資料倉儲環境中資料品質的評估機制*

An Evaluation Mechanism for Assuring Data Quality in Data Warehouse Environment

朱雨其　楊珊珊　楊鏡樵
Yu-Chi Chu, Shan-Shan Yang, Chen-Chau Yang

國立台灣科技大學　電子工程系
Dep. of Electronic Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan, Republic of China
ycchu@sun.epa.gov.tw　ccyang@selab3.et.ntust.edu.tw

摘　要

本文旨在探討資料倉儲環境中，如何評估系統內部所儲存資料之品質的方法，以及在提昇資料品質與降低成本支出間找尋合理之平衡點，期能以最低的成本花費，提供最優良的資料品質。我們基於品保/品管的考量，倡議一項資料品控制作業流程，並將其與資料倉儲系統的發展相溶合。在資料品質保證機制方面，我們採行二階段的作法，首先是建立資料品質的實體模型，再運用資料品質成本/效益評估模型，尋求資料品質"最適使用"的狀態。資料品質實體模型是以屬性爲基礎並延用傳統的實體關係模型加以擴充，用以將資料品質的品質因子融入資料倉儲系統的設計過程，使資料被區分爲一般應用的屬性資料及提供品質訊息的品質資料。藉由品質資料所提供的訊息可以識別資料的良窳與適用與否。其次，經由成本/效益評估模型的運算，可在有限成本控制下針對最嚴重的不良資料作優先的修正處理。本文倡議的方法不僅能加强屬性資料的詮釋性，免除資料誤用及語意的混淆，還能提供診斷性之資訊以期發現錯誤發生的原因及出處。

關鍵詞:資料品質，資料倉儲，資料淨化，成本/效益模型

ABSTRACT

The purpose of this paper is to explore the balance between the benefits and cost of the improvement of data quality in the environment of data warehouse. Based upon the QA/QC consideration, we propose an operation process for data quality control, and incorporate the process with the development process of data warehouse systems. Concerning the data quality assurance, we adopt a two-phase approach. We first construct the data quality model, then employ the cost/benefit model to determine the "fit to use" condition of data quality. The data quality model is built upon ER-Model, that is expanded with the consideration of data quality indicators. Therefore, the datum in data warehouse can be classified into attribute data and quality data. In terms of the information provided by quality data, we can identify the diversity of poor data which are not fit to use. In addition, we can also use cost/benefit model to find out the most fatal poor data and set the priority for correction under limited cost. Our approach not only can enrich the interpretation of attribute data, avoid of data misusing, as well as data misunderstanding, but also can provide the diagnostic information to figure out the reasons and the sources of errors.

Keywords: data quality, data warehouse, data cleansing, cost/benefit model

1　緒論

資料倉儲系統(Data Warehouse System)近年來不論在學理上或實作上都引起廣泛的研究興趣[6, 3, 4, 14]，一般而言，由於資料倉儲對繁瑣的大量資料具有整合與彙總的功能，因此對於建構在某個特定議題之下(Subject-Oriented)的資料倉儲系統，具有輔助分析整體趨勢與決策支援的功效；目前資料倉儲系統已有逐漸成爲各機構組織用來整合異質性資訊源重要策略的趨勢。事實上，資料倉儲可視爲一種整合性的資料儲存體，其內部所儲存的資料乃是由多個分散式、自主性以及異質性的資訊源中，萃取並整合而來的。我們預見未來會有越來越多的機構組織開始建構資料倉儲系統以做爲決策分析使用，相對的更突顯出資料倉儲中資料品質問題的重要性。換句話說，決定資料倉儲系統是否能夠開發成功以及有效運作發揮實際功效的一項關鍵因素，必須取決於資料倉儲系統內部所儲存資料的品質是否足以適用，資料是機構組織中的重要資源(產)之一，更應該加以重視並對其資料品質做妥善有效的分析與管理。但是在目前絕大部分的資料倉儲相關研究與實作發展中，大都將焦點著重於加强資料倉儲

*本項研究係由國科會補助研究計畫經費(NSC88-2213-E011-004)

