以字串特徵做為文本資料之錯誤偵測

劉吉軒、鄭雍瑋
國立政治大學資訊科學系
jsliu@cs.nccu.edu.tw

摘要

資訊擷取是從自然語言文本中辨識出特定的主題或事件的描述，進而萃取出相關主題或事件元素中的對應資訊。然而資訊擷取的結果會有錯誤情況發生，若單只依靠人工的方式進行錯誤的檢查及更正，將會是耗費大量人力及時間的工作。在本論文中，我們提出一種字串特徵為主的錯誤偵測方法，以資料描述的概念進行字串外表特徵的捕捉與轉換，再透過 C4.5 或 SVM 機器學習分類方法，自動建構適當的二元資料分類模型，進而達到辨別正確與錯誤資料的目的。

實驗結果顯示，本研究所提出的錯誤偵測方法，可以有效偵測出資訊擷取成果中不正確的值組，確保高品質的資訊擷取成果產出，促使資訊擷取技術更廣泛的實際應用。

關鍵詞：錯誤偵測、資料描述、資訊擷取

1. 緒論

資訊擷取(Information Extraction)是自然語言處理中的一個技術，能從自然語言文本中辨識出特定主題或事件描述，進而抽取核心資訊所對應的文字資料，如人、事、時、地、物等，而將原始非結構化的資料轉換成結構化資訊，並彙整成資料庫，

提供進一步的資訊加值處理與應用的可能性。對於大部分的資訊加值應用而言，資料的正確性應是最基本的條件。目前大多數資訊擷取研究皆致力於提升其擷取系統效能及各領域的應用性，但即使是最新的資訊擷取技術的擷取輸出，仍然是包含有某種程度的錯誤。因此，在進行任何資訊加值的應用之前，必須先將擷取結果加以驗證並更正錯誤。一般而言，資訊擷取技術所產生的資料量是相當龐大的，可能包括數以萬計的值組。若要以人工方式進行錯誤偵測及更正，甚至還得比對原始文件做進一步確認，這將會是耗費大量人力及時間的工作。因此，資料驗證所需的成本在部分資訊擷取應用上是一大障礙。可惜的是，現今大部分資訊擷取研究中，較少學者去針對擷取結果，進行驗證、偵測及更正錯誤資料的探討。我們認為，若能針對資訊擷取成果發展出適當的錯誤偵測機制，以確保高品質的資訊擷取成果產出，將可促成資訊擷取技術更廣泛的實際應用。

從資料清理的觀點來看，資訊擷取所產出資料集合的基本特性，是描述某項主題的文本資料。這種特性不利於使用一般的數據分析方法，如統計、分類與分群演算法。文本資料通常得透過文法和語義分析，並且必須對該領域主題事先進行領域知識的定義。這種需特別去定義領域知識的方式，不但會增加開發成本，並且也限制了應用性。為了增強資訊擷取效能與應用性，必須發展較通用的文本資料描述技術。此外，在資料清理的相關領域中，較少學者去針對以中文文字資料進行清理、錯誤偵測等研究。因此，我們希望能夠發展一套針對中文資料錯誤偵測機制，不但能提昇資訊擷取的效用以外，更能對於以中文為主的資料清理技術有所幫助。

2. 相關研究

資料清理(data cleansing)是一種針對資料集合進行識別、移除錯誤與不一致資料，進而改善資料品質的技術 [4]。不論是單一資料來源的資料庫，或是異質性資料來源的資料倉儲，都可以透過資料清理的技術來改善其資料產出的品質。資料清理技術以資料整合分析與資料稽核為主，其目的在於對資料進行分析，取得正確資料的特徵與規則後，便可偵測出異常及矛盾的資料，而指出資料庫中錯誤與不一致的情形。資料描述 (data profiling) 和資料探勘 (data mining) 這兩種相關的技術，對於資料分析有很大的幫助 [3]。資料描述的重點在於描述各個屬性值的資訊，譬如資料型態、長度、資料範圍、值組頻率、空值組出現情況以及字串規則等。資料探勘則是從大量的資料屬性中挖掘出有價值的資訊，藉由統計及人工智慧的技術，將資料做深入分析，找出相關資料的特定規則。

Galhardas和Raman兩位學者分別提出許多相關的資料清理技術[1] [5]，可以解決屬性遺失值(missing value)、雜訊資料(noisy data)、資料完整性、資料一致性等問題，藉此提高資料品質。例如，處理屬性遺失值的方法有忽略法(ignore tuple)、填補法(unknown)、平均法(mean)及線性回歸法(linear regression)等。忽略法就是不理會此資料，或者將此種資料刪除；填補法則是遇到遺失值的屬性欄位補上一特殊的代表值，如unknown；平均法是將所有相關的屬性值加總除上筆數，用此數值填入遺失的屬性欄位；線性回歸法則是利用統計的方法來找出最合適的值填補遺失值。

對於雜訊資料的處理，則有儲存槽法(bin)、叢聚法(clustering)等。儲存槽法的作法是將一連續的資料分割成離散資料，如年齡可分割成20 歲以下、20-40 歲、40-60 歲、60 歲以上等範圍。叢聚法則有分割(partition)、階層(hierarchy)、密度(density)等技術，主要是利用資料間的相似程度來加以分類。資料的一致性則是針對同名異物或同物異名的問題，這種現象通常發生於資料來源多重時。例如「motherboard」，主機板或母板都可代表，此時為了資料的一致，需要選定一代表值。另外在選取資料時，到底那份資料較能完全的符合使用者的需求，則屬於資料完整性的問題。藉由本體論，可判斷資料所包含的特徵個數和本體論所定義的特徵個數比率，計算資料的完整程度，可提供資料選取的一個指標。

