申請日期：	94.3.14	案號：	94104602
類別：	G06F 17/21		

（以上各欄由本局填註）

公告本

發明專利說明書

518480

一、發明名稱	中文	用以標準化電子文件內文字資訊之方法
	英文	METHOD OF STANDARDIZING CHARACTER INFORMATION IN ELECTRONIC DOCUMENTS
二、發明人	姓名（中文）	1. 中居 治彥 2. 木戶 彰夫 3. 榎木 義彥 4. 織田 哲治
	姓名（英文）	1. HARUHIKO NAKAI 2. AKIO KIDO 3. YOSHIHIKO ENOMOTO 4. TETSUJI ORITA
	國籍	1. 日本 2. 日本 3. 日本 4. 日本
	住、居所	1. 日本國神奈川縣川崎市麻生區白山2-2-1-414 2. 日本國東京都豐島區西池袋2-35-10 3. 日本國神奈川縣津久井郡城山町中沢595-4 4. 日本國神奈川縣橫濱市綠區十日市場854-16-304
三、申請人	姓名（名稱）（中文）	1. 美商萬國商業機器公司
	姓名（名稱）（英文）	1. INTERNATIONAL BUSINESS MACHINES CORPORATION
	國籍	1. 美國
	住、居所（事務所）	1. 美國紐約州阿蒙市新果園路
	代表人姓名（中文）	1. 傑拉德 羅森賽
	代表人姓名（英文）	1. GERALD ROSENTHAL

本案已向

國(地區)申請專利	申請日期	案號	主張優先權
日本 JP	1999/07/23	特願平11-209094	有

有關微生物已寄存於　　寄存日期　　寄存號碼

無

五、發明說明（1）

[發明範圍]

本發明與透過以來自對應的標準字型組的字元替代在電子文件中利用一非標準字型組的字元，以標準化電子文件中字元資訊的方法有關。

[發明背景]

照慣例，一電子文件中使用的字型之選擇委任給產生此一文件的人。安裝在電子文件處理設備-例如文書處理器等-中的字型，也因機器而有所不同，且此等機器通常受限於只能處理一特定語言。想要產生含有許多語言的文件、或想要使用未包括在一基本字型上的字元之文件生產者，因此必須定義此等字元的字型為外部字元，以在一電子文件中使用此等字型。這不是於列印在紙上的文件之交換期間的問題，而已變成對網際網路上電子文件的交換之激增、和在電子圖書館中電子文件的註冊期間的主要缺點。

電子文件的生產者和讀者雙方都必須有相同的字型組和字元碼，以便可靠地交換字元資訊。然而，考慮目前在每個平台上可使用的字型組不同的情形，在資訊的交換中所使用的格式，例如在網際網路線上傳遞的格式、和在一電子圖書館當中或在一公司當中儲存在集中的檔案中之資料的格式，用作字元資訊的標準字型之標準化是必需的。

[本發明解決的問題]

字型替代在有關的電子文件產生系統中已經可實行，但在此替代中，字元碼資訊儲存為就像以另一字型替代的字

五、發明說明（2）

型資訊。舉例來說，外部字元字型通常定義為獨立的字型，然後它通常從字元定義的順序決定字元的索引。因此，即使當使用了大的字型組例如，舉例來說包括來自全世界的所有基本字元之Unicode字型（包括未被一般的電子文件產生系統所支援，遍及數千字的日本工業規格補充漢字組），因為字元在字型（字元編碼）當中的索引不同，字型替代不能施行。

使用者將必須在更換字型時，手動地改變字元碼在一電子文件當中的編碼值。為了要達成這個，使用者將需要知道使用在最初的電子文件中之字型索引，和對應於要替代的那些字元之字元的索引。當電子文件積聚在一電子圖書館中時，所積聚文件的生產者之數目是一持續增加的大數目，而使用在所有文件中之字型組的儲存、和在字型組當中字元的索引之儲存，使得可以一次一個手動地標準化的那些文件在實務上不可能。

