限於商品資訊的匹配，只要是基於文本的文本匹配都可以通過本申請的方法實現。

下面通過具體的實施例說明本申請文本匹配的實現過程。

實施例一：

本申請實施例一提供的文本匹配方法，針對每個週期

的每個新增文本，計算每個新增文本與每個原始文本之間、以及任意兩個新增文本之間的相似度。即確定與新增文本相關的相似度數據。例如：在商品推薦過程中使用時，則是根據當前週期內發佈的商品資訊獲取新增文本。並根據新增文本確定與當前週期內發佈的商品資訊相匹配的所有商品（資訊包括此前發佈的商品資訊和當前週期內發佈的商品資訊）。

本申請實施例一提供的文本匹配方法的流程如圖2所示，執行步驟如下：

步驟S11：週期性收集用戶發佈的內容資訊，根據用戶發佈的內容資訊得到當前週期內的新增文本。

收集用戶發佈的內容資訊的週期可以根據需要設定。根據收集到的各個用戶在當前週期內發佈的內容資訊，可以生成相關的文本，即爲當前週期的新增文本。收集到新增文本後將其儲存至資料庫中，則資料庫中當前儲存有上個週期就已經儲存的原始文本和當前週期內存入的新增文本。

例如：用戶通過用戶端發佈商品資訊，伺服器週期性的獲取各個用戶端發佈的商品資訊，其中設定的週期可以是一天、一星期或幾個小時等。

優選的，在收集到用戶發佈的內容資訊後，根據設定的輸入過濾規則，對收集到的用戶發佈的內容資訊進行過濾。

對收集到的用戶發佈的內容資訊進行過濾可以根據內

容資訊的品質是否符合設定的品質評估閾值，發佈內容資訊的用戶是否是設定的合格用戶等設置的過濾規則中的一個或多個，對收集到的用戶發佈的內容資訊進行過濾。或者根據其他設置的輸入過濾規則，對收集到的用戶發佈的內容資訊進行過濾。在對收集到的用戶發佈的內容資訊進行過濾後，根據過濾後內容資訊生成當前週期內的新增文本。

仍以商品資訊的匹配爲例，在獲取到用戶端發佈的商品資訊時，對商品資訊進行過濾，例如：過濾掉沒有提供圖片或沒有其他設定的必要資訊的商品。

上述通過對收集到的內容資訊進行過濾，得到新增文本，可以提高收集得到的用戶發佈的內容資訊的可用性，提高了用於匹配的新增文本的品質，從而可以獲得更佳的匹配結果；同時也進一步減少匹配過程的計算量，提高了匹配速度。

仍以商品資訊的匹配爲例，在獲取到用戶端在當前週期內發佈的商品資訊後可以得到當前週期內的新增文本。例如：發佈的一個MP3的商品資訊包括：名稱MP3、顏色紅色、型號XX以及功能描述等相關資訊，則根據用戶發佈的商品資訊，得到一個新增文本。

步驟S12：對輸入的新增文本進行分詞，提取關鍵字。

即針對輸入的每個新增文本，將文本內容劃分爲若干詞語，並提取用於文本匹配的若干關鍵字，提取得到的若

干關鍵字可以生成一個分詞向量。

例如：發佈的一個MP3的商品資訊包括：名稱MP3、顏色紅色、型號XX和功能描述等資訊，則將得到的文本分詞後，可以從中提取出MP3、紅色等關鍵字，這些關鍵字可以組成一個分詞向量。

步驟S13：根據預先儲存的詞頻表計算從新增文本中提取的每個關鍵字在資料庫中當前儲存的各文本中的權重。

該步驟具體計算每個關鍵字在資料庫中儲存的每個文本（包括當前週期的新增文本和上一個週期儲存的原始文本）中的權重，具體可以通過查詢詞頻表中每個關鍵字在文本中的出現頻率，實現計算關鍵字在該文本中的權重。

其中，詞頻表根據各個詞語在資料庫中儲存的每個文本中的出現頻率週期性更新。這裏的各個詞語是指所有詞頻表中詞語，針對這些詞語預計算出來的詞頻，而不僅僅包含當前輸入的新增文本分詞後劃分出的關鍵字的詞頻。

詞頻表在建立時，針對資料庫中已儲存的所有文本進行統計，得到每個詞語在各個文本中出現次數的詞頻表，在後續可以通過更新的方式來添加和減少更新後的結果。每個收集週期，詞頻表都可以根據各個關鍵字在資料庫中的當前儲存的各文本中的出現頻率週期性更新，具體包括兩種情況：

情況一：根據資料庫中的當前儲存的所有文本直接更新詞頻表。

每次輸入新增文本後，統計各個詞語在輸入的新增文本和資料庫中儲存的原始文本中的出現頻率，得到包含各個詞語在資料庫中當前儲存的每個文本中的出現頻率的詞頻表。由於計算詞頻的運算量是與輸入資料量成線性關係的，因此，即使採用對資料庫中儲存的所有文本進行統計來更新詞頻表，其運算量也不會很大，時間也不長。

情況二：根據新增文本和原來詞頻表中儲存的內容更新詞頻表。

每次輸入新增文本後，統計各個詞語在輸入的每個新增文本中的出現頻率，根據統計得到的結果與詞頻表中儲存的各個詞語在資料庫中儲存的原始文本中的出現頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的詞頻表。具體實施例中，若預先儲存的詞頻表中未記錄新增文本分詞後得到的各詞語的詞頻，則以情況一該方案更新詞頻表。若預先儲存的詞頻表中已記錄新增文本分詞後得到的各詞語在原始文本中的詞頻，則以情況二該方案更新詞頻表。

上述根據預先儲存的詞頻表計算分詞提取的每個關鍵字在資料庫中的當前儲存的各個文本中的權重，具體包括：

根據詞頻表，分別確定選定關鍵字在資料庫中當前儲存的每個文本中的出現次數。以及

確定資料庫中當前儲存的的所有文本與包含有選定關鍵字的文本的數量比。

根據選定關鍵字在每個文本中的出現次數和上述計算得到的數量比，分別計算每個關鍵字在每個文本中的權重。

步驟S14：根據計算得到的每個關鍵字在資料庫中當前儲存的各個文本中的權重，計算每個新增文本與資料庫當前儲存的各個文本的相似度。

計算每個新增文本與資料庫中當前儲存的各個文本的相似度，包括：計算輸入的任意兩個新增文本之間的相似度、以及計算每個新增文本和資料庫中儲存的每個原始文本的相似度。

計算每個新增文本與資料庫中當前儲存的各文本的相似度，具體包括：

將待計算相似度的文本中的每個關鍵字的權重組成權重向量。權重向量由上述計算出的各個關鍵字在該文本中的權重組成。

針對每個新增文本，分別計算該新增文本的權重向量與資料庫中當前儲存的各文本的權重向量的內積，得到該新增文本與資料庫中當前儲存的各文本的相似度。

由於資料庫中的原始文本之間的相似度在上一次輸入上一個週期的新增文本時已經計算過，因此，本次只計算新輸入的新增文本之間、以及新輸入的新增文本與資料庫中的原始文本之間的相似度，從而大大減少了運算量。