在某些資料庫的欄位中，正確值組皆為唯一性的資料，若欄位內有重複值組則屬於錯誤資料。因此，唯一性偵測 (uniqueness detection)技術[6]的目的，在於自動偵測出目標欄位之值組是否為唯一性，以判斷資料的正確與否。這種技術以七個欄位屬性特徵為主，包括 data type、attribute length、whether a default value is specified、whether null is permitted 、distinctness ratio、min/max data length ratio 與 order of distinctness ratio with its relation。從訓練資料中擷取出這七個屬性特徵後，再透過 C4.5 演算法訓練出決策樹，便可透過此決策樹判定該欄位是否為唯一性，達到偵測錯誤資料之目的。

本研究針對資訊擷取系統之輸出資料，提出一種錯誤資料的偵測方式，以篩選

出錯誤的資訊擷取結果，進而搭配人工的查驗與更正，完成資料清理的工作。資訊擷取的結果是文本中的部分字串資料，通常無法適用於處理數值資料的統計方法。另外，欄位中的正確值組通常也包含重複的值組，例如，許多人有同樣的職位，因此，也不適用於唯一性偵測的技術。我們採用資料描述(data profiling)的技術概念，針對中文字串，提出一組字串特徵，以描述資訊擷取出的字串值組，再藉由相同欄位中的正確值組與錯誤值組的描述差異，達到錯誤偵測之目的。

3. 資訊擷取之錯誤偵測

在資料庫領域的資料品質議題上，資料錯誤的情形通常是因為人工輸入疏失及多重資料來源的資料不一致。而資訊擷取結果的資料錯誤情形，則是資訊擷取技術本身在文本資料的辨識與選取上所發生的錯誤，其資料錯誤的形式並不一樣。從資料取得的觀點來看，資訊擷取是針對選定的主題，從大量文本中辨識主題描述的存在、萃取字串資料，建立各屬性(欄位)中的字串集合，最後彙整而成結構化的資料庫資料。在這資料庫中，每一筆資料代表從文本中辨識出的一個主題個體(subject instance)，該筆資料中每一個欄位的值組，則是該主題屬性在文本中所應對應的描述字串。資料的錯誤通常是因為主題辨識、選取字串、對應欄位時發生錯誤。

一個完美的資訊擷取結果資料庫，是經過資訊擷取技術正確而沒有遺漏的辨識、選取、與對應，最後完整的匯集了文本集合中的所有主題資訊。任何不是完美的資訊擷取結果資料庫，就是有錯誤的資料庫。我們將其錯誤或異常情形分為以下四種情況：

1. 遺失個體(missing entities)：資訊擷取技術無法從文本中辨識出存在的主題個體，而造成資料庫中缺失了該主題個體的整筆資料。
2. 遺失值組(missing values)：資訊擷取技術可以從文本中辨識出存在的主題個體，但對於部分的屬性，卻因辨識失誤而忽略了對應字串的選取，造成該屬性在資料庫中的值組缺失。
3. 重複個體(duplicates)：資料庫中存在描述同一主題個體的多筆資料。重複個體的發生，有可能是資訊擷取技術在資料對應輸出時產生錯誤，也有可能是由於在文本集合中，同一主題個體的描述多次出現。
4. 不正確值組(invalid values)：資訊擷取技術可以從文本中辨識出主題個體的存在，但對於部分的屬性，卻因辨識失誤，而在字串選取或對應至欄位時產生錯誤，造成此欄位中的值組是不符合原始文本描述的屬性資訊。

針對一個資訊擷取結果建立的資料庫而言，遺失個體的偵測幾乎是不可能的，因為在不比對文本集合的情形下，並沒有任何的資訊可以判斷該主題個體的缺失。遺失值組的偵測也有同樣的困難，除非在主題資訊的定義上，已知某一屬性為必定存在。在具備此資訊的情形下，遺失值組的偵測是相當容易而直接的。重複個體是主題個體的重複，可以關鍵屬性的值組重複偵測出來。以上三種資料錯誤或異常情形，在偵測上完全是直接可以或不可以，並不具備偵測方法的研究議題。相對的，不正確值組的偵測就必須分析、判斷欄位中每一個值組的正確性，分析方式與判斷的適當性決定偵測的準確性，其結果的好壞差距可能非常大。因此，我們以不正確

值組的錯誤偵測方法為研究目標。給定由資訊擷取結果所彙整的資料庫，對於每一屬性的值組集合，我們將透過資料描述的方法，以機器學習方式建立資料分析與判斷模型，再依此模型去分析、偵測出不正確值組。這些錯誤偵測技術將能有助於降低資訊擷取結果的人工檢驗成本，提升資訊擷取技術之加值應用可行性。