的查詢及索引結構方面[4, 9, 11, 12]，而忽略了最基本的資料品質問題。據估計約有六成以上的資料倉儲系統宣告失敗，其主要的原因就是因爲這些存在於系統中的資料品質問題[8]，系統開發人員沒有給予充分的時間與努力來解決、淨化資料品質的問題而導致失敗。

資料品質的問題在傳統資料庫與資料倉儲中最大的不同處在於資料倉儲中的資料主要是做爲決策支援所用，而非操作性的交易應用。因此儲存在資料倉儲中的資料通常是歷史性資料，透過歷史性資料在時序上的變化與比較不僅可做爲決策分析的依據，同時也引發了我們對資料品質的評估與時序關係變化的研究動機。事實上，資料倉儲中資料的品質問題也絕非是突然間憑空形成的，必定是經過長時間的延續下累積而成，因此資料的品質與時間之間必然存在著某種關係，這也是促使我們以時間爲基礎，開始進一步分析與探討改善資料品質的模型建構方法。

本文旨在探討有關於資料品質的定義以及各種造成資料品質問題的歧異性，以及這些問題的重要影響與解決方法，並提出評估資料品質的一套機制，以達到資料倉儲系統中資料品質保證的目的。第二章主要描述資料品質的意義及品質模型的整體架構，第三章則從設計與實現階段兩方面來討論資料品質的控制處理機制，第四章提出一個資料品質控制處理機制中的成本／效益評估模型，第五章作結論並闡明未來研究方向。

2 資料品質定義與品保作業

2.1 資料品質的階層性

就資料品質本質上的觀點而言，在使用資料庫或資料倉儲中的資料時，使用者最關切的莫過於資料是否適於使用“fit to use”，是以目前不論是學理或實務上，大部分都將資料品質的意義定位在“適於使用”的目標上[6]。由這項定義繼續延伸，必須再明確推演出使資料適於使用的基本要素，基於這項需求，可將資料品質再細分成四個層面(Dimension)來討論分析[13]，每個層面又可再細分爲若干個資料品質參數(Data Quality Parameter)，資料品質參數的主要作用是讓使用者評估資料倉儲中的資料品質。品質參數的形成與選擇目前雖無一定之規則，但仍須能充分表示出原始資料特徵的先決條件。圖1是構成資料品質定義之階層圖，資料品質的四個層面與品質參數說明如下：

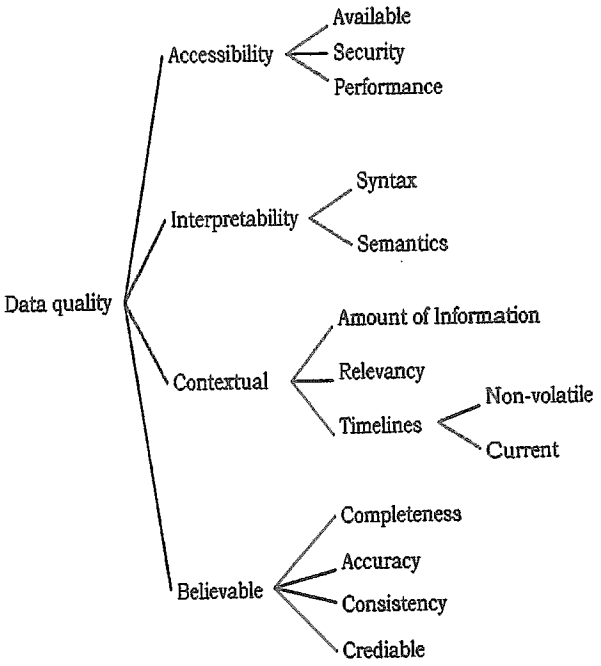

Figure 1: 資料品質定義之階層圖

1. **存取性(Accessibility)**：對使用者而言，具有良好資料品質特性的資料倉儲應具有輕易取得所需資料以做進一步分析操作的能力。其次是安全性(Security)的考量，對於機密性的資料爲確保其安全性與隱密性，必須有效限制使用者的存取。乍看之下，資料的存取性與安全性考量在某些情形下是相衝突的，但若完全無法取得這些保密性資料，則分析人員將無法研究解決資料不適用的問題，需求資料的管理者也無法做出相關的決策。是以發展一套合理的機制有效限制機密資料的存取是必要的。當資料量開始逐漸增加時，還可再加上效能(Performance)參數來評估資料的使用是否足以負擔。