結果，在相關技術中處理電子文件的字元資訊之電子圖書館和在公司當中集中的檔案，唯一的選擇是如文件所製作一樣積聚它們，基本上放棄嘗試標準化電子文件。因為生產者和使用者之間字型環境的差異，因而出現無法瞭解的字元。這在交換電子文件時造成不便，且表示由其他系統所做的電子文件之顯示和處理，不能由限制於第0層（Tier-0）來源等的系統施行。日本專利權公告出版第Hei. 7-319854號中，揭露了一種外部字元管理系統，以用一種有效率的方式製作和散佈外部字元字型檔案。然而，這種

五、發明說明 (3)

技術是用來管理封閉網路環境中的外部字型，且不能應用在本發明的目標之電子文件中字元資訊的標準化。

因為本發明要解決上述問題，因此本發明的一個目標在於提供一種標準化出現在電子文件中字元資訊的方法，能夠標準化使用在電子文件的資訊之累積和轉換中的字型，使用對每個平台或電子文件產生系統不同的多種字型，而不損害資訊的品質。

[解決問題的方法]

本發明與藉由以來自一對應的字型組之字元取代電子文件中利用一非標準字型組的字元，標準化電子文件中的字元資訊之方法有關，也就是本發明是一種標準化電子文件中字元資訊的方法，包含步驟：藉由比較在一電子文件中使用的字型、和要提供作為替代字型的目標字型組當中的字型，自動地產生在實際字型替代期間引用的字型比較表；提出一自動地產生的字型比較表給使用者，並讓使用者修正比較表中的錯誤；和根據修正的字型比較表，實際地更換電子文件中的字型。

依照本發明，使用外部字元轉換電子文件為一標準字型組例如，舉例來說，Unicode字型，和含有一些外國語言片段的電子文件之轉換是可能的，和相似字元與外國語言文件的資訊之轉換和累積是可能的。

在本發明一較佳的範例中，一電子文件由一來源、電子文件中使用的一字型組、供施行標準化的一目標字型組、在一先前的轉換中產生的一比較表、和描述限制字元比較

五、發明說明（4）

的物件之一規則組的字型物件資訊構成，而在自動地產生字型比較表的步驟中，與對映每一漢字部首有關的一規則組被輸入，一字型比較表候選清單被輸出。與相似字元之間的對映有關的加權資訊，也可輸出當一參考檔案。字型比較表候選清單是一做為元件群組的清單，包含來源字型當中的一個字元，和在一與來源字型相容的目標字型當中的多個字元。優先次序程度資訊可增加給目標字型當中的多個字元。字型比較表可以是一清單，用來當作在一群來源字型組和這個來源字型組當中的字元碼、和一群目標字型組和這個目標字型組當中的字元碼之間一對應關係的元件。這些情況中的任何一種可應用於自動地產生字型比較表的步驟。

在本發明一較佳的範例中，在自動地產生字型比較表的步驟中的字型比較，可自動地使用光學字元識別(OCR)技術實施。此外，修正字型比較表中的錯誤之步驟可以是一程序，其中顯示每一項目的字型比較表的候選清單，且使用者從候選清單選擇一字元。描述來源電子文件結構的字型比較表和一規則組被輸入，而來源電子文件中使用的字型和字元碼之標準化，可在字型替代步驟中施行。要提供做為替代的字型組可以是Unicode字型的一個字型組。本發明適合應用在這些情況中的任何一種。

[具體實施例]

圖1是說明本發明標準化一電子文件中字元資訊的方法之概念的流程圖。本發明現在將依照圖1描述。首先-藉由

五、發明說明 (5)

比較在電子文件中使用的字型、和字型組當中將要替代的字元(字型)，以便製作字型比較表的一個候選清單-，執行一字型比較表自動產生步驟，以自動地產生在實際字型替代期間參考的一個字型比較表。接著，自動地產生的字型比較表提出給使用者，使用者執行一字型比較表修正階段，以修正字型比較表中的錯誤，和製作一新的字型比較表。最後，根據修正過的新字型比較表，執行一字型替代階段，以實際地替換電子文件中的字型，和獲得一標準化的電子文件。

可利用本發明標準化電子文件中字元資訊的方法之範圍有電子圖書館、文件管理系統、中間伺服器支援的手持裝置(普遍的計算環境)例如PDAs、網路出版、和網瀏覽器等。現在考慮，在電子文件中包括由一使用者所製作的外部字元標準化成一Unicode字型的情況。在此情況，使用者不只標準化由使用者所獨立地定義的那些外部字元到Unicode之內，而且也標準化那些一般字元到Unicode之內是必需的。關於一般字元，當電子文件的字型間存在之字型索引比較表已經由使用者製作，舉例來說，在微軟mincho和Unicode字型之間，標準化可根據這個比較表以直接的方式施行。

