(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 201027375 A1
(43)公開日： 中華民國 99 (2010) 年 07 月 16 日

(21)申請案號：098133263 (22)申請日： 中華民國 98 (2009) 年 09 月 30 日
(51)Int. Cl.： *G06F17/30 (2006.01)*
(30)優先權：2008/10/20 日本 2008-270028
(71)申請人：萬國商業機器公司 (美國) INTERNATIONAL BUSINESS MACHINES CORPORATION (US)
美國
(72)發明人：守屋豐 MORIYA, YUTAKA (JP) ；照井文彥 TERUI, FUMIHIKO (JP)
(74)代理人：陳長文
申請實體審查：無 申請專利範圍項數：20 項 圖式數：11 共 48 頁

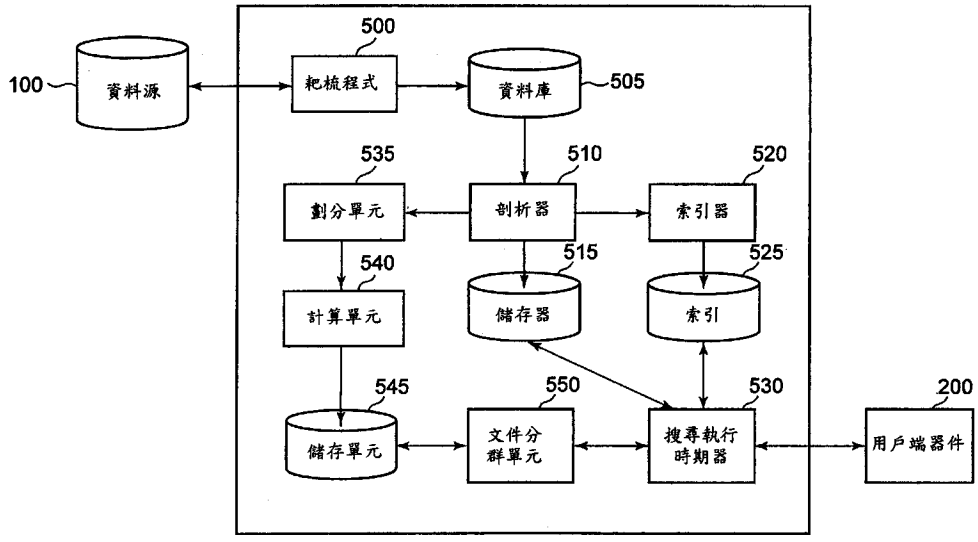
(54)名稱

搜尋系統，搜尋方法及程式

SEARCH SYSTEM, SEARCH METHOD AND PROGRAM

(57)摘要

提供易於在一搜尋結果之文件當中尋找真正所需之一文件的一種搜尋系統及一種搜尋方法。此搜尋系統包括：一劃分單元，其根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊；及一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊，而自該儲存單元 545 提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。



100：資料源/資料庫
200：用戶端器件
500：耙梳程式
505：資料庫
510：剖析器
515：儲存器
520：索引器
525：索引
530：搜尋執行時期器
535：劃分單元
540：計算單元
545：儲存單元
550：文件分群單元

(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 201027375 A1
(43)公開日： 中華民國 99 (2010) 年 07 月 16 日

(21)申請案號：098133263 (22)申請日： 中華民國 98 (2009) 年 09 月 30 日
(51)Int. Cl.： *G06F17/30 (2006.01)*
(30)優先權：2008/10/20 日本 2008-270028
(71)申請人：萬國商業機器公司 (美國) INTERNATIONAL BUSINESS MACHINES CORPORATION (US)
美國
(72)發明人：守屋豐 MORIYA, YUTAKA (JP) ；照井文彥 TERUI, FUMIHIKO (JP)
(74)代理人：陳長文
申請實體審查：無 申請專利範圍項數：20 項 圖式數：11 共 48 頁

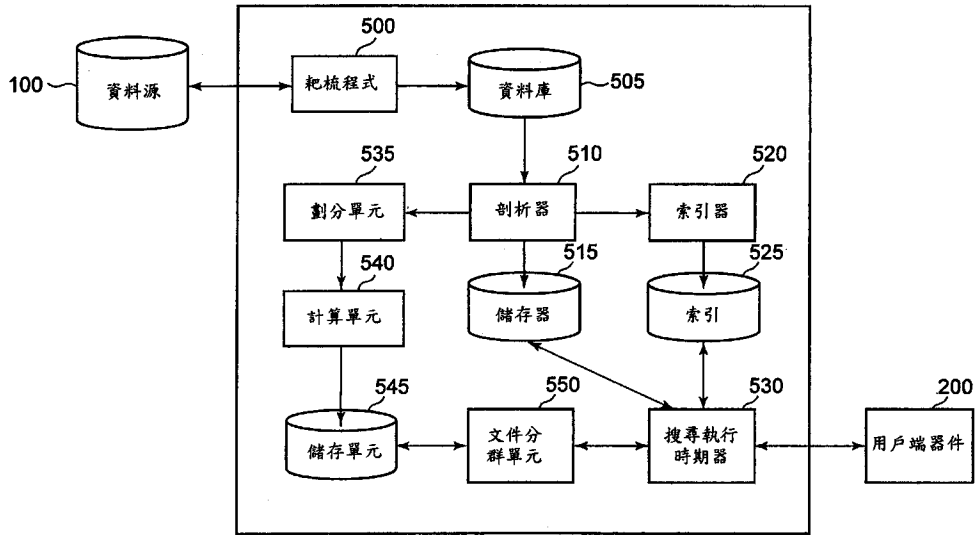
(54)名稱

搜尋系統，搜尋方法及程式

SEARCH SYSTEM, SEARCH METHOD AND PROGRAM

(57)摘要

提供易於在一搜尋結果之文件當中尋找真正所需之一文件的一種搜尋系統及一種搜尋方法。此搜尋系統包括：一劃分單元，其根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊；及一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊，而自該儲存單元 545 提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。



100：資料源/資料庫
200：用戶端器件
500：耙梳程式
505：資料庫
510：剖析器
515：儲存器
520：索引器
525：索引
530：搜尋執行時期器
535：劃分單元
540：計算單元
545：儲存單元
550：文件分群單元