步驟S15：根據計算得到的相似度確定資料庫中當前儲存的每個文本的相關文本。

上述計算獲取到的每個新增文本和資料庫中當前儲存的各個文本之間的相似度之後，根據具體需求，既可以確定與每個新增文本具有一定相關性的相關文本，也可以確定與資料庫中當前儲存的每個文本具有一定相關性的相關文本了。其中，與每個新增文本相關的文本可以是新獲取到的其他新增文本也可以是儲存的原始文本。與資料庫中當前儲存的每個文本相關的文本可以是新獲取到的新增文本也可以是儲存的原始文本。其中原始文本與原始文本之間的相似度在之前的週期內已經確定並儲存在資料庫中。也就是說在本實施例中，在確定相關文本時，涉及到資料庫中原始文本和原始文本之間的相似度時，直接使用上一次儲存的相似度。

其中，與每個文本具有一定相關性的相關文本的確定，具體包括下列兩種確定方式：

方式一：通過設定閾值確定符合設定條件的相關文本。

針對待確定相關文本的新增文本或資料庫中當前儲存的文本，確定與該新增文本或資料庫中當前儲存的文本的相似度大於或大於等於設定閾值的至少一個文本爲該新增文本或資料庫中當前儲存的文本的相關文本。

方式二：通過排序獲取設定數量的相關文本。

針對待確定相關文本的新增文本或資料庫中當前儲存的文本，根據資料庫中資料庫中當前儲存的每個文本與待確定相關文本的新增文本或資料庫中當前儲存的文本的相

似度大小排序，確定相似度較高的設定數量的文本作爲待確定相關文本的新增文本或資料庫中當前儲存的文本的相關文本。

在確定了新增文本或資料庫中當前儲存的文本得相關文本之後，儲存在資料庫中，用作後續的商品推薦或其他過程中使用。以用於商品推薦爲例：

在獲取到包括用戶的點擊行爲、瀏覽行爲、用戶購買行爲、收藏網頁上展示的商品等等用戶操作行爲時，根據用戶操作行爲涉及的商品所對應的文本，從資料庫中獲取該文本的相關文本，將獲取到的相關文本對應的商品推薦給用戶。其中，涉及的商品所對應的文本和該文本的相關文本，根據商品的發佈時間不同，可能是新增文本也可能是原始文本。

實施例二：

本申請實施例二提供的文本匹配方法，針對每個週期輸入新增文本後資料中儲存的每個文本，計算任意兩個文本之間的相似度，其流程如圖3所示，執行步驟如下：

步驟S21：週期性收集用戶發佈的內容資訊，根據用戶發佈的內容資訊得到當前週期內的新增文本。

同步驟S11，此處不再贅述。

步驟S22：對輸入的新增文本進行分詞，提取關鍵字。

同步驟S12，此處不再贅述。

步驟S23：根據預先儲存的詞頻表計算從新增文本中提取的每個關鍵字在資料庫中的當前儲存的各文本中的權重。

同步驟S13，此處不再贅述。

步驟S24：根據計算得到的每個關鍵字在資料庫中當前儲存的各文本中的權重，計算資料庫中任意兩個文本的相似度。

計算資料庫中任意兩個文本的相似度，包括：計算輸入的任意兩個新增文本之間的相似度、計算每個新增文本和資料庫中儲存的每個原始文本的相似度、以及計算任意兩個原始文本之間的相似度。計算任意兩個文本的相似度，具體包括：

將待計算相似度的文本中的每個關鍵字的權重組成權重向量。

針對每個文本，分別計算該文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該文本與資料庫中儲存的各文本的相似度。

該方式在詞頻更新之後重新計算每個文本之間的相似度，從而能夠獲取到準確的相似度值，使後續比較匹配的結果更準確。

步驟S25：根據計算得到的相似度確定資料庫中當前儲存的每個文本的相關文本。

該步驟確定相關文本時，和步驟S15類似的也包含兩種方式。所不同的是在本實施例中，在確定相關文本時，

涉及到資料庫中原始文本和原始文本之間的相似度時，也是用本次計算得到的相似度。

確定相關文本後在商品推薦過程中的應用也與步驟S15類似。

實施例三：

本申請實施例三提供的文本匹配方法，針對實施例一和實施例二的方案進行改進，增加輸出過濾的過程。具體包括：

在實施例一的步驟S14計算相似度之後和步驟S15確定相關文本之前增加輸出過濾的步驟，在實施例二的步驟S24計算相似度之後和步驟S25確定相關文本之前增加輸出過濾的過程，其流程如圖4所示，執行步驟如下：

步驟S31：獲取計算得到的每個新增文本與資料庫中當前儲存的各個文本的相似度，或計算得到的資料庫中任意兩個文本的相似度。

針對兩個文本的相似度的過濾，可以根據後續相關文本確定的不同要求，對不同文本的相似度進行過濾，因此，針對實施例一計算新增文本和資料庫中當前儲存的各個文本之間的相似度時，獲取的是計算得到的每個新增文本與資料庫中的資料庫中當前儲存的每個文本的相似度。針對實施例二計算任意兩個文本之間的相似度時，獲取的是計算得到的資料庫中任意兩個文本的相似度。

步驟S32：根據設定的輸出過濾規則，對資料庫中當

前儲存的待確定相關文本的每個文本相關的相似度數據進行過濾。

對待確定相關文本的每個文本相關的相似度數據進行過濾，去除不符合設定條件的文本資料時，可以根據相似度的大小，去除與待確定相關文本的每個文本相似度小於設定閾值的文本；也可以根據相似度的大小排序，去除與待確定相關文本的每個文本相似度較低的設定數量的文本。當然也可以設置其他的輸出過濾規則對輸出文本進行過濾。

通過對待確定相關文本的每個文本相關的相似度數據進行過濾，減少匹配過程中需要匹配的文本的數量，從而進一步了提高匹配速度和效率。

實施例四：

本申請實施例四提供的文本匹配方法，具體提供實現文本匹配的一個具體實現示例，其實現原理如圖5所示，其流程如圖6所示，執行步驟如下：

步驟S41：週期性在資料層採集用戶發佈的內容資訊。

其中，用戶發佈的內容資訊的採集是在資料層完成的。資料表中的資料在資料層進行更新，更新根據設定的週期進行。

資料層是資料的提供層和儲存層，爲資料的應用層提供資料，最終用於前臺展現。同時，資料層爲底層的演算

法層提供輸入資料，也接受演算法層的運算結果。這一層包括資料庫和一些儲存檔。

例如，將採集到的用戶發佈的商品資訊中的商品名稱作爲文本資料，下面的匹配對比是基於得到的文本資料的內容進行的。例如：採集到發佈的商品資訊爲MP3，則找到包含MP3的其他文本作爲匹配文本。