2. **詮釋性(Interpretability)**：詮釋性層面的品質參數主要用來描述資料本身的定義，資料的定義明確清晰將有助於使用者或分析人員提高對資料的了解度，了解的層面應包含資料的格式、內容與主要用途等。因此在詮釋性層面中包含了資料格式(Syntax)與資料語意(Semantics)等兩個品質因子，用來定義資料項的屬性。資料的格式必須明確定義出資料的屬性單位及精確度，資料語意則記錄資料的內容與主要用途。

3. **議題關聯性(Contextual)**：資料是否切於議題，存乎於資料的適用性。資料是否適於使用則可由資料量的多寡、資料之相關性(Relevancy)及合時性(Timeliness)等三方面來探討，其中合時性品質參數又進一步將資料的時間性質區分爲非揮發性(Non-volatile：資料項的使用無時效限制)與具時效性(Current：資料項被存入資料庫的時間及其有效期限)。匯入資料倉儲的資料量多寡應視應用爲基礎，而非將所有資訊源中的資料完全整合匯入系統中，大量的資料雖有助於決策分析使用免除相關資料不存在的窘境，但過多的資料也將造成使用者無法在合理時間中存取資料而使系統效能降低以及資源浪費的缺失。

4. **可信度(Believable)**：資料倉儲的資料除了要與議題具有相關性外，還要能取得使用者

的信任，這樣的資料才具有可利用的價值。使用者對資料的可信度，一般可由完整性(Completeness)、一致性(Consistence)、正確性(Accurate)及可靠性(Credible)等四個品質參數來衡量[2, 7]。

一般而言，資料品質的概念是屬於多維度的，而品質參數的訂定也最好要符合各種不同的資料型態，來自各個不同的應用領域中，都有著共同的品質參數用以評估其資料品質的需求。這些品質參數除了可用來標示資料項之外，還可以標註於品質因子上，以確保品質因子的品質。這種多維度與多階層品質參數及品質因子的概念，便可用來解析資料品質的特徵。

2.2 品保作業流程及機制

相較於原有傳統的二維式表格資料庫，資料倉儲可在資料屬性欄位附加上品質因子的相關資料，並依據應用需求藉由品質因子做爲未來評估篩選的準則。擴展後的資料倉儲在容量上必定會大於原有的舊系統，但是以長遠的眼光來看，擴展後的品質資料倉儲在使用上卻更能提供較大的彈性，使用者在實際的操作應用上並不會感覺到這些品質相關資料的存在，藉由對品質資訊的淨化與篩選，相對地對於資料的品質卻提供了一層更大的保障以及可信度。爲使資料品質的問題在系統開發的設計階段、完成階段到日後的使用及維護階段都在嚴格控管之下，我們提出一個將資料品質控制的作業流程，以構成一種資料品質保證機制(Data Quality Assurance Mechanism)，並與資料倉儲系統發展的架構相融合。

一般而言，匯入資料倉儲中的資料來源，除了分散於各部門的本地資料庫之外，還包含了一些外部的資訊源。如果這些資訊來源本身所儲存的資料，夾帶了一些錯誤的訊息，那麼這些錯誤的訊息也會被一併整合、匯入資料倉儲系統中，即使原始資料沒有任何瑕疵與錯誤，後續資料的維護與更新，也有可能造成資料錯誤或資料品質惡化的情況發生。資料品質的控制機制是對即將匯入品質資料倉儲中的資料進行品質驗證處理，其控制機制與流程如圖2所示。詳細的作業過程說明如下：