六、發明說明：

【發明所屬之技術領域】

本發明係關於一種能夠在搜尋結果當中偵測所偵測之文件中的哪些文件包括重疊內容的搜尋系統，針對其之一種搜尋方法，及一種實施該方法的電腦可讀程式。

【先前技術】

搜尋引擎可用作搜尋儲存於一與網路(諸如，網際網路)連接之資料庫中的文件的系統。一些搜尋引擎具有自複數個文件搜尋特定字元串之全文搜尋功能。

配備該全文搜尋功能之此全文搜尋引擎經分類成循序搜尋類型及索引類型，其中循序搜尋類型引擎逐一掃描複數個文件的內容以搜尋字元串。然而，當必須搜尋極大數目個文件時，如此使循序搜尋花費長時間來進行搜尋，索引類型搜尋引擎預先建立具有由一字元串、一文件位置、一更新時間、一出現頻率及其類似者組成之一表結構的一索引，且在搜尋時存取該索引，因此實現快速搜尋。

用於索引類型搜尋引擎之索引具有各種格式，通常包括一反向索引，其具有由多個字組成之一可變長度記錄及包括該等字之一文件檔案ID。

現參看圖1及圖2，在下文中例示三個文件、對應於該等文件之一反向索引，及保存收集之文件之一資料結構。圖1A至圖1C中所說明之皆為電子郵件文件的文件分別具有1至3之文件檔案ID。圖2A說明由用作一關鍵字之一字及包括該字之一ID組成的一反向索引，其中包括字「PHP」、

「鈴木」(英文「Suzuki」)及「代碼コード」(英文「code」)的文件與其相關聯。圖2B說明用以儲存所收集之文件之資料結構的項實例，其中用作一關鍵字之一字及對應於該字之文件的內容彼此相關聯。在圖2B中，字列於左邊欄位中，且對應於選定字之文件內容展示於右邊欄位中。

全文搜尋引擎傳回與一搜尋字匹配之一字出現所在的文件的一群組作為搜尋結果。判斷文件之間的相似性的此等技術整體上描述於(例如)專利文件1至4中。

此等技術並不考慮哪些字元串包括文件中與搜尋字匹配的字。因此，當搜尋結果包括大量文件時，難以在不強加負荷之情況下尋找真正所需的文件。舉例而言，當搜尋字存在於文件之範本中時，將傳回使用該範本之所有文件，因此強加自搜尋結果尋找在其正文(main body)中包括搜尋字的文件作為真實目標的負荷。在本文中，範本指代文件之頁首或頁尾、Web網站之選單、電子郵件之簽名或其類似者。

在電子郵件之狀況下，回覆郵件或轉發郵件在該電子郵件結尾常包括其原始郵件的複本。若複本部分包括一搜尋字，則甚至當該郵件之正文不包括該搜尋字時，傳回之搜尋結果亦將包括該郵件。當一搜尋必須針對在其正文中包括該搜尋字的郵件而進行時，此狀況造成無意義資料(noise)。

因此，若可將在其正文中於相同字元串中包括該搜尋字

的文件收集至一個群組中，則待評估之文件的數目得以減少，因此易於尋找真正所需的文件。

舉例而言，已提議一種在考慮搜尋字之出現位置的情況下偵測具有重疊內容之文件的技術(參看專利文件4)，其針對包括於所偵測之搜尋結果中的該等文件中之每一者擷取並比較包括搜尋關鍵字的字元串。

圖3說明描述於專利文件4中之搜尋引擎的組態。搜尋引擎10與保存待搜尋之文件之一資料源20連接且進一步與一用戶端器件30連接，該用戶端器件30輸出由使用者所輸入之詢問(查詢)以獲取搜尋結果。

搜尋引擎10具備：一資料庫11，其將多個文件記錄於其中；及一耙梳程式(crawler)12，其以定期間隔獲取資料源20上之文件以建立索引。耙梳程式12重複請求用於索引建立之文件的複本、追蹤包括於該文件中之連接及收集另一文件的操作。當耙梳程式12發現新文件時，該耙梳程式12在資料庫11中記錄該新文件。當耙梳程式12發現文件不再可用時，則該耙梳程式12自資料庫11刪除該文件。

搜尋引擎10具備一剖析器13，其自由耙梳程式12所獲取及在資料庫11中記錄之文件擷取本文，且擷取諸如段落之格式資訊。剖析器13執行語法分析，且將作為語法分析之結果而擷取之本文及格式資訊輸入至稱為儲存器14之資料結構，該儲存器14儲存收集之文件。

搜尋引擎10具備一索引器(indexer)15，該索引器15基於由剖析器13所擷取之本文及格式資訊而建立索引。索引器

15使用作一關鍵字之一字與包括該字之文件的ID(如上文所描述)相關聯，且將其儲存在索引16中。

搜尋引擎10進一步具備：用作搜尋伺服器之一搜尋執行時期器(search run time)17，其回應於自用戶端器件30所接收之包括搜尋字的查詢而搜尋包括作為關鍵字之一搜尋字的文件；一查詢相關資訊產生器件18，其接收來自搜尋執行時期器17之搜尋結果，自該儲存器14獲取包括搜尋字的文件，及產生包括該搜尋字之字元串；及一查詢相關資訊比較器件19，其比較經產生之字元串與該搜尋結果中之文件。

搜尋引擎10使查詢相關資訊產生器件18針對每一搜尋及每一搜尋結果產生包括搜尋字的字元串，且使查詢相關資訊比較器件19比較該等字元串，因此偵測整體上彼此匹配的文件及作為相關文件之包括彼此匹配之若干取樣部分的文件。

[專利文件1]美國專利第6,230,155號

[專利文件2]美國專利第6,658,423號

[專利文件3]美國專利第6,978,419號

[專利文件4]美國專利第6,615,209號

習知搜尋引擎處置具有相同內容之不同文件作為個別搜尋結果，因此有可能在文件收集或索引建立時預先排除具有相同內容或相似內容的此等文件。然而，習知搜尋引擎僅可判斷文件或其若干部分具有相同內容或相似內容，但不可基於部分一致性來判斷文件具有相同內容或相似內