3.1 字串特徵

資訊擷取結果中的屬性值組，是從文本中選取的部份字串。如果這些部分字串是正確的屬性值組，它們應是對應到同一個主題元素，並且通常以一至多個字詞形式呈現，表達同種語意類別的資訊。我們提出一種資料描述的方式，以字串的外表形式上的特徵，做為區別字串類別的依據。我們假設同一個主題元素的正確屬性值組，會有類似的或接近的字串外表特徵。因此，我們可以依據字串特徵的相同或相異，來判斷屬性值組的正確或錯誤。

我們所定義的字串特徵是描述字串的外表特徵，而不考慮其文字意義。依據語言和主題領域的不同，字串可以在字元層級與字詞層級顯示出不同的外表特徵。我們將重點擺在字元層級以及中文字串。同樣的觀念，也可適用於字詞層級與其他語言文本。針對此目的，我們共定義出六個字串特徵：

1. string cardinality (以下簡稱 S_c)：字串中的字元個數。
2. string prefix (以下簡稱 S_p)：字串前 k 個字元，k 是可設定的參數。
3. string suffix (以下簡稱 S_s)：字串後 k 個字元，k 是可設定的參數。
4. string entity (以下簡稱 S_e)：字串的所有字元序列。

5. string numeral (以下簡稱 S_n)：字串是否包含代表數字的字元，輸出結果為 true or false。
6. string format (以下簡稱 S_f)：字串內容所屬的資料型態。

我們以 *SF* 代表六個字串特徵的集合：$SF = \{S_c, S_p, S_s, S_e, S_n, S_f\}$，透過 *SF* 可用來評估資料庫中每個屬性值組(v_i)，$SF(v_i) = (S_c(v_i), S_p(v_i), S_s(v_i), S_e(v_i), S_n(v_i), S_f(v_i))$。以人事異動主題中之單位欄位為例，假設值組內容為「台北市政府」，則其字串特徵為如表一所示。

表一、字串特徵範例

SF(台北市政府) *and* *k*=1					
S_c	S_p	S_s	S_e	S_n	S_f
5	台	府	台北市政府	false	string

3.2 字串特徵之數值轉換

如前所述，我們對於不正確值組的偵測是建立於其異常字串特徵的假設上。進一步的說，我們假設一個欄位中的正確值組會有相同的或相當類似的字串特徵。因此，正確值組的字串特徵會經常出現而甚為普遍及常見。相對的，如果一個值組的字串特徵是少見的，就代表其字串特徵是異常的，也就可能是不正確的值組。這個假設是基於統計學上的多數法則(majority rule)。對於一個欄位中的值組集合，我們對字串的每一個特徵，計算每一個特徵值出現的百分比，再以其百分比轉換成一個適當的數值，做為後續判斷的依據。在資訊擷取結果的資料庫中，任何一個值組會具有六個字串特徵之轉換數值。理想上，一個正確值組的六個字串特徵值在該欄位的值組集合中都是相當常見的，而具備六個較大的特徵轉換數值。如果一個值組的數個特徵轉換數值都是較小的，代表其字串特徵值在該欄位的值組集合中都是較少見的，可以推論其可能的異常或錯誤。

首先，我們定義 $S_j, j \in \{1,2,3,4,5,6\}$,是上一小節中 SF 中各個字串特徵，資料庫中每個值組 v_i 之各別特徵值為 $S_j(v_i)$，其所占的百分比為 $P_{rob}(S_j(v_i))$，轉換成的數值為 $S_j'(v_i)$。要將特徵值出現的百分比轉換成一個適當的數值，有許多可能的方式。在本研究中，我們提出兩種相當直接的字串特徵數值轉換方式，都是將特徵值出現的百分比對應到一個固定等份 w 的區間，而 w 是可設定的參數。我們以 $T(w)$ 代表一個對應函數，將百分比對應到等份依序排列之區間數值，也就是百分比乘以 w 的結果取整數再加 1，但最大不超過 w。譬如 $T(10)$ 便是當百分比值為介於 0% 與 10% 之間時，對應的轉換數值為 1、百分比值介於 10% 與 20%之間時，對應的轉換數值為 2，依此類推到特徵轉換數值最大為 10。

第一種方式為個別百分比轉換，每一個字串特徵值依其出現的百分比個別的進行轉換，特徵轉換數值 $S_j'(v_i)$ 公式為：$S_j'(v_i) = P_{rob}(S_j(v_i)) \cdot T(w)$。第二種方式為累計百分比轉換，先將每一個字串特徵值依其出現的百分比由小到大排列，累加其百分比值後，再進行轉換。以 G 代表特徵值所佔百分比不大於 $P_{rob}(S_j(v_i))$ 的群組，特徵轉換數值 $S_j'(v_i)$ 的公式為：$S_j'(v_i) = \sum_{\forall i \in G} P_{rob}(S_j(v_i)) \cdot T(w)$。

我們以中文姓名屬性值組和字元個數特徵為例說明。假設資料庫中姓名屬性所有值組的字元個數的集合為{1, 2, 3, 4, 5, 6}，各自所佔的百分比為{1.5%, 11%, 79%, 5%, 3%, 0.5%}。假設 w 參數為 10，個別百分比轉換方式得到的結果如表二所示，其中字元個數為{1, 4, 5, 6}的百分比都是介於 0% 與 10% 之間，分別轉換之後的轉換數值都是 1，字元個數為{2}的百分比都是介於 10% 與 20% 之間，轉換之後的轉換數值是 2，字元個數為{3}的百分比都是介於 70% 與 80% 之間，轉換之後的轉換數值是 8。累計百分比轉換方式得到的結果如表三所示，其中字元個數為{6}的百分比最小，轉換之後的轉換數值是 1。接著是字元個數為{1}的百分比，累計字元個數為{6,1}的百分比之後的轉換數值仍是 1。再來是字元個數為{5}的百分比，累計字元