1. **資料品質的需求分析：** 這項過程與傳統關聯式資料庫設計過程中的資料需求分析類似，惟系統分析人員仍有必要去了解取自各資訊源之異質性資料所代表的語意與格式，做爲建構資料倉儲中各項異質性資料間語意釐清與一致性表示法的解決方案。更進一步分析，爲使品質因子中所包含的品質資料也在品質控制下，也可應需要爲品質因子附加上第二層或第三層以上的品質因子，以達到品質保證的目的。每個品質因子都包含了相關的品質資料，例如：異質性資料的來源(可能包含一個以上的資訊源)，資料的時效性等限制。

2. **建構以屬性為基礎的資料品質模型：** 本項作業將資料品質控制的概念併入了傳統資料庫

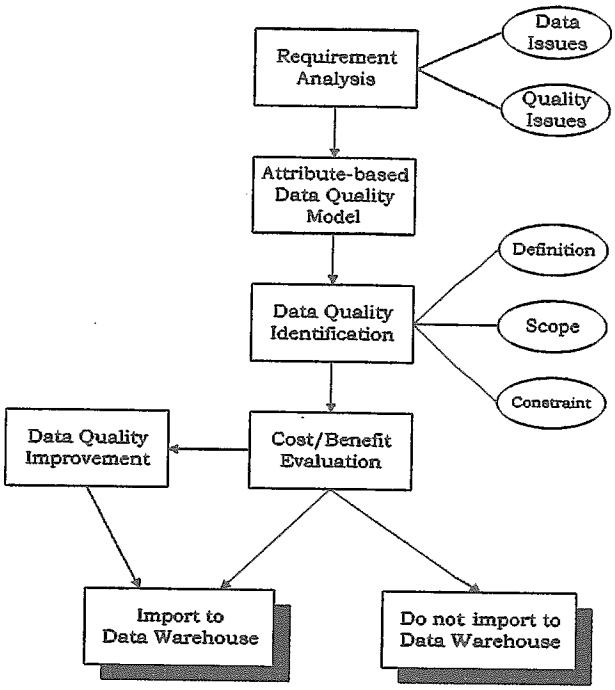

Figure 2: 資料品質控制作業流程

的設計中，並以屬性基礎的資料品質模型定義品質資料在資料倉儲中的儲存結構與方法。品質資料的加入除依據品質因子附加於需求的異質性資料欄位上(附加的原則是以屬性爲基礎)之外，還須將這兩種資料結合成一個單一的元素。由於品質資料主要的作用是用來詮釋資料藉以提供品質方面的訊息，因此與其相對應的資料具有密切之關係，是以對資料所進行的存取動作包含新增、刪除或修改等，也須對相對應的品質因子進行適當的新增、刪除或修改等處理。

3. **識別不良資料：** 包含有品質因子的資料可進一步分析處理，找出資料庫中不符合品質標準的資料。不符合品質標準的資料未必就是錯誤的資料，只是在資料倉儲的多使用者環境中，每個使用者對資料品質所需求的評估標準未盡相同，例如：資料的時效限制，使用者A對於資料表格D中當月的資料，其時效性的需求爲一年(西元一九九九年)，而使用者B卻需要資料表格D中過去十年的資料(由西元一九八九年到一九九九年)，由此可見，使用者A與使用者B對資料時效性的評估標準便不一樣，一旦資料不在標準範圍內，對使用者來說這樣的資料便不適用了。因此系統分析人員有責任在資料需求分析階段時就取得所有的品質標準，並在此階段中訂出不良資料的定義、標準範圍與條件，藉以識別出不良的資料並加以修正，使資料倉儲中的資料品質能夠適用於每個使用者。

4. **成本/效益評估：** 前項步驟中所識別出在範圍標準外的不良資料，是否可以或應該進行資料品質的淨化處理？就現實的角度來說，當

提昇資料品質所需的成本過高，在經濟上求取零缺點的資料是不切實際的。況且，零缺點的資料在有些時候可能是既不需要，同時也無法達成的。在絕大部分的情況下，降低資料錯誤率的最佳解通常是降低成本的最差解，因此根據實際需求在成本與品質尋求一平衡點，以最低的成本達到最高的品質是必要的。