容。

當搜尋字出現在Web網站之選單中時，習知搜尋引擎傳回包括該選單之所有頁。儘管可藉由預先指定並非出現作為文件之特性的字及字元串來限制傳回的頁，但此等字及字元串必須在指定之前為已知的。

另外，習知搜尋引擎在不考慮文件之間的關係的情況下傳回搜尋結果。因此，需要使用者逐一做出關於包括於傳回之搜尋結果中的所有文件是否為真正需要之文件的判斷。

【發明內容】

為了應對以上所陳述之問題，根據本發明，將組成一文件之本文劃分成複數個區塊，標註包括一搜尋字之區塊，且在包括於一搜尋結果中之文件當中，將包括此等區塊之具有相同內容的文件進行分群，使得具有相同內容或具有相似內容之文件可基於部分一致性而經判定，且可在考慮該等文件之間的關係的情況下傳回搜尋結果。

更具體言之，當建立一索引時，將待搜尋之一文件中的本文劃分成複數個區塊。一區塊可為一句子、一段落或其類似者。針對如此獲得之區塊中的每一者計算一雜湊值。該雜湊值為對應於一字元串之一數值。與該文件相關聯而保存此雜湊值與該文件中之區塊的位置資訊。

接著，當執行搜尋時，針對搜尋結果中之每一文件，根據表示其中出現搜尋字之區塊之位置的相應位置資訊而提取一雜湊值，且將具有相同雜湊值之文件分群並輸出。

為實施此情形，本發明提供一種搜尋系統，其包括：一劃分單元，其根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊；及一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊而自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。

該劃分單元可根據該劃分資訊以下列方式中之至少一者劃分該文件：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分。該額外資訊可包括一HTML文件中之一HTML標籤。該劃分單元可不僅使用一種類型之劃分資訊而且使用複數種類型之劃分資訊來劃分一文件。舉例而言，當使用一特定搜尋字時，可使用針對每一段落之劃分資訊，且當使用不同於該特定搜尋字之另一搜尋字時，可使用針對每一句子之劃分資訊。以此方式，所使用之複數種類型的劃分資訊允許在使用者或系統判斷使用針對每一句子之劃分資訊的分群不適當時使用不同於針對每一句子之劃分資訊的劃分資訊(例如，針對每一段落之劃分資訊)以用於分群。

該文件包括一符記串，在該符記串中複數個字或符記經循序排序，且每一區塊中所包含之一字元串包括至少一符

記。因此，可由符記之數目來表示每一區塊的位置。該位置資訊可包括自該文件中之一前置符記至每一區塊之一前置符記的一符記次序。該位置資訊可進一步包括自該文件中之該前置符記至每一區塊之一結束符記的一符記次序。此等兩個符記數目可用作組成該區塊之符記串之前置符記至結束符記的範圍。

另外，亦可由字元之數目來表示每一區塊的位置。在此狀況下，該位置資訊可包括自該文件中之一前置字元至每一區塊中所包含之一字元串的一前置字元的字元之一數目。該位置資訊可進一步包括自該文件中之該前置字元至每一區塊之一結束字元的字元之數目。此等兩個字元數目可用作組成該區塊之字元串之前置字元至結束字元的範圍。

當該區塊中所包括之一字元串包括一指定字元類型時，該計算單元可藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值。在電子郵件中，當引述所接收內容時，常添加標記「>」。接著，基於已排除此標記「>」之字元串來計算雜湊值，藉此可將具有相同雜湊值之文件分群。

該文件分群單元可包括一排序單元，該排序單元根據一搜尋記分對包括於一群組中之複數個文件進行排序。藉此，可以搜尋記分之次序來排列包括於該群組中之該複數個文件。

本發明亦可提供一種由以上所陳述之搜尋系統所執行的

搜尋方法。此方法包括由劃分單元、計算單元、儲存單元及文件分群單元所執行的處理步驟。

此搜尋方法可經組態為一程式且可藉由執行該程式而體現。此程式可儲存於一記錄媒體中以用於提供。

本發明之搜尋系統、搜尋方法、程式及記錄媒體使得易於自搜尋結果尋找真正所需之文件，因此減少搜尋所需文件之麻煩且縮短搜尋時間。

【實施方式】

下文借助於特定實施例來描述本發明，該等特定實施例並不意欲將本發明限於以下實施例。

圖4例示一網路系統，其包括保存待搜尋之文件的一資料源、發出搜尋請求之一用戶端器件，及包括接收該搜尋請求及執行搜尋處理之一搜尋引擎的一伺服器器件。此圖式僅說明一個資料源100、一個用戶端器件200及一個伺服器器件300。然而，兩個或兩個以上此等器件可與網路400連接。資料源100與伺服器器件300可直接連接。

資料源100可為可保存文件之任何器件，其可為針對每一項目收集資料且管理相同或另一伺服器器件的資料庫。資料源100可為(例如)保存文件且由另一使用者使用之PC。

當資料源100為資料庫時，關連式資料庫可用於該資料庫，其包括作為基本資料類型之複數個關係，其中使用諸如等號及不等號之關連式運算子及諸如邏輯積、邏輯和及邏輯非之邏輯運算子產生用以獲取經儲存資料的詢問。該

資料庫可直接建構於由作業系統(OS)提供之檔案系統上，或可使用資料庫管理系統(DBMS)來建構。

用戶端器件200可為可輸出搜尋請求之任何器件，其可為配備能夠自藉由使用者輸入之搜尋字產生搜尋請求且經由網路進行詢問之應用程式的PC。此PC配備一允許使用者輸入搜尋字之鍵盤、一指定輸入位置且給出開始搜尋之指令的滑鼠、一顯示輸入螢幕及搜尋結果之顯示器件、用於與網路連接之一網路介面、儲存應用程式之一HDD、一在其上讀出應用程式以用於執行的RAM、一執行應用程式之CPU，及其類似者。除應用程式外，Web瀏覽器可用以實現經由網路之通信。

伺服器器件300亦可具有與用戶端器件200之硬體組態類似的硬體組態，其配備用於與Web瀏覽器通信的一Web伺服器及用於處理自用戶端器件200接收之搜尋請求的一搜尋引擎。