個數為{6,1,5}的百分比之後的轉換數值仍是 1。再來是字元個數為{4}的百分比，累計字元個數為{6,1,5,4}的百分比之後的轉換數值為 2。再來是字元個數為{2}的百分比，累計字元個數為{6,1,5,4,2}的百分比之後的轉換數值為 3。最後是字元個數為{3}的百分比，累計字元個數為{6,1,5,4,2,3}的百分比之後的轉換數值為 10。

表二、個別百分比轉換

$S_c(v_i)$	$P_{rob}(S_c(v_i))$	$S_c'(v_i)$
6	0.5 %	1
1	1.5 %	1
5	3 %	1
4	5 %	1
2	11 %	2
3	79 %	8

表三、累計百分比轉換

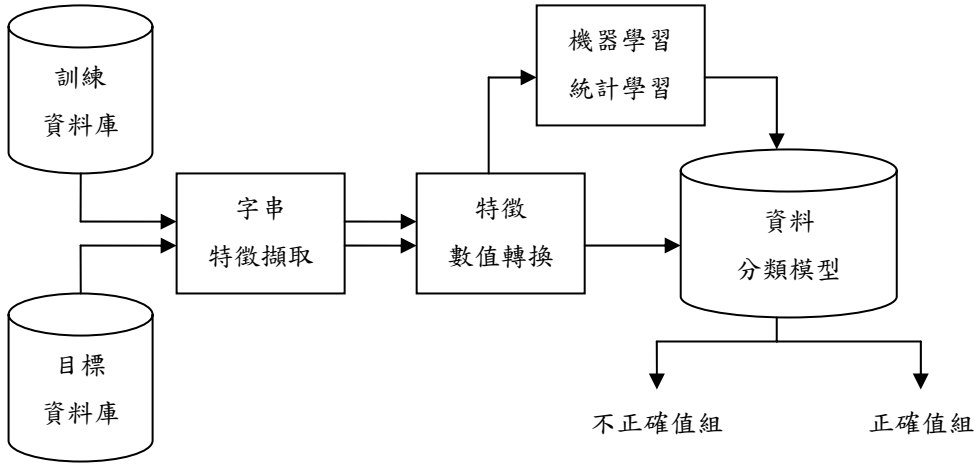
$S_c(v_i)$	$\sum_{\forall i \in G} P_{rob}(S_c(v_i))$	$S_c'(v_i)$
6	0.5 %	1
1	2 %	1
5	5 %	1
4	10 %	2
2	21 %	3
3	100 %	10

如前所述，我們的基本假設是正確的值組會有常見的字串特徵，其特徵出現的頻率較高，轉換後得到的數值較大。因此，特徵轉換數值的大小差異提供了一個推論及判斷值組為正確或錯誤的依據。特徵轉換數值的大小差異愈大時，代表少數值組的異常性愈明確，判斷其為錯誤值組的正確性機會愈高。若一值組集合在某一特徵上並沒有少數與多數的差異，而是平均的分布，則特徵轉換數值的大小差異就會不明顯，使得該特徵不容易做為區別正確與錯誤值組的依據。兩種轉換方式所提供的切割層面不同，對後續判斷與分類的準確性的影響也不同，我們將以實驗進行比較。

3.3 資料規則模型與錯誤偵測架構

資料或值組的錯誤偵測基本上可以視為二元分類的問題，對每一個值組進行正確或錯誤兩個類別的分類。為了自動建立有效的分類模型，我們採用機器學習或統計學習技術，透過含有已知分類結果的訓練資料，歸納出字串特徵及特徵轉換數值的分類規則或分類面。此分類模型即可用於目標資料中，對每一個值組進行分類，

達成判斷、辨識錯誤資料的目的。本研究所提出針對資訊擷取結果的錯誤偵測方法架構如圖一所示。



圖一、錯誤偵測方法架構圖

在自動建立資料分類模型的過程中，我們分別採用監督式機器學習(supervised machine learning)技術中的 C4.5 決策樹，及統計學習理論(statistical learning theory)中的支持向量機(support vector machine)。我們的目的是利用現有的自動學習技術，驗證字串特徵及特徵數值轉換做為錯誤資料偵測依據的成效。採用兩種不同技術的用意在於相互佐證錯誤偵測之成效，同時也驗證字串特徵之方法可有效搭配適當的自動分類學習技術。基於以上的考量，我們在 C4.5 決策樹及支持向量機的技術上，都是直接以公開可取得的軟體為主(本研究使用的 SVM 軟體及參數設定係參考國立台灣大學 LIBSVM 網站[2])，在實驗的過程中，並不做最佳化的調整。因此，相關的實驗結果只用以驗證字串特徵及特徵數值轉換做為錯誤資料偵測依據的成效，而無關 C4.5 決策樹及支持向量機兩者的比較。