最後的程序是將提昇品質的資料訊息傳送給系統分析人員，以做爲未來系統資料品質潤飾的基礎。對於符合適用標準的資料，可直接匯入資料倉儲中。相對的被識別出的不良資料經過效益評估後，可分爲X與Y兩個集合，其中X集合中的不良資料可在合理的成本效益中，予以提昇資料品質後，匯入品質資料倉儲中。剩餘無法在合理成本支出中獲得品質改善或無法找出相關修正方法的不良資料便構成Y集合，Y集合中對於未被處理的不良資料是否要匯入品質資料倉儲中的決擇，則依據使用者導向原則，由使用者基於主觀、客觀或其他人爲外在因素自行決定。

3 資料品質模型設計與實現

3.1 結合資料品質的資料模型

Wang氏等倡議以屬性爲基礎的品質資料表示方法[13]，針對資料的屬性在資料欄位附加上品質因子，這些品質因子就是實際資料所欲達到的目標、特徵以及產生的過程，使用者則依據實際應用需求來評估資料的品質。附有品質因子的資料欄位可聯結至相關的品質資訊，其儲存結構改變了原始關聯資料庫中資料欄位值必須是單一值的限制，使每項有品質因子的資料欄位都以< Attribute , Quality_Key >的序對方式儲存資料。改變原始資料儲存方式所需付出的代價，便是對品質因子的設定、儲存及擷取必須做進一步的處理，以免造成資料因新增或刪除所形成的異常現象，同時還要修改SQL的查詢語言結構使其包含對品質資料的處理。我們將[13]的品質資料表示方法加以改良，使其能保留原有的優點，並免除品質因子的額外處理工作與SQL語言結構的修改。

由於資料品質實體模型將使用者對資料品質的觀點與需求併入了資料倉儲的設計中，因此對於不同使用範疇的使用者，可依據其使用權限或需求的資料品質標準，做爲篩選資料的標竿，一旦外界環境對資料品質的需求有所改變，也可隨時再重新訂定資料品質的標準，進行資料篩選與淨化處理，這使得資料品質的驗證評估在資料倉儲發展與應用的任一階段中皆可進行處理，而非僅限於設計階段或應用階段始可進行[10]。

表1爲含有品質資訊的資料庫示例，原有的資料表格分店< ID, Addr, Owner, Profit_est. >，經過資料品質需求分析之後，便在屬性欄位附加上品質因子，品質因子中包含有對資料相關的品質訊息，如表1所示的例子中顯示，在屬性欄位分店"獲利評估"(Profit_est)上附加了三個品質因子：< Entry_date, Evaluator, Entry >，分別用以表示資料鍵入日期、獲利評估分析人員，與資料輸入人員。使得擴展後的屬性欄位"獲利評估"可進一步取得相關的品質資料。

3.2 實作構建與釋例

在資料需求分析過程中最後所得附有資料品質因子的需求文件，將做爲實現階段中所用，使屬性資料的品質訊息確實併入資料倉儲中。由於每個屬性可能會有任意數目的資料品質因子，爲使屬性與其相關的資料品質因子儲存於資料倉儲時即發生關聯性，必須發展一套機制將欲檢驗的資料品質因子併入資料倉儲的綱要設計中，以便將兩者互相連結在一起。每個附加有資料品質因子的屬性在屬性欄位上會再附加上"Δ"的標記，以表示該屬性下的所有資料都可參考至相關的品質資料，屬性上所附加的資料品質因子是以層級爲單位，同一層級的資料品質因子視爲同一品質綱要，品質綱要中的主索引稱爲品質索引是由該屬性的主索引再附加上"Δ"的標記所形成，以達到唯一識別及參考的目的。經過資料品質需求分析後，在屬性< Profit_est.> 附加上資料品質因子，擴展後所形成的綱要便如表2所示變成屬性< Profit_est. Δ>，屬性< Profit_est.Δ> 中的資料可根據主索引加上"Δ"記號形成品質索引，參考至品質綱要< Profit_est.Δ> 中的相關品質資料(如表3)，其中在表3中又可見屬性<SourceΔ >欄位上附加了"Δ"的標記，表示屬性<SourceΔ> 也附加上了資料品質因子，因此也可依照同樣的品質索引參考至表4有關於屬性< SourceΔ> 的相關品質資料。由此可見，在屬性欄位上所附加的"Δ"標記可用來指示該項屬性的資料是否有關於品質訊息的資料，未加上"Δ"標記的屬性則表示不包含有相關的品質訊息。就這個例子而言，品質資料的存在甚至可以再予以擴展，收集更明確的相關資料，使品質資料提供更明確的資訊，例如：<Evaluator > Monica其相關工作經驗爲二十年，工作考績爲A；而<Evaluator> Bill其相關工作經驗爲三年，工作考績爲B。所以在查閱資料的同時並附上這些品質資訊，使用者可以更容易了解品質資料所要提供的訊息，以做爲決策分析的參考，詳細實作過程請參閱[15]。這與[13]所提的資料品質架構中，將每個附有資料品質因子的屬性經過擴展後形成<Attribute , Quality_key> 序對的多資料值形式相比較下，會繼續保留其優點而剔除掉要重新設計資料庫與增加SQL語言查詢功能的缺點。