步驟S42：對採集到的用戶發佈的內容資訊進行過濾。

在過濾層進行用戶發佈的內容資訊的過濾，根據設定輸入過濾規則，對採集到的用戶發佈的內容資訊進行過濾。也就是說由過濾層對演算法層的輸入和輸出做過濾處理，該步驟的輸入過濾涉及到的是對演算法層輸入的過濾，過濾後提供給演算法層。後續步驟中的輸出過濾涉及到的是對演算法層的計算結果進行過濾，提供給資料層。

其中，設定的過濾規則包括實施例一中所描述的：內容資訊的品質是否符合設定的品質評估閾值，發佈內容資訊的用戶是否是設定合格用戶等等。

例如：過濾去掉資料品質低的內容資訊。即將內容資訊品質低於設定的品質評估閾值的內容資訊去除。從而避免在文本匹配中，有的文本來源於低品質的商品資訊，這類商品資訊，通常品質評分值比較低，比如沒有提供圖片，或其他必要的資訊，這類商品被推薦和點擊的意義不大。因此，這類商品資訊一般品質評分值低於設定的品質評估閾值，在進行文本匹配運算之前就會被過濾剔除掉。

又例如：過濾掉不合格用戶的內容資訊，不合格用戶包括網路爬蟲，機器人，和不合格的物理用戶等等。

可以通過判斷發佈內容資訊的用戶的訪問次數是否超過設定的訪問閾值，例如網路爬蟲，機器人，他們的行爲有明顯的特徵，他們通常在一段時間內異常活躍，他們提供的資料，可視爲噪音，予以剔除。此時可以設定一個訪問閾值，當訪問次數大於該閾值認爲是網路爬蟲或機器人。

也可以通過判斷用戶的信用值、有效期限等來判斷是否是合格的用戶。從而去除包括低信用的用戶，過期的用戶，還有不活躍的用戶（一般指設定時間範圍內沒有操作行爲的用戶，如最近的一個月沒有登錄，一個月沒有行爲資料等），這些不合格的用戶發佈的內容資訊可視爲無效資訊，予以剔除。

輸入過濾的目的是在系統採集到待輸入的文本資料後，對輸入的文本資料的過濾處理，過濾掉噪音，不合格用戶資料和低質量數據等，使輸入的文本資料減少。

步驟S43：根據過濾後的內容資訊得到當前週期的新增文本。

在對收集到的用戶發佈的內容資訊進行過濾後，根據過濾後內容資訊生成當前週期內的新增文本，從而提高了新增文本的品質。

步驟S44：根據過濾後輸入的新增文本進行相似度計算。

過濾後的新增文本會被輸入到演算法層，用於相似度的運算，以及更新詞頻表。

其中，更新詞頻表的原理如圖7所示。

當新增文本輸入後，演算法層擁有包含此前各週期內輸入的原始文本和當前週期輸入的新增文本在內的資料庫中當前儲存的所有文本。此時可以直接根據資料庫中當前儲存的所有文本更新詞頻表，也可以根據資料庫中當前儲存的所有文本與原始文本對比得到的新增文本，獲取新增的資料檔案來更新詞頻表。

新增文本與資料庫中儲存的各文本之間的相似度計算，以及資料庫中當前儲存任意兩個文本之間的相似度計算過程分別參見實施例一和實施例二的描述。

其中，根據預先儲存的詞頻表計算分詞提取的每個關鍵字在資料庫中的各文本中的權重的過程具體包括：

首先，確定選定關鍵字在資料庫中每個文本中的出現次數。即針對每個文本，分別確定選定的關鍵字的出現次數。

具體可以通過詞頻表的到，詞頻表中詞語出現次數可以通過詞頻-反向文檔頻率（term frequency–inverse document frequency，TF-IDF），即第i個關鍵字在第j個文本中出現的次數可以通過下列公式計算得到：

$$TF_{i,j} = \frac{f_{i,j}}{\max f_{z,j}}$$

其中，$f_{i,j}$是第i個關鍵字k_i在第j個文本d_j中出現的次數，$\max f_{z,j}$表示$f_{i,j}$中的最大值，i，j爲正整數。詞頻表根據該公式更新，而使用過程中需要確定時可以直接查詢詞頻表。

在使用上述公式時，可以根據實際情況對$f_{i,j}$和$\max f_{z,j}$的值進行限定。例如：可以設置$f_{i,j}$和$\max f_{z,j}$的值爲1，來表示將文本中多次出現的同一個關鍵字視爲出現了一次。

其次，確定資料庫中的儲存的所有文本與包含有選定關鍵字的文本的數量比。具體通過下列公式確定：

$$IDF_i = \log \frac{N}{n_i}$$

其中，N是資料庫中所有文本的個數，n_i表示出現了第i個關鍵字k_i的文本數量。

上述確定詞頻和確定數量比的過程順序不分先後，也可以同時執行。

然後，根據選定關鍵字在每個文本中的出現次數和上述計算得到的數量比，分別計算每個關鍵字在每個文本中的權重。如關鍵字k_i在文本d_j中的權重定義爲：

$$w_{i,j} = TF_{i,j} \times IDF_j$$

上述得到每個關鍵字在每個文本中的權重後，就可以構建權重向量，計算任意兩個文本的相似度了。

例如：針對文本d_j構建的包含關鍵字i=1、2、……、k

的權重向量爲：

$$W(d_j) = (w_{1j}, \ldots, w_{ij}, \ldots, w_{kj})$$

通過下列向量內積公式計算文本d_j和文本d_m得到相似度：

$$u(d_j, d_m) = \cos(\overrightarrow{W(d_j)}\,\overrightarrow{W(d_m)}) = \frac{\overrightarrow{W(d_j)} \bullet \overrightarrow{W(d_m)}}{\|W(d_j)\|_2 \times \|W(d_m)\|_2} = \frac{\sum_{i=1}^{K} w_{i,j} w_{i,m}}{\sqrt{\sum_{i=1}^{k} w_{i,j}^2} \sqrt{\sum_{i=1}^{k} w_{i,m}^2}}$$