伺服器器件300可具有與用戶端器件200之硬體組態類似之如上文所描述的硬體組態。現參看圖5，在下文中簡要例示伺服器器件300之硬體組態。在圖5之硬體組態中，伺服器器件300具備一記憶體310、至少一處理器320、一記憶體控制單元330、一頻道子系統340、至少一控制器350，及至少一輸入/輸出器件360。

記憶體310儲存經由輸入/輸出器件360所輸入之資料及程式，且回應於藉由處理器320或頻道子系統340作出之位址指定而將儲存於該位址處的資料或其類似者發送至處理

器320或頻道子系統340。

處理器320控制整個裝置，且執行至少一OS。OS控制裝置中之程式及輸入/輸出處理之執行。記憶體控制單元330經由匯流排與記憶體310、處理器320及頻道子系統340連接。此記憶體控制單元330允許自處理器320或頻道子系統340發出之請求臨時儲存於佇列中且藉由預定時序發送至記憶體310。

頻道子系統340與控制器350中之每一者連接，且控制輸入/輸出器件360與記憶體310之間的資料傳送，以便減少處理器320之處理負載。藉此，可並列執行由處理器320進行之計算處理及由輸入/輸出器件360進行之輸入/輸出處理，因此改良處理程序效率。

控制器350控制由輸入/輸出器件360進行之資料傳送的時序或其類似者。輸入/輸出器件360經由控制器350、頻道子系統340及記憶體控制單元330將資料傳送至記憶體310/傳送來自記憶體310的資料。當輸入/輸出器件360、HDD、顯示器、鍵盤、印表機、通信器件及其他儲存器件可用時，輸入/輸出器件360中之一者直接與資料庫100連接或經由網路400連接。

為實施藉由伺服器器件300進行之搜尋處理，提供其上記錄有程式的記錄媒體且使其與輸入/輸出器件360中之一者連接。接著，將程式經由控制器350、頻道子系統340及記憶體控制單元330發送至記憶體310，且儲存至記憶體310。經儲存之程式安裝於同樣經由相同器件與輸入/輸出

器件360連接之HDD中，且由處理器320適當地讀出以用於執行。

作為其上儲存有之程式的記錄媒體，軟性磁碟、CD-ROM、DVD、SD卡、快閃記憶體或其類似者為可用的。此程式包括用於執行搜尋處理且輸出搜尋結果之程式。此程式安裝於同一HDD中，該程式由處理器320適當地讀出以用於執行，因此實施搜尋引擎之功能。

圖6為展示伺服器器件300經組態為搜尋系統的功能方塊圖。與圖3中所說明之習知搜尋引擎類似，此搜尋系統包括：一耙梳程式500，其作為週期性地獲取文件之獲取單元；一資料庫505，其作為儲存經獲取之文件之儲存單元；一剖析器510，其作為自該文件擷取本文且擷取諸如段落之格式資訊的擷取單元；一儲存器515，其作為儲存經擷取之本文及格式資訊之儲存單元；一索引器520，其作為自該本文及格式資訊建立索引之建立單元；一索引525，作為保存經建立之索引之保存單元；及一搜尋執行時期器530，其用作回應於自用戶端器件200接收之包括一搜尋字的搜尋請求而搜尋包括作為關鍵字之該搜尋字之文件的搜尋單元。

圖3中所說明之習知搜尋引擎包括查詢相關資訊產生器件18及查詢相關資訊比較器件19。另一方面，圖6中所說明之搜尋系統包括一劃分單元535、一計算單元540、一儲存單元545，及一文件分群單元550。

由於耙梳程式500、資料庫505、剖析器510、儲存器

515、索引器520、索引525及搜尋執行時期器530之每一功能已於上文加以描述，因此下文詳細描述劃分單元535、計算單元540、儲存單元545及文件分群單元550。

劃分單元535接收由剖析器510所擷取之本文及格式資訊，且根據由使用者所指定之劃分資訊將該本文劃分成複數個區塊。劃分資訊展示將如何劃分本文，其可選自下列方式中之至少一者：劃分成每一句子；劃分成每一段落；在一空行(null line)處劃分；及基於經添加至文件之額外資訊而劃分。當選擇每一句子時，本文將劃分成每一句子。可使用複數種類型之劃分資訊。舉例而言，當使用特定搜尋字時，可使用針對每一段落之劃分資訊，且當使用不同於該特定搜尋字之另一搜尋字時，可使用針對每一句子之劃分資訊。設定複數種類型之劃分資訊，使得可使用此資訊進行劃分，藉此當使用者或系統判斷使用針對每一句子之劃分資訊來進行分群不適當時，則針對每一段落之劃分資訊可用於分群。以此方式，使用複數個準則之劃分為有效的，因為其能夠調整分群在搜尋期間的粒度。在本文中，額外資訊可包括HTML文件中之HTML標籤。當建立索引時，可進行此種劃分。

計算單元540藉由將一雜湊函數應用於區塊中所包括之一字元串來計算每一區塊的雜湊值。該雜湊函數自資料產生特定範圍之數值，且藉由應用該雜湊函數所獲得之雜湊值為對應於每一字元串之數值。可使用Java®語言之標準方法(諸如，hashCode())來計算雜湊值。此處，hashCode()

為用以傳回雜湊值之方法。

雜湊函數之一實例包括添加經指派給字元串之每一字元的字元碼(例如，數值)的函數。此狀況下之字元碼包括ASCII字元碼。上文所陳述之實例僅為一實例，且任何已知計算公式及演算法可用以得到雜湊值。

儲存單元545儲存由計算單元540所計算之雜湊值與文件中之區塊的位置資訊。將在下文詳細描述區塊之位置資訊。

對於藉由基於搜尋字進行搜尋所獲得的每一文件，文件分群單元550根據關於包括搜尋字之區塊的位置資訊而自儲存單元545提取相應雜湊值。接著，文件分群單元550分群具有相同雜湊值之文件，且輸出該等文件作為搜尋結果。將因此輸出之搜尋結果發送至搜尋執行時期器530，且搜尋執行時期器530將搜尋結果傳回至用戶端器件200。當Web瀏覽器接收搜尋結果時，用戶端器件200使顯示器件顯示該搜尋結果。