4. 實驗評估

本研究選擇政府人事任免公報之擷取結果做為實驗對象，此公報分別以任命或免職的命令，記載政府各部門人事異動情形。政府人事任免公報的範例如圖二所示。

在過去的研究中[7]，我們針對此主題領域，採用型態辨識資訊擷取技術，處理約 20 年份的公報文本，共萃取出超過 10 萬筆人事異動資料，彙集而成包括姓名、組織單位、職位、職等、異動原因和日期等屬性的資料庫。目前大約二分之一的資料(西元 1995 年到 2004 年)已完成人工檢驗與校正，我們以這些資料做為訓練資料與測試資料的來源。

> ・・・
> 任命鄒擅銘為國史館臺灣文獻館簡任第十職等組長。
> 任命吳文慎為交通部臺灣區國道新建工程局人事室簡任第十職等主任，林渭鵬為經濟部水利處人事室簡任第十職等主任。
> 任命鍾萬梅為行政院客家委員會簡任第十二職等處長，黃崇烈為簡任第十一職等副處長。
> 行政院國家科學委員會科學工業園區管理科長曹常通另有任用，應予免職。
> ・・・

圖二、政府人事任免公報範例

4.1 評估方法

基本上，我們對資訊擷取結果的錯誤偵測方法是在建立一個適當的二元分類器，以指出資訊擷取結果中的合格資料與異常資料。因此，我們採用標準的 2 x 2 confusion matrix 做為偵測結果的效能指標。表四為以 true positive rate、true negative rate、false positive rate 和 false negative rate 四種量度所組成的 2 x 2 confusion matrix。

表四、2 X 2 confusion matrix

	分類為 正確資料	**分類為 錯誤資料**
原來為 正確資料	number of true positives (TP)	number of false negatives (FN)
原來為 錯誤資料	number of false positives (FP)	number of true negatives (TN)

這四種量度的定義如下，其中目標樣本是指原為正確類別的資料，非目標樣本是指

原為錯誤類別的資料：

1. True positive rate：TP-rate = TP/(TP+FN)，是指目標樣本分類正確的樣本數目比率。
2. True negative rate：TN-rate = TN/(FP+TN)，是指非目標樣本分類正確的非樣本數目比率。
3. False positive rate：FP-rate = FP/(FP+TN) = 1 – TN，是指非目標樣本分類成目標樣本的錯誤樣本數目比率。
4. False negative rate：FN-rate = FN/(TP+FN) = 1 – TP，是指目標樣本分類成非目標樣本的錯誤樣本數目比率。

二元分類器的效能目標是最大化 TP-rate、TN-rate 或最小化 FN-rate、FP-rate。一個理想的二元分類器能夠使 TP-rate 與 TN-rate 為 1。然而，在大多數的實際情形下，TP-rate 與 TN-rate 會是損益平衡關係，當我們對二元分類器進行調整，而能使 TP-rate 增加時，也同時會造成 TN-rate 的降低或 FP-rate 的提升。

4.2 實驗結果

為了能對字串特徵做為錯誤偵測依據的效能做不同層面的評估，本研究進行了四組實驗。第一組「字串特徵數值轉換」實驗是針對本方法所採用的兩種字串特徵數值轉換方式，比較其差異與優劣。第二組「訓練及目標資料範圍」實驗，乃根據訓練及目標資料組成範圍的不同，分為單一年份及合併年份，以觀察訓練資料的範圍，是否影響資料分類模型的表現。第三組「字串特徵數值轉換參數」實驗，將會改變 w 參數(區分的群組數)，以比較及分析其結果。第四組「訓練資料組成」實驗，是調整訓練資料的正反案例筆數，分別為 1:1 到 N:1 的各種組成方式，比較其資料分類模型效能之差異。

4.2.1 字串特徵數值轉換

本實驗中訓練資料的建立是於所有資料中，隨機取樣正反案例各 100 筆，並且隨機取樣三次，得到三組不同的訓練資料。實驗的進行，是以設定 *w* 參數為 100，並且以三組不同的訓練資料分別得到的分類結果，取其平均值，做為整體的實驗結果。表四為兩種字串特徵數值轉換方式的實驗結果，每一組資料中的數據分別代表 TP-rate/FP-rate。二元分類器的效能目標是最大化 TP-rate 和最小化 FP-rate，我們可以發現不論是由 C4.5 或 SVM 所建立的分類模型，「累計百分比轉換」方式的整體效果會比「個別百分比轉換」方式好。兩種轉換方法的差異，在於前者對於特徵值區分群組的切割較細，提供了機器分類法學習出更加細緻的分類規則。

此外，以表五中的各項實驗數據可以發現，對於「姓名」欄位的偵測表現不如其他欄位理想。這是由於該欄位的資料內容變化性較大，造成資料偵測模型較難精準的反映整體資料特性。而「單位」、「職稱」欄位的資料變化程度小於「姓名」資料，因此偵測的效果較好。至於「職等」、「年份」與「總統」欄位的資料內容皆是較為單純，因此透過訓練資料，就能建立出完整的分類模型，而得到非常理想的偵測效能。但是 SVM 分類法仍有些許誤判的情況，這點將在後續的小節說明。