步驟S45：對輸出文本之間的相似度數據進行輸出過濾。

對輸出資料的過濾參照實施例三的描述，其主要目的是過濾掉相似度比較低（例如相似度對比分數低）的結果或相似度排名靠後的若干文本資料。

例如，將一個待匹配的文本稱爲左列文本（即Left Offer），與之匹配的文本稱爲右列文本（Right Offer）。Left Offer和Right Offer是成對比較的結果的表示，也可以說每對比較，第一個文本稱爲Left Offer，第二個文本稱爲Right Offer。

那麼針對一個待匹配的Left Offer，過濾掉Right Offer排名靠後的、相似度比較低的若干文本。

輸出過濾是在計算相似度後先進行一次過濾，以便減少後續輸出相關文本時，所需要選擇的文本數量。

對文本的過濾可以在過濾層實現，可選的也可以在演算法層實現。

步驟S46：根據過濾後的文本之間的相似度數據輸出資料庫中當前儲存的各個文本的相關文本。

關於匹配文本的確定過程參見上述實施例中的描述。在獲取相關文本後，則可以實現對每個Left Offer，只輸出相似度最高的幾個（top N，根據不同的規則可配置）Right Offer。

當需要進行商品推薦時，將用戶操作行爲涉及的商品對應的文本作爲Left Offer，查找資料庫中儲存的該Left Offer對應的Right Offer，將查找到的Right Offer對應的商品推薦給用戶。

實施例五：

本申請實施例五根據本申請上述實施例提供的上述文本匹配方法，構建一種文本匹配裝置，該裝置可以設置在網路設備，例如上述的伺服器中，用於文本的匹配。該裝置的結構如圖8所示，包括：收集模組10、分詞模組20、權重確定模組30、詞頻更新模組40、相似度確定模組50和文本比較模組60。

收集模組10，用於週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中。

分詞模組20，用於對輸入的新增文本進行分詞，並提取關鍵字。

權重確定模組30，用於根據預先儲存的詞頻表計算提

取的每個關鍵字在資料庫中的各文本中的權重。

優選的，上述權重確定模組30，具體包括：第一確定單元301、第二確定單元302和權重計算單元303。

第一確定單元301，用於根據詞頻表，分別確定選定關鍵字在資料庫中每個文本中的出現次數。

第二確定單元302，用於確定資料庫中儲存的文本與包含有選定關鍵字的文本的數量比。

權重計算單元303，用於根據選定關鍵字在每個文本中的出現次數和第二確定單元302確定出來的數量比，分別計算每個關鍵字在每個文本中的權重。

詞頻更新模組40，用於根據各個詞語在資料庫中的各文本中的出現頻率週期性更新詞頻表；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本。

優選的，上述詞頻更新模組40，具體用於：每次輸入新增文本後，統計各個詞語在輸入的新增文本和資料庫中儲存的原始文本中的出現的頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表；或每次輸入新增文本後，統計各個詞語在輸入的每個新增文本中的出現的頻率，根據統計得到的結果與詞頻表中儲存的各個詞語在資料庫中的儲存的原始文本中的出現頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表。

相似度確定模組50，用於根根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資

料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度。

優選的，上述相似度確定模組50，具體包括：向量生成單元501和相似度計算單元502。

向量生成單元501，用於將待計算相似度的文本中的每個關鍵字的權重組成權重向量。

相似度計算單元502，用於針對每個新增文本，分別計算該新增文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該新增文本與資料庫中儲存的各文本的相似度；或針對資料庫中儲存的每個文本，分別計算該文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該文本與資料庫中儲存的各文本的相似度。

文本比較模組60，用於根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。

優選的，上述文本比較模組60，具體用於：針對待確定相關文本的每個文本，確定與該文本的相似度大於或大於等於設定閾值的至少一個資料庫中儲存的文本的相關文本；或針對待確定相關文本的每個文本，根據資料庫中各文本與待確定相關文本的文本的相似度大小排序，確定相似度較高的設定數量的資料庫中儲存的文本作爲待確定相關文本的文本的相關文本。

優選的，上述文本匹配裝置，還包括：輸入過濾模組70，用於根據設定的輸入過濾規則，對當前週期內收集到用戶發佈的內容資訊進行過濾，根據過濾後內容資訊得到

當前週期內的新增文本，輸入給分詞模組20。

輸入過濾單元70，具體用於根據內容資訊的品質是否符合設定的品質評估閾值和/或發佈內容資訊的用戶是否是設定的合格用戶，對該收集到的內容資訊進行過濾。

優選的，上述文本匹配裝置，還包括：輸出過濾模組80，用於根據相似度確定模組50計算得到的每個新增文本與資料庫中的每個文本的相似度，或計算得到的資料庫中任意兩個文本的相似度；對待確定相關文本的新增文本或資料庫中儲存的文本相關的相似度數據進行過濾，去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度小於設定閾值的文本，或去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度較低的設定數量的文本，提供給文本比較模組60。文本比較模組60再根據過濾後的文本確定新增文本或資料庫中儲存的各文本的相關文本。

本申請實施例提供的上述文本匹配方法及裝置，可以通過軟體實現，也可以通過硬體實現。例如使用C語言、linux作業系統，應用分散式集群，比如簇（cluster），或Hadoop（一種分散式系統架構）集群等硬體實現。上述方式在各種文本的匹配過程中均可使用，例如可應用在用於電子交易的資源（sourcing）平臺中對商品相關的文本資料進行匹配，以便爲用戶提供關聯商品。

本申請實施例提供的上述文本匹配方法及裝置，通過建立和更新詞頻表的方式避免了現有技術中任意兩個文本的匹配都需要對所有文本進行計算的問題，具體爲關鍵字

的權重不再依賴與全局資料運算得到總體變數，而依靠詞頻表即可實現，從而減少了匹配運算工作量，提高了系統性能。

且通過使用詞頻表可以僅計算部分文本之間的相似度或計算全部文本之間的相似度，因此即使只針對更新後的新增文本進行計算，也能獲取到準確的匹配運算結果，而只計算更新的部分使得運行時間大大縮短，實現了大資料量文本匹配計算過程中增量演算法實現過程。

該方式適用於所有文本的匹配，具有很強的通用性和普遍適用性，其匹配過程實現簡單，且資料傳輸和採集也可以只針對更新部分，很好的解決網路系統瓶頸問題。

上述方法，在輸入資料之前進行輸入匹配，在匹配運算之後進行輸出匹配，從而進一步減少了匹配運算的處理資料量。上述方法採用層次化、模組化的結構，達到了可擴展，易於維護的目的。

顯然，本領域的技術人員可以對本申請進行各種改動和變型而不脫離本申請的精神和範圍。這樣，倘若本申請的這些修改和變型屬於本申請之申請專利範圍及其等同技術的範圍之內，則本申請也意圖包含這些改動和變型在內。