本發明標準化電子文件中字元資訊的方法，也可利用來做外部字元的標準化。首先，對每一外部字型執行字型比較表自動產生階段，而對每一外部字型獲得一相符或相似的Unicode字型，以便字型比較表的一個候選清單暫時地

五、發明說明 (6)

做為字型比較表。字型比較表的候選清單通常是每一外部字元有多個Unicode字型。然後，以提供給使用者的字型比較表修正清單，執行字型比較表修正階段，而使用者修正比較表中的錯誤，也就是使用者從候選清單選擇一字型，而當沒有Unicode字型對應到一外部字元時，分配一相似的Unicode字型，而當沒有對應的字型時註記為一Unicode字型外部字元。Unicode也支援數千個日本工業規格補充漢字，以便由使用者所做的幾乎所有外部字元可對應到Unicode字型。最後，根據校訂的字型比較表執行實際地更換電子文件中的字型之字型替代階段，而可獲得標準化為Unicode字型的電子文件。

每個階段的細節描述如下。

(1)字型比較表自動產生階段：

在這個階段中，與其為來源文件的電子文件有關的一規則組、在這電子文件中使用的字型組、實行標準化的目標字型組、在其後的轉換中製作的一比較表、限制字元比較的物件和對每一漢字部首的對映之規則組("一鍵部首"和"二鍵部首"，漢字的某些部份是否分開，和決定一個字元是否判定為相同部首或一不同部首、或不同字元的其他外觀)，被輸入當成上述字型比較指定資訊，並輸出一字型比較表候選清單。在本發明的較佳實施中，在這個階段的執行中評估的相似字元間對映的加權資訊，被輸出當做一參考檔案，且可在下一執行期間參考。

字型比較表是一清單，其用在一群來源字型組和這個字

五、發明說明（7）

型組（字型索引）中的字元碼、和一群目標字型組和在這個字型組當中對應字元的字元碼之間的對應關係當作元件。字型比較指示資訊包含資訊：指定表示哪一來源的哪一字型組對應到哪一目標的哪一字型組的字型群組、和在一來源字型組當中哪一字元要接受比較、和構成目標字型當中的比較之一字型群組。字型比較表候選清單是一清單，將來源字型當中一字元的群組、和可對應於此一字元的目標字型當中的多個字元當成元件。在本發明的較佳實施中，優先次序層次資訊加入到目標字型中的字元，以便在下一個階段中輔助手動地定義字型比較表。

字元比較使用光學字元辨識（OCR）技術透過下列程序實行。

①目標字型組製作要比較的字元群組的樣式。

②從電子文件挑選一個字元，並檢查它的碼值。

③如果是樣式資訊比較的目標之字元的碼值；

A. 從來源字型組製作這個字元的樣式。

B. 把所獲得的樣式與在①中所獲得群組的樣式比較，並加入相似樣式的一個群組到比較表候選清單。在此較佳實施中，在候選清單當中的優先次序層次資訊在這個時間加入。

④重複②和③中的程序。

（2）字型比較表修正階段：

在這個階段中，在先前階段中所獲得字型比較表的候選清單、和這個階段的先前執行所獲得的結果之字型比較

表，被輸入且一最終字型比較表被輸出。這個階段主要是一程序，其中在先前階段中所獲得的字型比較表之候選清單顯示為每一項目，然後供使用者從清單中選擇一個。在本發明的較佳實施中，當使用者所選擇的候選與先前處理中所製作的字型比較表中一項目矛盾時，或當使用者指定多對一或一對多對映時，這個階段的處理系統使用一警告告知使用者，以便使用者能再考慮。此外，本發明的較佳實施也可具有一功能，在呈現字型比較表的候選清單給使用者時，顯示候選字元的屬性（字元名稱、字元意義、字元類型、和可在所敘述字元中參考的其他資訊）。