表五、字串特徵數值轉換方式之比較

	C4.5 個別百分比轉換	C4.5 累計百分比轉換	SVM 個別百分比轉換	SVM 累計百分比轉換
姓名	84.91% / 33.57%	90.06% / 29.06%	87.11% / 60.79%	92.12% / 57.19%
單位	84.58% / 13.56%	88.58% / 7.69%	76.54% / 5.99%	79.61% / 4.00%
職等	100.00% / 0.00%	100.00% / 0.00%	86.21% / 5.04%	89.47% / 4.86%
職稱	87.44% / 11.36%	91.13% / 3.81%	80.58% / 2.01%	81.81% / 1.43%
總統	100.00% / 0.00%	100.00% / 0.00%	100.00% / 4.17%	100.00% / 2.58%
年份	100.00% / 2.88%	100.00% / 0.00%	93.96% / 5.00%	97.25% / 3.41%

4.2.2 訓練及目標資料範圍

本實驗建立兩種不同之訓練及目標資料範圍，分別為「單一年份」及「合併年份」，其他實驗參數包括 *w* 參數為 100，訓練正反案例筆數各為 100。「單一年份」

中，訓練資料為西元 2001 年，目標資料則分別為西元 2002 年、2003 年和 2004 年，因此共進行三次個別實驗後，再計算其整體平均之 TP-rate 和 FP-rate。而「合併年份」中，訓練資料與目標資料皆從十年份之資料庫取出(西元 1995 年到西元 2004 年)，而訓練資料為隨機取出三次，分別進行模型建立及資料分類測試，再計算其整體平均之 TP-rate 和 FP-rate。

表六為訓練及目標資料範圍的實驗結果，欄位中的數據仍分別代表 TP-rate/ FP-rate。整體而言，從「合併年份」資料範圍中所建立的分類模型，比從「單一年份」資料範圍中所建立的分類模型能得到較好的偵測結果。這是因為「合併年份」的資料範圍提供了較為全面的資料變化採樣空間，使得學習出的資料分類模型可靠度較高。至於各欄位的實驗結果與上一小節情況相似，對於資料變化程度越大的資料分類的準確度較低，反之亦然。

表六、訓練及目標資料範圍之比較

	C4.5 單一年份	C4.5 合併年份	SVM 單一年份	SVM 合併年份
姓名	95.78% / 37.49%	90.06% / 29.06%	76.33% / 57.29%	92.12% / 57.19%
單位	95.83% / 13.24%	88.58% / 7.69%	90.83% / 21.66%	79.61% / 4.00%
職等	100.00% / 0.00%	100.00% / 0.00%	85.23% / 5.62%	89.47% / 4.86%
職稱	89.62% / 17.45%	91.13% / 3.81%	58.86% / 16.11%	81.81% / 1.43%
總統	100.00% / 0.00%	100.00% / 0.00%	100.00% / 3.98%	100.00% / 2.58%
年份	100.00% / 2.56%	100.00% / 0.00%	95.07% / 4.44%	97.25% / 3.41%

4.2.3 字串特徵數值轉換參數

本實驗的目的是比較*w*參數，也就是特徵值出現的百分比對應的等份區間數目，依序設為5、10、25、50、100等變化下的不同分類表現。根據前面小節的實驗結果，本實驗在字串特徵數值轉換方式上，採用「累計百分比轉換」，而訓練與目標資料範圍，則是採用「合併年份」，訓練的正反案例筆數皆為100，並且隨機取樣三次建立三份測試資料，分別進行模型建立及資料分類測試，再計算其整體平均之

TP-rate和FP-rate。表七為*w*參數變化之實驗結果。

表七、字串特徵數值轉換參數之比較

欄位：姓名					欄位：職稱				
w	**TP-rate**		**FP-rate**		***w***	**TP-rate**		**FP-rate**	
	C4.5	**SVM**	**C4.5**	**SVM**		**C4.5**	**SVM**	**C4.5**	**SVM**
100	90.06%	92.12%	29.06%	57.19%	**100**	91.13%	81.81%	3.81%	1.43%
50	87.30%	84.82%	27.50%	51.56%	**50**	91.13%	86.11%	3.81%	6.65%
25	85.82%	72.95%	27.19%	28.75%	**25**	91.13%	89.89%	3.81%	9.33%
10	74.91%	80.17%	27.50%	40.00%	**10**	91.13%	88.12%	3.81%	8.79%
5	71.03%	87.05%	25.93%	49.56%	**5**	91.13%	89.19%	3.81%	11.40%
欄位：單位					**欄位：總統**				
w	**TP-rate**		**FP-rate**		***w***	**TP-rate**		**FP-rate**	
	C4.5	**SVM**	**C4.5**	**SVM**		**C4.5**	**SVM**	**C4.5**	**SVM**
100	88.58%	79.61%	7.69%	4.00%	**100**	100.00%	100.00%	0.00%	2.58%
50	94.46%	85.76%	19.01%	6.75%	**50**	100.00%	100.00%	0.00%	2.58%
25	90.06%	89.32%	9.36%	9.00%	**25**	100.00%	100.00%	0.00%	2.58%
10	82.45%	85.70%	4.50%	9.50%	**10**	100.00%	100.00%	0.00%	2.58%
5	81.71%	78.19%	3.94%	3.75%	**5**	100.00%	100.00%	0.00%	2.58%
欄位：職等					**欄位：年份**				
w	**TP-rate**		**FP-rate**		***w***	**TP-rate**		**FP-rate**	
	C4.5	**SVM**	**C4.5**	**SVM**		**C4.5**	**SVM**	**C4.5**	**SVM**
100	100.00%	89.47%	0.00%	4.86%	**100**	100.00%	97.25%	0.00%	3.41%
50	100.00%	88.51%	0.00%	3.47%	**50**	100.00%	97.25%	0.00%	3.41%
25	100.00%	88.51%	0.00%	3.47%	**25**	100.00%	97.25%	0.00%	3.41%
10	100.00%	88.64%	0.00%	3.47%	**10**	100.00%	97.25%	0.00%	3.41%
5	100.00%	88.51%	0.00%	3.47%	**5**	100.00%	97.25%	0.00%	3.41%