【圖式簡單說明】

圖1爲本申請實施例一中文本匹配系統的結構示意圖；

圖2爲本申請實施例一中文本匹配方法的流程圖；

圖3爲本申請實施例二中文本匹配方法的流程圖；

圖4爲本申請實施例三中文本匹配方法的流程圖；

圖5爲本申請實施例五中文本匹配實現原理的示意圖；

圖6爲本申請實施例五中文本匹配方法的流程圖；

圖7爲本申請實施例五中詞頻表更新的原理示意圖；

圖8爲本申請實施例中文本匹配裝置的結構示意圖。

【主要元件符號說明】

10：收集模組

20：分詞模組

30：權重確定模組

301：第一確定單元

302：第二確定單元

303：權重計算單元

40：詞頻更新模組

50：相似度確定模組

501：向量生成單元

502：相似度計算單元

60：文本比較模組

70：輸入過濾模組

80：輸出過濾模組

發明專利說明書

（本申請書格式、順序，請勿任意更動，※記號部分請勿填寫）

※申請案號：099140210

※申請日：099 年 11 月 22 日 **※IPC 分類：** G06F 17/30 (2006.01)
G06F 17/27 (2006.01)

一、發明名稱：（中文／英文）

文本匹配方法及裝置

二、中文發明摘要：

本申請公開了一種文本匹配方法及裝置，該方法包括：根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；對輸入的新增文本進行分詞並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。通過建立和更新詞頻表的方式避免了現有技術中每次匹配都需要對所有文本進行計算的問題，減少了匹配運算工作量，提高了系統性能。

三、英文發明摘要：

七、申請專利範圍：

1. 一種文本匹配方法，其特徵在於，包括：

週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；

對輸入的新增文本進行分詞，並提取關鍵字；根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；該詞頻表根據各個詞語在資料庫中的各文本中的出現頻率週期性更新；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本；

根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；

根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。

2. 如申請專利範圍第1項所述的方法，其中，該詞頻表根據各個關鍵字在資料庫中的各文本中的出現頻率週期性更新，具體包括：

每次輸入新增文本後，統計各個詞語在輸入的新增文本和資料庫中儲存的原始文本中的出現的頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表；或

每次輸入新增文本後，統計各個詞語在輸入的每個新增文本中的出現的頻率，根據統計得到的結果與詞頻表中

儲存的各個詞語在資料庫中的儲存的原始文本中的出現頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表。

3. 如申請專利範圍第2項所述的方法，其中，該根據預先儲存的詞頻表計算分詞得到的每個關鍵字在資料庫中各文本中的權重，具體包括：

根據詞頻表，分別確定選定關鍵字在資料庫中每個文本中的出現次數；以及

確定資料庫中的儲存的文本與包含有選定關鍵字的文本的數量比；

根據選定關鍵字在每個文本中的出現次數和該數量比，分別計算每個關鍵字在每個文本中的權重。

4. 如申請專利範圍第1項所述的方法，其中，該計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度，具體包括：

將待計算相似度的文本中的每個關鍵字的權重組成權重向量；

針對每個新增文本，分別計算該新增文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該新增文本與資料庫中儲存的各文本的相似度；或針對資料庫中儲存的每個文本，分別計算該文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該文本與資料庫中儲存的各文本的相似度。

5. 如申請專利範圍第1項所述的方法，其中，該根據

計算得到的相似度確定資料庫中儲存的各文本的相關文本，具體包括：

針對待確定相關文本的每個文本，確定與該文本的相似度大於或大於等於設定閾值的至少一個資料庫中儲存的文本爲該文本的相關文本；或

針對待確定相關文本的每個文本，根據資料庫中各文本與待確定相關文本的文本的相似度大小排序，確定相似度較高的設定數量的資料庫中儲存的文本作爲待確定相關文本的文本的相關文本。

6. 如申請專利範圍第1-5項之任一項所述的方法，其中，該根據計算得到的相似度確定資料庫中儲存的各文本的相關文本之前，還包括：

根據計算得到的每個新增文本與資料庫中的每個文本的相似度，或計算得到的資料庫中任意兩個文本的相似度；對待確定相關文本的新增文本或資料庫中儲存的文本相關的相似度數據進行過濾，去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度小於設定閾值的文本，或去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度較低的設定數量的文本。

7. 如申請專利範圍第1-5項之任一項所述的方法，其中，該根據當前週期內收集的內容資訊得到當前週期內的新增文本之前，還包括：

根據設定的輸入過濾規則，對當前週期內收集到用戶發佈的內容資訊進行過濾，根據過濾後內容資訊得到當前

週期內的新增文本。

8. 如申請專利範圍第7項所述的方法，其中，該根據設定的輸入過濾規則，對當前週期內收集到用戶發佈的內容資訊進行過濾，具體包括：

根據內容資訊的品質是否符合設定的品質評估閾值和/或發佈內容資訊的用戶是否是設定的合格用戶，對該收集到的內容資訊進行過濾。

9. 一種文本匹配裝置，其特徵在於，包括：

收集模組，用於週期性收集用戶發佈的內容資訊，根據當前週期內收集的內容資訊得到當前週期內的新增文本並儲存到資料庫中；

分詞模組，用於對輸入的新增文本進行分詞，並提取關鍵字；

權重確定模組，用於根據預先儲存的詞頻表計算提取的每個關鍵字在資料庫中的各文本中的權重；

詞頻更新模組，用於根據各個詞語在資料庫中的各文本中的出現頻率週期性更新詞頻表；資料庫中的文本包括當前週期儲存的新增文本和之前儲存的原始文本；

相似度確定模組，用於根據計算得到的每個關鍵字在資料庫中的各文本中的權重，計算每個新增文本與資料庫中的各文本的相似度，或計算資料庫中任意兩個文本的相似度；

文本比較模組，用於根據計算得到的相似度確定資料庫中儲存的各文本的相關文本。

10. 如申請專利範圍第9項所述的裝置，其中，該詞頻更新模組，具體用於：

每次輸入新增文本後，統計各個詞語在輸入的新增文本和資料庫中儲存的原始文本中的出現的頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表；或

每次輸入新增文本後，統計各個詞語在輸入的每個新增文本中的出現的頻率，根據統計得到的結果與詞頻表中儲存的各個詞語在資料庫中的儲存的原始文本中的出現頻率，得到包含各個詞語在資料庫中的每個文本中的出現頻率的的詞頻表。