（3）字型替代階段：

先前階段輸出的字型比較表、描述電子來源文件的結構之規則組、和描述目標電子文件的結構之規則組（其可能與來源相同的格式）被輸入，而在來源電子文件中使用的那些字型和字元碼被標準化。當在來源和目標之間電子文件的格式和結構不一致時，電子文件格式的轉換也可同時地在這個階段中實行。

如在本發明的目的中所陳述，在上面詳細地描述的本發明標準化電子文件中字元資訊的方法，除了能夠標準化-使用因每個平台或電子文件產生系統而不同的各種字型-使用在電子文件的資訊之累積和轉換中的字型，而不損害資訊的品質，也可用來達成下列描述的各種目標。

（1）使用各種字元碼的電子文件之字元碼，可透過轉換成普遍地了解的字元碼Unicode，輕易地製作成遵循網際

五、發明說明（9）

網路標準的文件描述語言例如XML，而然後電子文件可在網路上出版。

（2）電子圖書館中和公司中集中的檔案之資訊檢索的效率，可藉由標準化使用各種字元碼的電子文件之字元碼成為特定的字元碼而改良。

（3）藉由反向地轉換使用標準化的字元碼和字元字型累積和交換的電子文件，成為此用戶環境獨特的字元碼和字型，即使在在沒有資訊交換中所使用的字型的環境中，電子文件可使用相似的字元顯示。

（4）藉由反向地轉換使用標準化的字元碼和字元字型累積和交換的電子文件，成為對用戶環境獨特的字元碼和字型，用戶環境中資訊處理的效率可增加。

（5）在字元碼和字型的標準化期間參考的比較表之半自動化的結果，使用者工作量可大大地減少，而標準化那些文件所必需的工作量可減少到一更行的層次。

（6）舉例來說，當一第0層系統中有一些資源（當有一些字型或當沒有轉換表和函數，以便用一有效的方式處理來自一原始電子文件的字元碼之文件成為這個系統的字元碼）要顯示和處理在另一系統中製作的文件時，改變電子文件的格式成為可在一用戶環境中處理的格式，是藉由將要最佳化成有一些資源的用戶環境之文件的標準化，交託給此電子文件的存取期間之存取路徑中的一個中間伺服器。

（7）字元碼和字型的比較表之自動產生的效率，可藉由

五、發明說明（10）

替換字型和字元碼增進，其在相關的技術中，當參考一先前標準化的各別文件時只能獨立地替換。

（8）在對映時犯錯的可能性，可藉由使用過去的對映經驗、藉由自動地產生字元碼、和在參考一過去標準化的各別文件時使用一字型對映表減少。下列情況可視為對映錯誤的因素。

·當在目標字型組中存在形式相似的多個相似字元時，有可能對映錯誤的字元或在字型的對映中可能出現不一致。

·當在來源字型組中存在形式相似的多個相似字元時，有可能多於一個字元對映一個目標字元。

（9）在對映已經在先前定義的來源和目標字型組當中字元（字型）的比較，可藉由敘述自動地實行的字元比較之字型群組避免。

（10）由於在對映已經在先前定義的來源和目標字型組當中字元（字型）的比較，可藉由敘述自動地實行的字元比較之字型群組避免，使用者不打算產生而產生（舉例來說，一日本工業規格X0208層次1日本工業規格字元被對映為一日本工業規格X0208層次2字元）的比較表之危險減少。

（11）對映的準確性由於不同字體的字型之比較而減少範圍，可藉由定義包含在來源中比較特定字型的目標字型而降低。

（12）藉由註記語言和語言所使用的字型組之間的關係，在比較表的自動產生期間引入語言學規則是可能的。這

五、發明說明 (11)

樣，藉由比較字元資訊所獲得、其為轉換的目的之字元，被當成與先前和後續字元相連接的單字，而可增進比較表的自動產生之準確性。

[本發明的優點]

如從上面的描述而變得顯而易見的，依照本發明，轉換使用外部字元的電子文件成為一標準字型組例如，舉例來說，Unicode字型、和轉換含有一些外國語言片段的電子文件是可能的，且相似字元的資訊和外國語言文件的轉換和累積是可能的。

圖式概述

圖1是一流程圖，舉例說明本發明標準化出現在電子文件中字元資訊的方法之概念。

四、中文發明摘要　（發明之名稱：用以標準化電子文件內文字資訊之方法）

[目的]