實驗結果顯示，*w* 參數較大時，所建立的資料分類模型的整體表現較好，至於其差異的大小則因欄位中內容變化的程度而異。例如，「姓名」欄位內容變化程度最大，*w* 參數提高，造成分群數目變多，使得資料分類模型在分類表現上的改善較為明顯。而「職等」、「總統」與「年份」欄位中，獨特值組的數目相當有限，造成該資料集合的特徵數值個數較少，因此當調整 *w* 參數時，不論區間數目增加或減少，

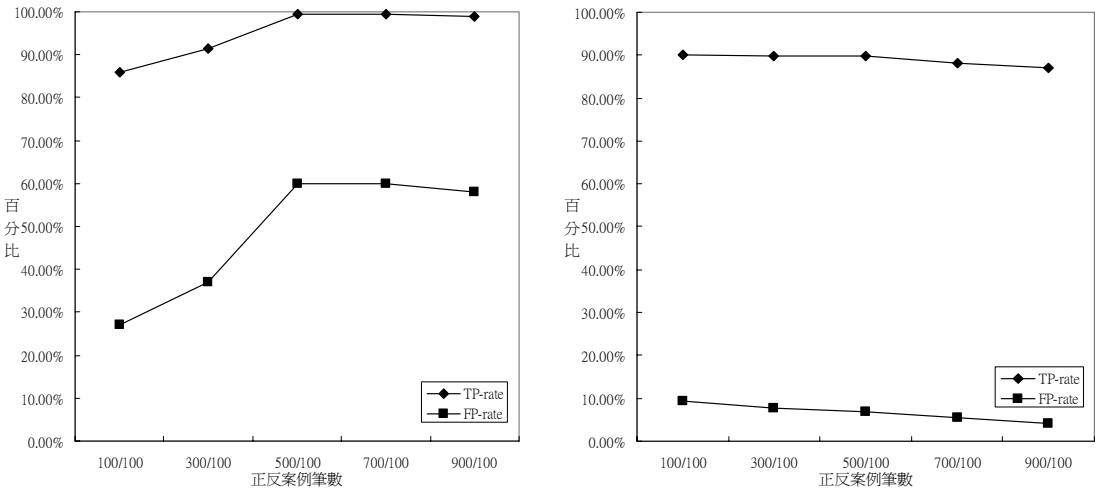
TP-rate 與 FP-rate 皆不受影響。至於「單位」與「職稱」欄位，其獨特值組的數目介於多與少的兩個極端之間，依照合理的推論，其分類表現應具備隨著 *w* 參數提升而改善的現象，但實驗結果所顯示的趨勢並不明顯，可能是本實驗之隨機採樣未能得到較完整之代表性，需要進一步的實驗驗證。另外，實驗結果也大致顯示 TP-rate 與 FP-rate 之損益平衡關係，即 TP-rate 若上升(TP-rate 變好)，則 FP-rate 也隨之上升(FP-rate 變壞)，兩者成反比關係。

4.2.4 訓練資料組成

本實驗是針對訓練資料中正反案例的組成變化，比較其建立之不同資料分類模型之表現。訓練資料的正反案例筆數組成分別為 100/100、300/100、500/100、700/100、900/100，並且隨機取樣三次，建立各種組成之三份測試資料，分別進行模型建立及資料分類測試，再計算其整體平均之 TP-rate 和 FP-rate。另外，*w* 參數設為 25，並使用「合併年份」為資料範圍。由於實驗資料是採用我們之前的資訊擷取成果，其擷取準確率約在 90%以上。因此，我們將正反案例比例逐步調整，從 1:1 到與資訊擷取成果中之正確的比例相同，進而觀察訓練案例比越趨近真實類別比例的表現變化。由於論文篇幅的限制，我們僅呈現較具代表性的實驗結果，包括以 C4.5 決策樹所建立之資料分類模型於「姓名」與「單位」兩個欄位上的表現變化。

如圖三所示，資料分類的表現大致上仍顯示 TP-rate 與 FP-rate 的損益平衡關係。在「姓名」欄位中，隨著正反案例比例的增加，TP-rate 的提升也帶來 FP-rate 的升高。「單位」欄位則是越接近 9:1 的比例，表現越好。「職稱」欄位在 3：1 的比例時，表現較好。而「職等」、「總統」與「年份」變化程度最小的欄位，其表現則是維持不變。每個欄位有不同的結果，是由於各欄位內容變化程度的差異相當大。「姓名」欄位的獨特值組數量非常龐大，所以正確案例筆數為 900 筆時，不但無法建立較完整的資料分類模型，反而造成 C4.5 決策樹所學習出的分類模型偏重於正確案例的一方，TP-rate 變好，但 FP-rate 也大幅增加。「單位」欄位在資料範圍中的獨特值組數