11. 如申請專利範圍第10項所述的裝置，其中，該權重確定模組，具體包括：

第一確定單元，用於根據詞頻表，分別確定選定關鍵字在資料庫中每個文本中的出現次數；

第二確定單元，用於確定資料庫中儲存的文本與包含有選定關鍵字的文本的數量比；

權重計算單元，用於根據選定關鍵字在每個文本中的出現次數和該數量比，分別計算每個關鍵字在每個文本中的權重。

12. 如申請專利範圍第9項所述的裝置，其中，該相似度確定模組，具體包括：

向量生成單元，用於將待計算相似度的文本中的每個關鍵字的權重組成權重向量；

相似度計算單元，用於針對每個新增文本，分別計算該新增文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該新增文本與資料庫中儲存的各文本的相似度；或針對資料庫中儲存的每個文本，分別計算該文本的權重向量與資料庫中儲存的各文本的權重向量的內積，得到該文本與資料庫中儲存的各文本的相似度。

13. 如申請專利範圍第9項所述的裝置，其中，該文本比較模組，具體用於：

針對待確定相關文本的每個文本，確定與該文本的相似度大於或大於等於設定閾值的至少一個資料庫中儲存的文本的相關文本；或

針對待確定相關文本的每個文本，根據資料庫中各文本與待確定相關文本的文本的相似度大小排序，確定相似度較高的設定數量的資料庫中儲存的文本作爲待確定相關文本的文本的相關文本。

14. 如申請專利範圍第9-13項之任一項所述的裝置，其中，還包括：

輸入過濾模組，用於根據設定的輸入過濾規則，對當前週期內收集到用戶發佈的內容資訊進行過濾，根據過濾後內容資訊得到當前週期內的新增文本。

15. 如申請專利範圍第9-13項之任一項所述的裝置，其中，還包括：

輸出過濾模組，用於根據該相似度確定模組計算得到的每個新增文本與資料庫中的每個文本的相似度，或計算

得到的資料庫中任意兩個文本的相似度；對待確定相關文本的新增文本或資料庫中儲存的文本相關的相似度數據進行過濾，去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度小於設定閾值的文本，或去除與待確定相關文本的新增文本或資料庫中儲存的文本相似度較低的設定數量的文本；

該文本比較模組具體用於：根據過濾後的文本確定資料庫中儲存的各文本的相關文本。

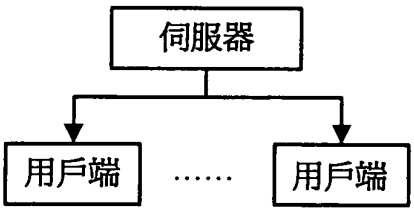

伺服器
用戶端
……
用戶端

圖1

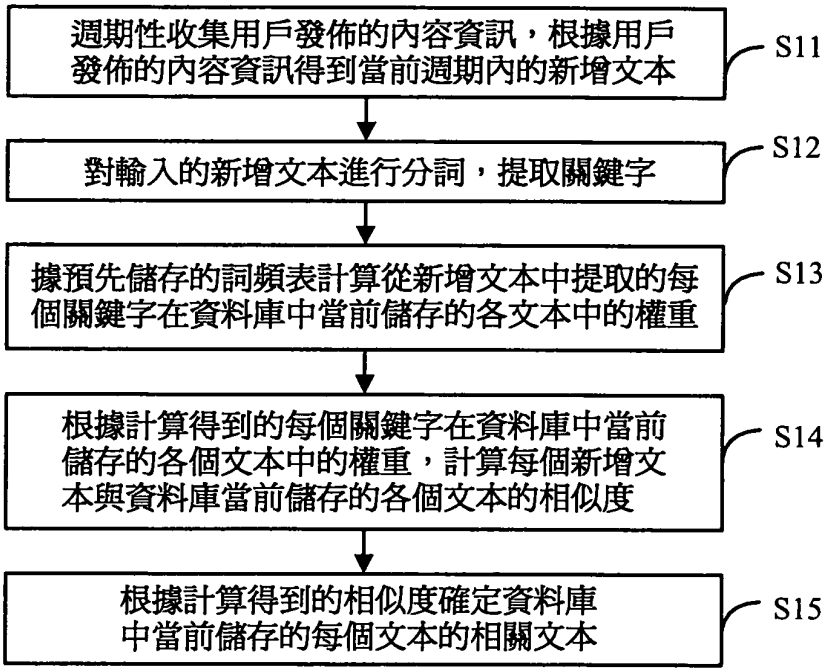

週期性收集用戶發佈的內容資訊，根據用戶發佈的內容資訊得到當前週期內的新增文本
S11
對輸入的新增文本進行分詞，提取關鍵字
S12
據預先儲存的詞頻表計算從新增文本中提取的每個關鍵字在資料庫中當前儲存的各文本中的權重
S13
根據計算得到的每個關鍵字在資料庫中當前儲存的各個文本中的權重，計算每個新增文本與資料庫當前儲存的各個文本的相似度
S14
根據計算得到的相似度確定資料庫中當前儲存的每個文本的相關文本
S15

圖2

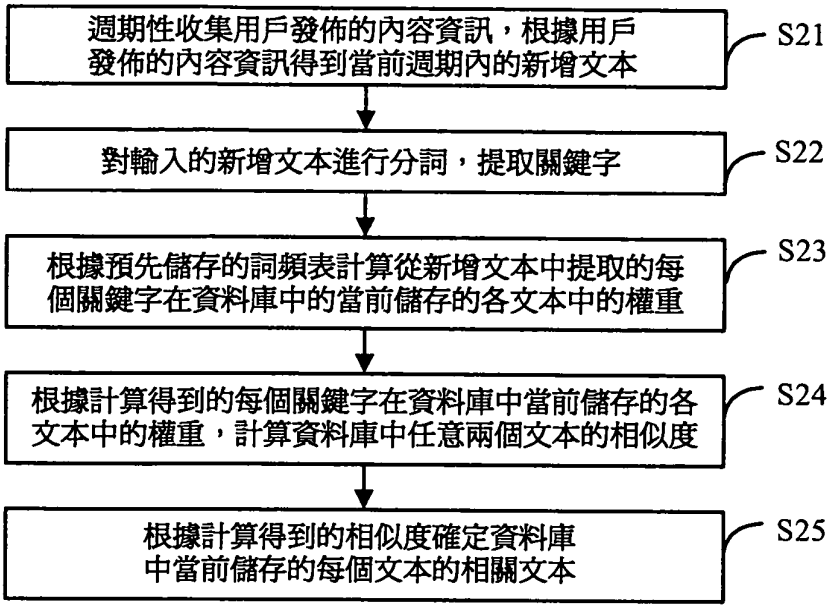

週期性收集用戶發佈的內容資訊，根據用戶發佈的內容資訊得到當前週期內的新增文本
S21
對輸入的新增文本進行分詞，提取關鍵字
S22
根據預先儲存的詞頻表計算從新增文本中提取的每個關鍵字在資料庫中的當前儲存的各文本中的權重
S23
根據計算得到的每個關鍵字在資料庫中當前儲存的各文本中的權重，計算資料庫中任意兩個文本的相似度
S24
根據計算得到的相似度確定資料庫中當前儲存的每個文本的相關文本
S25