提供一種標準化出現在電子文件中字元資訊的方法，能夠標準化電子文件的資訊之累積和交換中使用的字型，使用對每一平台或電子文件生產系統不同的各種字型，而不損害資訊的品質。

[構成]

一種標準化電子文件中字元資訊的方法，包含步驟：

藉由比較在一電子文件中使用的字型、和要提供作為替代字型的目標字型組當中的字型，自動地產生在實際字型替代期間引用的字型比較表；

提出一自動地產生的字型比較表給使用者，並讓使用者修正比較表中的錯誤；和

英文發明摘要　（發明之名稱：METHOD OF STANDARDIZING CHARACTER INFORMATION IN ELECTRONIC DOCUMENTS）

[Object] To provide a method of standardizing character information occurring in electronic documents that is capable of standardizing fonts used in accumulation and exchange of information for electronic documents made using a variety of fonts that differ for each platform or electronic document production system without detriment to the quality of the information.

[Constitution] A method of standardizing character information in electronic documents

四、中文發明摘要　(發明之名稱：用以標準化電子文件內文字資訊之方法)

根據修正的字型比較表，實際地更換電子文件中的字型。

英文發明摘要　(發明之名稱：METHOD OF STANDARDIZING CHARACTER INFORMATION IN ELECTRONIC DOCUMENTS)

comprising the steps of:

automatically generating a font comparison table referred to during actual font replacement by comparing a font used in an electronic document and a font within a target font set to be provided as a replacement font;

presenting an automatically generated font comparison table to the user and having the user amend errors in the comparison table; and

actually replacing the font in the electronic

四、中文發明摘要　（發明之名稱：用以標準化電子文件內文字資訊之方法）

英文發明摘要　（發明之名稱：METHOD OF STANDARDIZING CHARACTER INFORMATION IN ELECTRONIC DOCUMENTS）

document based on the amended font comparison table.

六、申請專利範圍

1. 一種標準化電子文件中字元資訊之方法，包含步驟：

藉由比較在一電子文件中使用的字型、和要提供作為替代字型的目標字型組當中的字型，自動地產生在實際字型替代期間引用的字型比較表；

提出一自動地產生的字型比較表給一使用者，並讓該使用者修正該字型比較表中的錯誤；和

根據修正的字型比較表，實際地替換該電子文件中的該字型。

2. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中一電子文件由一來源、該電子文件中使用的一字型組、供施行標準化的一目標字型組、在一先前的轉換中產生的一比較表、和描述限制字元比較的物件之一規則組的字型物件資訊構成，而在該自動地產生該字型比較表的步驟中，與對映每一漢字部首有關的一規則組被輸入，一字型比較表候選清單被輸出。

3. 如申請專利範圍第2項標準化電子文件中字元資訊之方法，其中與相似的字元之間的對映相關之加權資訊被輸出當成一參考檔案。

4. 如申請專利範圍第2項標準化電子文件中字元資訊之方法，其中該字型比較表候選清單是一作為元件群組的清單，包含在一來源字型當中的一字元、和在與該來源字型相容的一目標字型當中的多個字元。

5. 如申請專利範圍第4項標準化電子文件中字元資訊之方法，其中對該目標字型當中的該等多個字元加入優先次

六、申請專利範圍

序程度資訊。

6. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中字型比較表是一清單，用來當作在一群來源字型組和這個來源字型組當中的字元碼、和一群目標字型組和這個目標字型組當中的字元碼之間一對應關係的元件。

7. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中在自動地產生該字型比較表的該步驟中之字型比較，是自動地使用光學字元識別(OCR)技術施行。

8. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中修正該字型比較表中的錯誤之該步驟是一程序，其中顯示每一項目的字型比較表的候選清單，且該使用者從該候選清單選擇一字元。

9. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中輸入一字型比較表和一描述來源電子文件的結構之規則組，而該來源電子文件中使用的字型和字元碼之標準化，在該字型替代步驟中施行。