現參看圖7至圖11，將在下文詳細描述以上處理。圖7A至圖7C說明作為文件實例之三種類型的電子郵件。所有此等電子郵件實例包括一正文及包括簽名及其類似者之一簽名部分，且在該正文與該簽名部分之間存在一空線。在本文中，將「空線」指定為劃分資訊，且劃分單元535基於指定「空線」之劃分資訊在空線處將電子郵件劃分成兩個部分，亦即，正文部分及簽名部分。更具體言之，在耙梳程式500週期性地獲取文件且剖析器510執行對該文件之語

法分析之後，劃分單元535將經受語法分析之文件劃分成複數個區塊。

圖8說明雜湊函數應用於經劃分區塊中之每一者中所包括的一字元串且計算每一區塊之雜湊值的狀態。剖析器510使每一區塊中所包含之一字元串包括一符記(字，其中在該字之前及/或之後具有一空格)串。在圖8A中，在正文「附加PHP的原始代碼。拜託了。(PHPのソースコードを添付します。よろしくお願いします。)」(英文「A source code of PHP is attached. Thanks in advance.」)與簽名部分「------鈴木 Example Corp Japan XXX@example.co.jp」(英文「------ Suzuki Example Corp Japan XXX@example.co.jp」)之間存在一空線，使得該空線將本文劃分成兩個符記串。

計算單元540藉由將雜湊函數應用於每一符記串而計算作為相應數值之雜湊值。就以上實例而言，基於「附加PHP的原始代碼。拜託了。」(英文「A source code of PHP is attached. Thanks in advance.」)之計算導致「1234567890」，且基於「------鈴木 Example Corp Japan XXX@example.co.jp」(英文「------ Suzuki Example Corp Japan XXX@example.co.jp」)之計算導致「0987654321」。在本文中，雜湊值經計算為10進制數位值(10-digit value)，其並非一限制性實例，且可使用任何進制數位值。

文件中之字元在行之方向上自左向右排列。當結束該行

時，字元在下一行中自左向右排列。因此，文件中之符記以自左上角之符記至右下角之符記的次序排列。位置資訊可包括自文件中之前置符記至每一區塊中之一字元串的前置符記的符記次序。可(例如)由使用此次序及自文件中之前置符記至每一區塊中所包含之字元串的結束符記的符記次序的範圍來表示區塊之位置。此範圍可用作位置資訊。

在上文所陳述之實例「附加PHP的原始代碼。拜託了。」(英文「A source code of PHP is attached. Thanks in advance.」)中，包括十三個符記「PHP」、「の」、「ソースコード」、「を」、「添付」、「し」、「ます」、「。」、「よろしく」、「お願い」、「し」、「ます」及「。」。由於「PHP」為第一符記，因此此符記為0符記。由於最終「。」為第十三符記，因此位置資訊可為「0符記至12個符記」。在圖8A中，此等符記使用「@」之標記來組合，且表示為「1234567890@0符記至12個符記」及「0987654321@13個符記至24個符記」。儲存單元545儲存此資訊。

在上文所陳述之實例中，自文件中之前置符記至每一區塊之前置符記之符記的數目用作至每一區塊之前置符記的符記次序。然而，存在剖析器510實際上自一個字產生複數個符記之狀況。舉例而言，可自僅五個字產生六個符記，使得亦可以一字之時態變化(conjugated)形式進行搜尋。另一方面，搜尋系統傳回指示在何符記數目處命中出現之資訊，且因此使用基於如上文所陳述之符記的數目所

計算之位置資訊而提取的區塊可能與正確區塊不同。

為應對此情形，下文描述將「附加PHP的原始代碼。拜託了。」(英文「A source code of PHP is attached. Thanks in advance.」)之描述由一句子劃分成多個區塊且計算其位置資訊的實例。假設剖析器510產生十五個符記「PHP」、「の」、「ソースコード」、「を」、「添付」、「し」、「ます」、「ました」、「。」、「よろしく」、「お願い」、「し」、「ます」、「ました」及「。」。在本文中，兩個符記「ました」經產生為時態變化形式(「ます」之過去形式)，其實際上不包括於句子中。當針對每一句子劃分以上描述時，劃分單元535將其劃分成兩個區塊「附加PHP的原始代碼。(PHPのソースコードを添付します。)」(英文「A source code of PHP is attached.」)及「拜託了。(よろしくお願いします。)」(英文「Thanks in advance.」)。

當計算單元540計算雜湊值及位置資訊時，計算單元540如下計算自剖析器510獲得之自前置符記之符記的數目：符記「ます」不作為第七及第十三符記而計算，且符記「ました」不作為第八及第十四符記而計算，替代地，並行排列相鄰符記「ます」及「ました」，使得其作為第七及第十二符記共同地計算。

接著，就「附加PHP的原始代碼。」(英文「A source code of PHP is attached.」)之區塊而言，計算單元540使用該區塊之前置符記的次序及結束符記的次序計算「雜湊值

@0至7」，且就「拜託了。」(英文「Thanks in advance.」)之區塊而言，類似次序用以計算「雜湊值@8至12」，且經計算之值儲存於儲存單元545中。

只要符記串不改變，經計算之雜湊值便將始終相同。然而，當甚至一個符記不同時，將獲得不同雜湊值。參看圖8A及圖8B，由於符記之一部分在正文中不同，因此其雜湊值具有不同值「1234567890」及「2345678901」。然而，由於所有符記在簽名部分中相同，因此其具有相同雜湊值「0987654321」。在圖8C中，正文部分及簽名部分兩者至少部分地與圖8A及圖8B中之正文部分及簽名部分不同，其雜湊值與圖8A及圖8B之雜湊值不同。

由具有特定字元類型之一標記構成之一符記可排除在雜湊計算外。藉此，就字元串「你好(こんにちは)」(英文「Hello」)及「你好(>こんにちは)」而言，由於該等字元串僅在標記「>」之部分不同，但「你好」之部分為其共同的，因此可計算出相同雜湊值。當引述電子郵件之內容時，通常添加此標記「>」。因此，即使當引述所接收之電子郵件之內容且將標記「>」添加至其時，只要其他符記以相同方式排列，即可獲得相同雜湊值。此對搜尋電子郵件為有效的。當建立索引時，可執行以上所描述之處理。在本文中，在計算雜湊值時所排除之字元類型不限於當在電子郵件中引述內容時添加的「>」及「>>」，且任何字元類型可由使用者預先指定，藉此在排除指定字元類型之情況下執行計算。