圖3

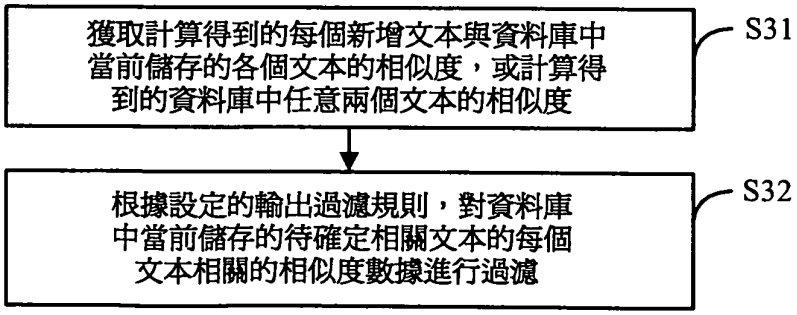

獲取計算得到的每個新增文本與資料庫中當前儲存的各個文本的相似度，或計算得到的資料庫中任意兩個文本的相似度
S31
根據設定的輸出過濾規則，對資料庫中當前儲存的待確定相關文本的每個文本相關的相似度數據進行過濾
S32

圖4

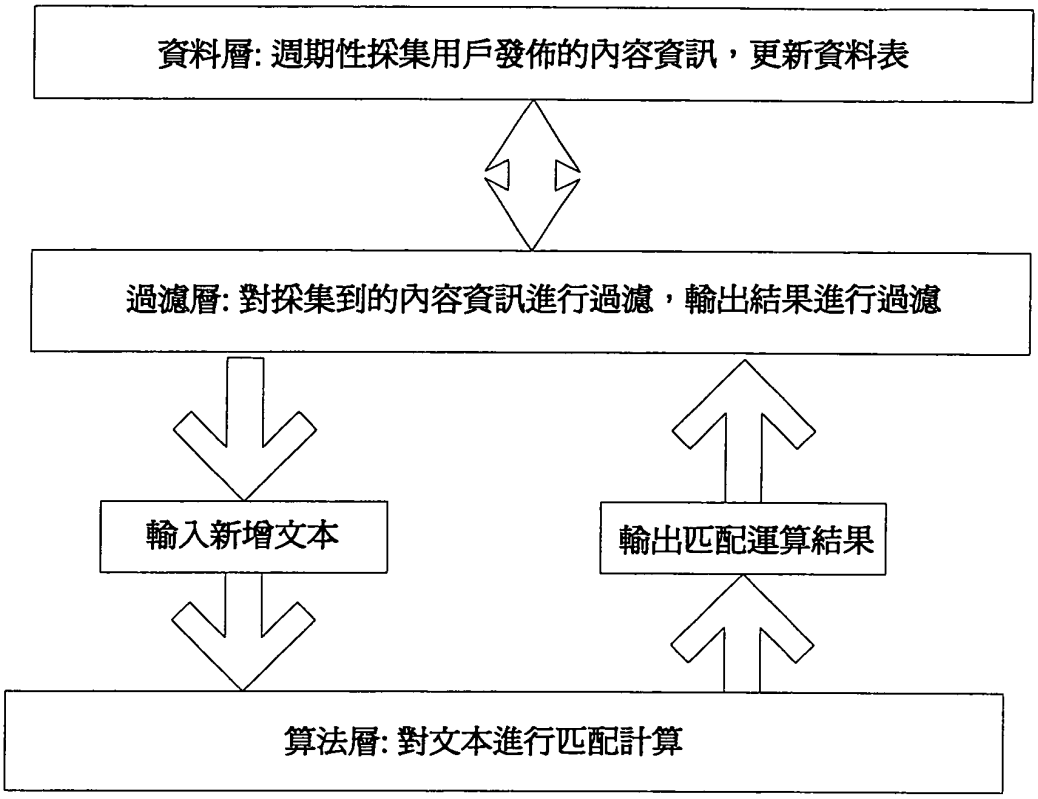

資料層: 週期性採集用戶發佈的內容資訊，更新資料表
過濾層: 對採集到的內容資訊進行過濾，輸出結果進行過濾
輸入新增文本
輸出匹配運算結果
算法層: 對文本進行匹配計算

圖5

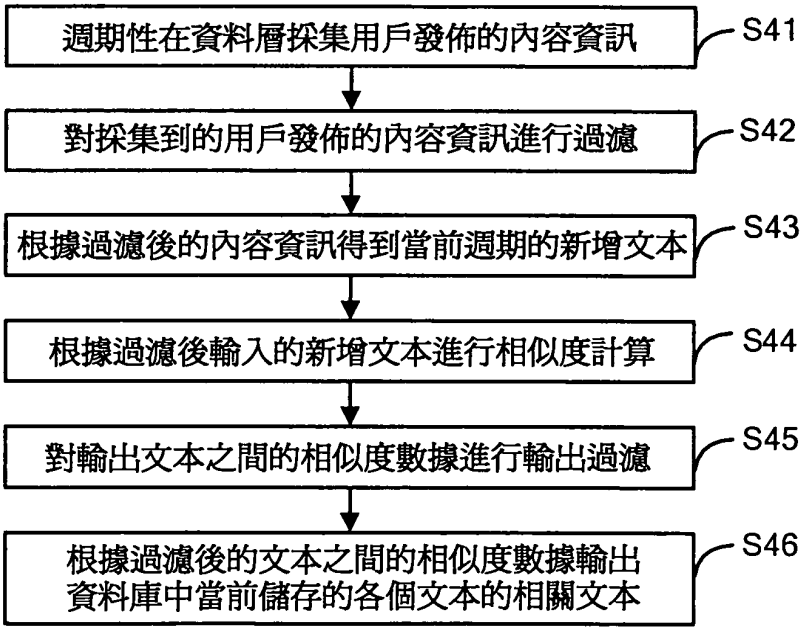

週期性在資料層採集用戶發佈的內容資訊
S41
對採集到的用戶發佈的內容資訊進行過濾
S42
根據過濾後的內容資訊得到當前週期的新增文本
S43
根據過濾後輸入的新增文本進行相似度計算
S44
對輸出文本之間的相似度數據進行輸出過濾
S45
根據過濾後的文本之間的相似度數據輸出
資料庫中當前儲存的各個文本的相關文本
S46