量約 1400 筆左右，所以當正確的訓練資料筆數越接近該數值時，所學習出的資料分類模型會更為正確與完整。



(a) 姓名欄位　　(b) 單位欄位

圖三、訓練資料組成變化對分類表現之影響

「職稱」欄位的獨特值組數量大約在 200 筆左右，所以在 3:1 的案例比時就有不錯的表現，當正確案例筆數增加，表現維持不變。至於「職等」、「總統」與「年份」欄位的資料就非常單純，獨特值組數量少於 30 筆，因此不但不受正反案例筆數的影響，並且整體效果接近完美。

4.3 實驗結果分析與討論

本研究所提出的以字串特徵做為錯誤偵測的方法，在實驗資料中的各項最佳表現如下：「姓名」欄位 TP-rate 約 85%、FP-rate 約 27%，「單位」欄位 TP-rate 約 87%、FP-rate 約 4%，「職稱」欄位 TP-rate 約 91%、FP-rate 約 3%，其他「職等」、「總統」與「年份」等欄位，皆能完美分類資料。實驗結果顯示，字串特徵能夠有效代表部份正反案例資料所展現之特徵，而建構出準確之資料分類模型，並與該欄位的內容特性相當吻合。這是由於字串特徵概念隱含了欄位內容之領域知識元素，再透過特徵數值轉換與 C4.5 決策樹的自動建構出適當的資料分類模型，而能展現有效的錯誤

偵測能力。

我們以表八做為說明。「姓名」欄位的分類模型之主要依據為 string cardinality (S_c)、string prefix (S_p)、string suffix (S_s)，分別是字元個數、字串前 k 個字元以及字串後 k 個字元。這個現象與姓名資料的特性相當吻合，譬如姓名資料的字元個數大都集中於 2 或 3，而字串前 1 個字元也是姓名中的姓氏，因此能夠適度的反映該資料之形式規則。「單位」欄位的決策樹是由這 string cardinality (S_c)、與 string suffix (S_s) 所組成。這個現象也與單位資料的特性相當吻合，不同單位名稱除了長度差異不大以外，最後一個字通常是「局」、「處」與「室」等常見的單位結尾詞。

表八、資料分類模型中之主要字串特徵

姓名	單位	職等	職稱	總統	年份
S_c、S_p、S_s	S_c、S_s	S_p、S_e、S_n	S_s、S_e	S_e、S_f	S_p、S_n、S_f

「職等」欄位則大多為「簡任第十職等」或「警正二階」等字串內容，這些字串大多含有數字類型的資料，因此造成 string numeral (S_n)字串特徵成為該欄位的決策樹主要節點之一。同時，字串開頭與完整字串內容都較為單純，所以 string prefix (S_p) 和 string entity (S_e) 也是該欄位的決策字串特徵。「職稱」欄位之資料內容，最後一個字通常是「員」、「官」與「長」等常見的職稱結尾詞，而「公務人員」與「警察官」的字串出現的次數相當高，因此 string suffix (S_s) 和 string entity (S_e)成為該欄位的主要決策字串特徵。

最後，「總統」欄位的內容只有兩任總統的名字，並且都是字串型態，因此 string entity (S_e) 和 string fromat (S_f)就可以反映出該欄位的特性。「年份」欄位的正確資料皆是數字型態以外，目標資料的年份為 84 年到 93 年，字串開頭為 8 或 9，也正好反映 string prefix (S_p)、string numeral (S_n) 和 string fromat (S_f)三個字串特徵的作用。

5. 結論

本研究提出以字串特徵為主的中文文本資料錯誤偵測機制，並以充分的實驗結果驗證其資料描述能力與錯誤偵測效能。此錯誤偵測機制能以後處理(post processing)的流程步驟，搭配一般的資訊擷取技術，確保高品質的資訊擷取成果產出，促成資訊擷取技術更廣泛的實際應用。另外，我們的研究成果也對於以中文為主的資料清理技術有所幫助，能應用於一般中文的大型商業資料庫上的錯誤偵測。

致謝聲明

本研究成果由國科會計畫 NSC 94-2422-H-004-002 及 NSC 95-2422-H-004-003 提供部分經費支持，特此致謝。

參考文獻

[1] Galhardas, H., Florescu, D., and Shasha, D. An Extensible Framework for Data Cleaning, *INRIA Technical Report*, 1999.

[2] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

[3] Muller, H., and Freytag, J. C. Problems, Methods, and Challenges in Comprehensive Data Cleansing. *Technical Report HUB-IB-164*, Humboldt University Berlin, 2003.

[4] Rahm, E. and Do, H.-H., "Data Cleaning: Problems and Current Approaches," *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol. 23, No. 4, December 2000.

[5] Raman, V. and Hellerstein, J. M. An Interactive Framework for Data Cleaning, *UC Berkeley Computer Science Division Report No. UCB/CSD00/1110*, September 2000.

[6] 李念秋，資料品質改善之研究：錯誤資料偵測技術之發展與評估，國立中山大學資訊管理系碩士論文，2002。

[7] 翁家緯，以型態辨識為主的中文資訊擷取技術研究，國立政治大學資訊科學系碩士論文，2003。