當用戶端器件200輸出搜尋請求時，搜尋執行時期器530基於包括於該搜尋請求中之搜尋字而自索引525搜尋由索引器520所建立的索引，且自儲存器515獲取由搜尋所獲得之文件的本文及格式資訊。搜尋執行時期器530將此資訊傳遞至文件分群單元550。

文件分群單元550針對搜尋結果中之每一文件基於包括該搜尋字之區塊的位置資訊自儲存單元545提取包括命中符記之區塊的雜湊值，且將具有相同雜湊值之文件分群為一個群組。

當基於輸入搜尋字執行搜尋時，搜尋執行時期器530傳回指示命中符記之序號的結果。在本文中，由於計算單元540計算符記串中之符記之序號作為位置資訊且儲存單元545儲存位置資訊，文件分群單元550基於自搜尋執行時期器530傳回之符記的序號而提取雜湊值，因此允許提取正確的雜湊值。

劃分單元535將包括複數個符記串之一文件劃分成複數個區塊，計算單元540基於每一區塊中所包含之符記串而計算每一雜湊值，且儲存單元545儲存經計算之雜湊值。當搜尋字包括於兩個或兩個以上區塊中時，基於包括於彼等兩個或兩個以上區塊中之符記串所計算的雜湊值可加總以提供文件之雜湊值，接著儲存該雜湊值。

當使用者經由用戶端器件200輸入「鈴木」(英文「Suzuki」)之搜尋字且提交針對其之搜尋請求時，搜尋執行時期器530搜尋索引525以獲得圖8A至圖8C中所說明之

三個文件作為搜尋結果。將圖8A至圖8C中所說明之文件分別稱作文件1至3。在文件1中，搜尋字「鈴木」(「Suzuki」)處於第十五符記處，且包括該符記之區塊的雜湊值為「0987654321」。在文件2中，搜尋字「鈴木」(「Suzuki」)處於第十七符記處，且包括該符記之區塊的雜湊值為「0987654321」，其與以上文件1之雜湊值相同。因此，將文件1及2分群至同一群組中。

在文件3中，搜尋字「鈴木」(「Suzuki」)處於第一符記處，且包括該符記之區塊的雜湊值為「3456789012」，其與文件1及2之雜湊值不同。因此，將文件3分群至一與文件1及2之群組不同的群組中。

經分群之文件可以任何顯示格式顯示為搜尋結果，只要其展示該等文件包括於一特定群組中。舉例而言，可如圖9B中所說明而顯示該等文件。在圖9B中所說明之搜尋結果中，正常顯示經分群至同一群組中之文件中的第一文件，且第二文件及後續文件藉由在其開始處添加之一分隔號(|)而向右縮排。藉此，使用者一看就可判斷搜尋結果之文件之間的關係。在本文中，經分群之文件的顯示不限於使用分隔號及縮排之以上式樣，且可藉由改變字元類型、添加識別標記或其類似者來展示該關係。

基於搜尋記分來排列經分群之文件。可如下獲得搜尋記分。自包括該搜尋字之文件的數目及所有文件的數目來計算表示所有文件當中有多少文件包括搜尋字之值，且將經計算之值與搜尋字出現的次數相乘，因此獲得搜尋記分。

因此，具有較大出現次數之文件具有較高記分，且具有較小出現次數之文件具有較低記分。

為與圖9B相比較，圖9A說明在不分群的情況下藉由搜尋執行時期器530基於搜尋字「鈴木」(英文「Suzuki」)進行之習知搜尋的結果。在圖9A中所說明之搜尋結果中，使用者必須評估搜尋結果中之每一者，而在圖9B中所說明之搜尋結果中，使用者可一眼就判斷出哪些結果重疊，使得可僅評估該等搜尋結果中的一者，因此易於尋找必要文件。

在至此所描述之實施例中，使用符記之次序來表示區塊之位置資訊。然而，表示位置資訊之方式不限於使用符記之次序，且其可使用經對準之字元的次序來表示。圖10A至圖10D說明電子郵件之四個實例，該等實例中之每一者被劃分成多個區塊，其中雜湊值與位置資訊彼此相關聯。

圖10中所說明之實例亦由劃分單元535在空線處劃分成多個區塊。圖10A及圖10B中所說明之文件1及2分別被劃分成兩個部分，該兩個部分為一正文部分及一簽名部分。圖10C及圖10D中所說明之文件3及4分別被劃分成四個部分及六個部分，該等部分分別包括複數個正文部分及簽名部分，其中添加標記「>」及「>>」至經引述的句子及簽名。

計算單元540自每一區塊中所包含之字元串計算雜湊值，使用一使用自文件之前置字元至字元串之前置字元的字元之數目及自文件之前置字元至字元串之結束字元的字

元之數目來表示的範圍作為位置資訊，且以相關聯方式將位置資訊及雜湊值儲存於儲存單元545中。參看圖10A中所說明之文件，其被劃分成係正文之「明天簽入db2jcc.jar。(db2jcc.jarを明日、チェックインします。) 」(英文「db2jcc.jar will be checked in tomorrow.」)及係簽名之「----田中」(英文「---- Tanaka」)，且「11111111」經計算以用於該正文，且「22222222」經計算以用於該簽名。由於該正文前不存在字元，其自第一個字元開始，且字元之數目為二十四，因此位置資訊為「1至24」。由於簽名自第二十五個字元開始，且字元之數目為六，因此位置資訊為「25至30」。

回應於來自用戶端器件200之搜尋請求，搜尋執行時期器530自索引525搜尋文件。在本文中，輸入「db2jcc.jar」作為搜尋字。搜尋執行時期器530搜尋包括此「db2jcc.jar」之文件，且將搜尋結果傳遞至文件分群單元550。文件分群單元550將各自具有一包括「db2jcc.jar」且具有相同雜湊值的區塊的文件分群至一個群組中。在此實施例中，由於文件1、3及4具有「11111111」之相同雜湊值，因此文件分群單元550將此等文件分群至同一群組中。由於在文件2中包括「db2jcc.jar」之區塊具有「33333333」之不同雜湊值，因此文件分群單元550將文件2分群至不同群組中。