圖6

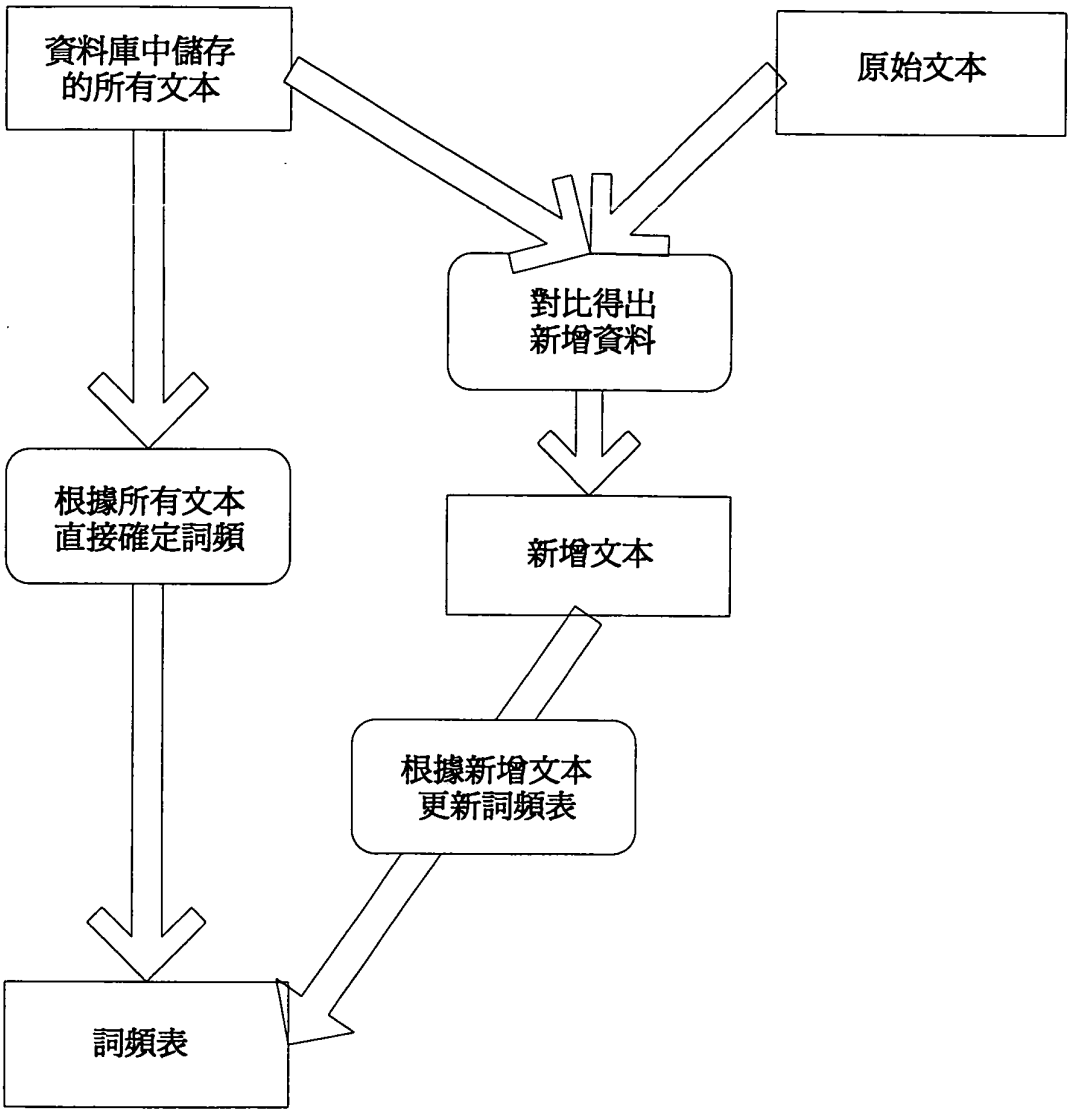

資料庫中儲存
的所有文本
原始文本
對比得出
新增資料
根據所有文本
直接確定詞頻
新增文本
根據新增文本
更新詞頻表
詞頻表

圖7

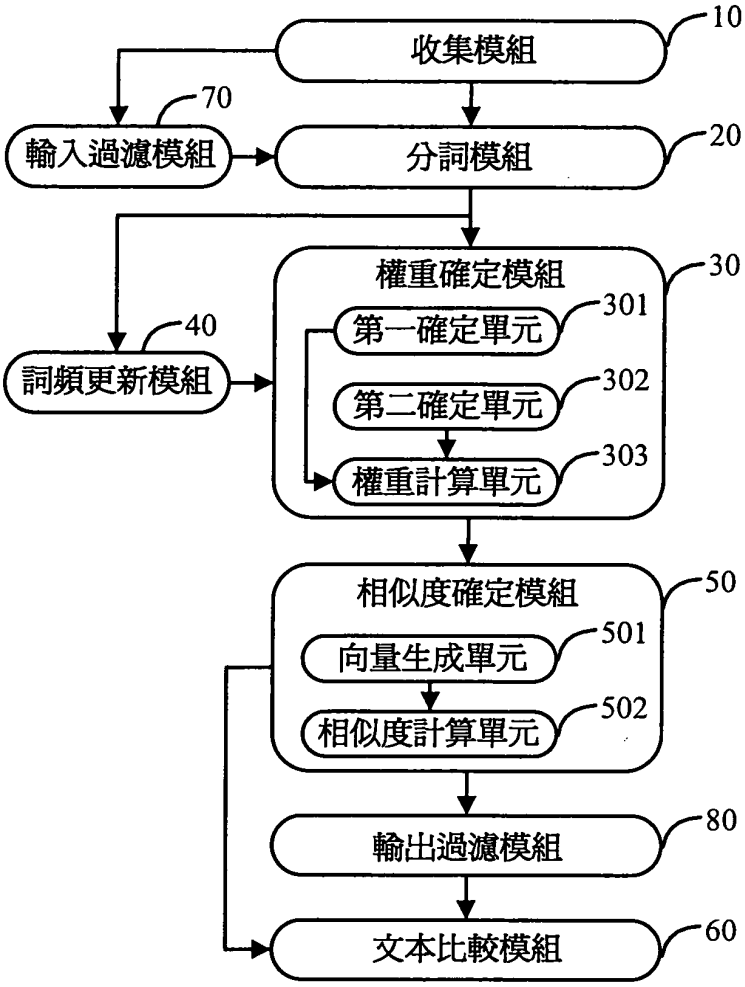

收集模組
10
輸入過濾模組
70
分詞模組
20
權重確定模組
30
第一確定單元
301
第二確定單元
302
權重計算單元
303
詞頻更新模組
40
相似度確定模組
50
向量生成單元
501
相似度計算單元
502
輸出過濾模組
80
文本比較模組
60

圖8

四、指定代表圖：

（一）本案指定代表圖為：第（2）圖。

（二）本代表圖之元件符號簡單說明：無

5

五、本案若有化學式時，請揭示最能顯示發明特徵的化學式：無