10. 如申請專利範圍第1項標準化電子文件中字元資訊之方法，其中要提供作為替代的該字型組，是一單一碼(Unicode)字型的字型組。

圖式

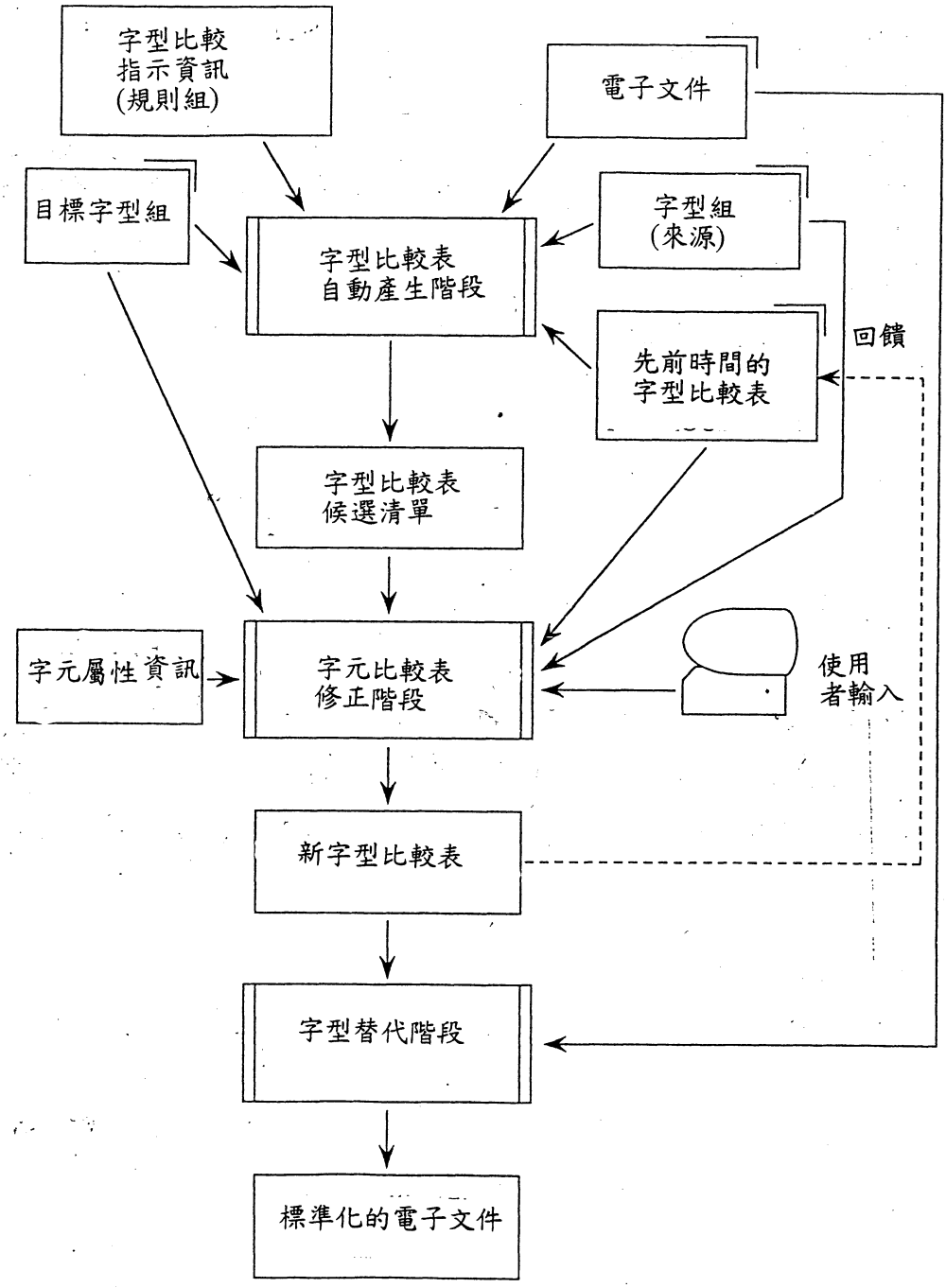
字型比較
指示資訊
(規則組)
電子文件
目標字型組
字型組
(來源)
字型比較表
自動產生階段
回饋
先前時間的
字型比較表
字型比較表
候選清單
字元屬性資訊
字元比較表
修正階段
使用
者輸入
新字型比較表
字型替代階段
標準化的電子文件

圖 1