文件分群單元550將經受分群之搜尋結果傳回至搜尋執行時期器530，且搜尋執行時期器530將搜尋結果發送至用

戶端器件200。圖11A及圖11B說明經受分群之搜尋結果的一實例。如一看搜尋結果便可見，屬於同一群組之第二文件及後續文件經縮排。在圖11中，將文件1、3及4分群至同一群組中，且將文件2分群至不同群組中。

根據本發明，將一待搜尋之文件劃分成複數個區塊，基於每一區塊中所包含之一字元串計算一雜湊值，且與區塊之位置資訊相關聯而儲存該經計算之雜湊值。因此，記憶體使用量增加達對應於雜湊值及位置資訊之儲存的量。記憶體使用量之顯著增加將大大降低處理器之處理速度。

因此，吾人研究記憶體使用量增加了多少。包括11,830個經儲存文件(電子郵件)及512,127個句子之一郵件語料庫(corpus)用作一資料源。在句子基礎上執行文件劃分，每一雜湊值具有8位元組之長度，且一表示自一文件之前置符記至一句子之前置符記的次序的符記數目及一表示自該文件之前置符記至該句子之結束符記的次序的符記數目用作位置資訊。

在此等條件下，在本發明中，用於儲存一索引之記憶體使用量在僅儲存該索引而不儲存雜湊值時為93,995,008個位元組，且在除該索引外還儲存雜湊值時為98,820,096個位元組。此意謂每一句子增加9.42個位元組，且記憶體使用量僅增加約5%。因此，可認為記憶體使用量並未大大增加，使得處理器之處理速度不受影響。

只要可自待搜尋之文件擷取本文，該等文件便可為任何文件，包括正文檔案、辦公室文件、電子郵件及其類似

者。此處注意，只要文件具有相同的所擷取之本文及劃分資訊，那麼即使在該等文件具有不同格式時，亦有可能判定該等文件是否彼此相關。因此，必須以相同方式執行劃分成區塊。此係因為，劃分之不同方式引起在對相關文件之判斷上的改變。

舉例而言，搜尋系統針對每一文件必須具有之資訊包括組成文件之以上所陳述的符記串及指示將如何劃分文件之劃分資訊，以及文件之識別資訊(例如，文件編號)及待包括於一雜湊值中之字元資訊。剖析器510接收符記串及文件之識別資訊，劃分單元535保存劃分資訊，且計算單元540保存待包括於雜湊值中之字元資訊。

在建立索引時所儲存及用於搜尋之資訊可包括區塊之雜湊值及位置資訊，以及文件之識別資訊。儲存單元545儲存此資訊，且文件分群單元550讀取此資訊。

雖然已詳細描述了本發明之搜尋系統及由該搜尋系統所執行之搜尋方法，但本發明不限於以上所描述之實施例，且另一實施例、添加、改變及刪除皆為可能的，只要其在對於熟習此項技術者顯而易見的範圍內。任何實施例將在本發明之範疇內，只要可自該實施例獲得本發明之效應。因此，本發明可經組態為可由電腦讀取之程式，且本發明可藉由使電腦執行該程式而體現為搜尋系統。該程式可藉由將其儲存於記錄媒體中而提供。

【圖式簡單說明】

圖1A至圖1C說明待搜尋之三個文件。

圖2A至圖2B說明用於圖1中所說明之文件的反向索引及用以保存收集之文件之資料結構的實例。

圖3說明習知搜尋引擎之例示性組態。

圖4例示一網路系統，其包括保存待搜尋之文件的一資料源、發出搜尋請求之一用戶端器件，及包括接收該搜尋請求及執行搜尋處理之一搜尋引擎的一伺服器器件。

圖5說明搜尋器件之例示性硬體組態。

圖6為展示伺服器器件經組態為搜尋系統的功能方塊圖。

圖7A至圖7C說明待搜尋之三個文件。

圖8A至圖8C說明雜湊函數應用於經劃分區塊中之每一者中所包括之一字元串且計算每一區塊之雜湊值的狀態。

圖9A至圖9B說明作為搜尋結果之經分群之文件。

圖10A至圖10D說明電子郵件之四個實例，該等實例中之每一者經劃分成多個區塊，其中雜湊值與位置資訊彼此相關聯。

圖11A至圖11B說明經受分群之例示性搜尋結果。

【主要元件符號說明】

10	搜尋引擎
11	資料庫
12	耙梳程式
13	剖析器
14	儲存器
15	索引器

16	索引
17	搜尋執行時期器
18	查詢相關資訊產生器件
19	查詢相關資訊比較器件
20	資料源
30	用戶端器件
100	資料源/資料庫
200	用戶端器件
300	伺服器器件
310	記憶體
320	處理器
330	記憶體控制單元
340	頻道子系統
350	控制器
360	輸入/輸出器件
400	網路
500	耙梳程式
505	資料庫
510	剖析器
515	儲存器
520	索引器
525	索引
530	搜尋執行時期器
535	劃分單元

540	計算單元
545	儲存單元
550	文件分群單元

201027375

發明專利說明書

（本說明書格式、順序及粗體字，請勿任意更動，※記號部分請勿填寫）

※ 申請案號： 98133263

※ 申請日： 98.9.30 ※IPC 分類：~~G06F~~ G06F 17/30 (2006.01)

一、發明名稱：（中文/英文）

搜尋系統，搜尋方法及程式

SEARCH SYSTEM, SEARCH METHOD AND PROGRAM

二、中文發明摘要：

提供易於在一搜尋結果之文件當中尋找真正所需之一文件的一種搜尋系統及一種搜尋方法。此搜尋系統包括：一劃分單元，其根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊；及一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊，而自該儲存單元545提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。

三、英文發明摘要：

The present invention provides a search system and a search method to make it easy to find out a document required truly among documents of a search result. This search system includes a division unit that divides a document to be searched into a plurality of blocks in accordance with designated division information, a calculation unit that calculates a hash value of each block by applying a hash function to a character string included in each block, a storage unit that stores the calculated hash value together with positional information on the block in the document, and a document grouping unit that fetches, for each document obtained by searching based on the search word, a corresponding hash value from the storage unit 545 in accordance with positional information on a block including the search word to group documents having the same hash value into one group and output the grouped documents as the search result.

七、申請專利範圍：

1. 一種搜尋系統，其基於一輸入搜尋字而搜尋文件且輸出一搜尋結果，該搜尋系統包含：

一劃分單元，其根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；

一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；

一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊；及

一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊而自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。

2. 如請求項1之搜尋系統，其中該劃分單元根據該劃分資訊以下列方式中之至少一者劃分該文件：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分。

3. 如請求項1之搜尋系統，其中該文件包括一符記串，在該符記串中複數個字或符記經循序排序，且關於每一區塊之該位置資訊包括自該文件中之一前置符記至每一區塊之一前置符記的一符記次序。

4. 如請求項1之搜尋系統，其中關於每一區塊之該位置資訊包括自該文件中之一前置字元至每一區塊之一前置字

元的字元之數目。

5. 如請求項1之搜尋系統，其中當該區塊中所包括之一字元串包括一指定字元類型時，該計算單元藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值。

6. 如請求項1之搜尋系統，其中該文件分群單元包括一排序單元，該排序單元根據一搜尋記分對包括於一群組中之複數個文件進行排序。

7. 一種搜尋方法，其由基於一輸入搜尋字而搜尋文件且輸出一搜尋結果之一搜尋系統執行，該方法包含以下步驟：

根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；

藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；

將該經計算之雜湊值與關於該文件中之該區塊的位置資訊儲存於一儲存單元中；及

針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊而自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。

8. 如請求項7之搜尋方法，其中該劃分步驟、該計算步驟及該儲存步驟係在建立由該搜尋系統在一搜尋期間所使用之一索引時執行，且該輸出步驟係在該搜尋時執行。

9. 如請求項7之搜尋方法，其中在該劃分步驟中，該文件係根據該劃分資訊以下列方式中之至少一者劃分：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分。

10. 如請求項7之搜尋方法，其中該文件包括一符記串，在該符記串中複數個字或符記經循序排序，且關於每一區塊之該位置資訊包括自該文件中之一前置符記至每一區塊之一前置符記的一符記次序。

11. 如請求項7之搜尋方法，其中關於每一區塊之該位置資訊包括自該文件中之一前置字元至每一區塊之一前置字元的字元之數目。

12. 如請求項7之搜尋方法，其中在該計算步驟中，當該區塊中所包括之一字元串包括一指定字元類型時，藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值。

13. 如請求項7之搜尋方法，其中該輸出步驟包括根據一搜尋記分對包括於一群組中之複數個文件進行排序的步驟。

14. 一種電腦可讀程式，其使一搜尋系統執行一搜尋方法，該搜尋系統基於一輸入搜尋字而搜尋文件且輸出一搜尋結果，該方法包含以下步驟：

根據指定劃分資訊將一待搜尋之文件劃分成複數個區塊；

藉由將一雜湊函數應用於每一區塊中所包含之一字元

串來計算每一區塊的一雜湊值；

將該經計算之雜湊值與關於該文件中之該區塊的位置資訊儲存於一儲存單元中；及

針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊而自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果。

15. 如請求項14之程式，其中該劃分步驟、該計算步驟及該儲存步驟係在建立由該搜尋系統在一搜尋期間所使用之一索引時執行，且該輸出步驟係在該搜尋時執行。

16. 如請求項14之程式，其中在該劃分步驟中，該文件係根據該劃分資訊以下列方式中之至少一者劃分：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分。

17. 如請求項14之程式，其中在該計算步驟中，當該區塊中所包括之一字元串包括一指定字元類型時，藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值。

18. 如請求項14之程式，其中該輸出步驟包括根據一搜尋記分對包括於一群組中之複數個文件進行排序的步驟。

19. 一種搜尋系統，其基於一輸入搜尋字而搜尋文件且輸出一搜尋結果，該搜尋系統包含：

一劃分單元，其基於指定劃分資訊將一待搜尋之文件劃分成複數個區塊，依據該指定劃分資訊以下列方式中

之至少一者劃分該文件：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分；

一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之一字元串來計算每一區塊的一雜湊值；

一儲存單元，其儲存該經計算之雜湊值與關於該文件中之該區塊的位置資訊，該文件包括一符記串，在該符記串中複數個字或符記經循序排序，關於每一區塊之該位置資訊包括自該文件中之一前置符記至每一區塊之一前置符記的一符記次序；及

一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊而自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果，

其中

當該區塊中所包括之一字元串包括一指定字元類型時，該計算單元藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值，且

該文件分群單元包括一排序單元，該排序單元根據一搜尋記分對包括於一群組中之複數個文件進行排序。

20. 一種搜尋系統，其基於一輸入搜尋字而搜尋文件且輸出一搜尋結果，該搜尋系統包含：

一劃分單元，其基於指定劃分資訊將一待搜尋之文件

劃分成複數個區塊，依據該指定劃分資訊以下列方式中之至少一者劃分該文件：劃分成每一句子；劃分成每一段落；在一空線處劃分；及基於經添加至該文件之額外資訊而劃分；

一計算單元，其藉由將一雜湊函數應用於每一區塊中所包含之來計算每一區塊的一雜湊值；

一儲存單元，其儲存該經計算之雜湊值與關於該區塊的位置資訊，該位置資訊包括自該文件中之一前置字元至每一區塊之一前置字元的字元之數目；及

一文件分群單元，其針對藉由基於該搜尋字而搜尋所獲得的每一文件，根據關於包括該搜尋字之一區塊的位置資訊自該儲存單元提取一相應雜湊值，以將具有相同雜湊值之文件分群至一群組中且輸出該等經分群之文件作為該搜尋結果，

其中

當該區塊中所包括之一字元串包括一指定字元類型時，該計算單元藉由將一雜湊函數應用於已排除該字元類型之一字元串而計算一雜湊值，且

該文件分群單元包括一排序單元，該排序單元根據一搜尋記分對包括於一群組中之複數個文件進行排序。