其次，在實現階段中將品質資料併入資料倉儲設計中後，尚須再考慮資料整合性的問題，以免除資料存取時所造成的資料異常現象。考量的方式應將屬性資料與其相關的品質資料視爲一體，亦即在對附有品質訊息的資料做新增、刪除或修改時，相對應的品質資料也要必須進行新增、刪除或修改。

Table 1: 結合資料品質的資料模型(資料品質方體)

ID	Addr.	Owner	Profit_est.	Entry_date	Evaluator	Entry
B001	Taipei	David	$1,000	1998-03-21	Monica	Mary
B002	Taichung	Mike	$3,000	1998-03-21	Bill	John
⋮	⋮	⋮	⋮	⋮	⋮	⋮
data attribute				data quality		

Table 2: 擴展參考至相關品質資料之資料表

ID	Addr.	Owner	Profit_est. Δ
B001	Taipei	David	$1,000
B002	Taichung	Mike	$3,000
⋮	⋮	⋮	⋮

Table 3: 第一階層之品質因子資料表格

Profit_est. Δ	Source Δ	Entry
B001 Δ	Monica	Mary
B002 Δ	Bill	John
⋮	⋮	⋮

4 資料品質成本效益評估

資料品質的成本效益評估是較爲主觀的一項課題，本章運用前述資料品質模型爲基礎，進行資料品質之成本與效益評估處理，進而決定識別出之不良資料在經濟的考量上是否有修正資料品質問題的必要性。修正過後的資料品質在經過一段時間差之後，必須對使用者的資料品質需求及外界影響因素(如：商業主題就必須考量市場變動因素)調查是否有新的異動，如有更新就必須從識別不良資料步驟開始，再重複進行資料品質的成本效益評估處理。如果資料品質的需求沒有改變，就可將修正過後的資料，予以整合轉換後匯入資料倉儲。

4.1 確定資料之良窳

資料品質的問題不僅侷限於資料值的正確與否，可能還包含了資料格式上、語意上及其它作業特徵(如：資料的時效性與資訊來源等因素)上的未滿足需求限制等因素。一般而言，系統中八成以上的品質問題大多出自於同一項缺點[3, 12]，是以能夠以最少的支出(包括時間、人力、物力與金錢)來獲得最大的改善，才能符合經濟效益的需求。本階段有下列兩項主要程序要完成：

1. **資料品質標準的訂定：** 使用者需求所定義出的資料品質因子，既是用來儲存與描述該項資料的相關品質訊息，便可藉以做爲篩選資料的工具。篩選的標準必須依據議題需求訂定出符合適用條件的品質標準定義、可接受的誤差範圍與相關的條件規則。舉例來說：某項資料被擴充附加上"時效性"的資料品質參數，在"時效性"的資料品質參數上有著"資料建立時間"的品質因子，其欄位中儲存著該項資料的建立日期，另外爲避免資料品質因子中的品質訊息也有資料品質上的問題，品質因子資料欄位應同時明確規範資料的格式(如：日期時間或文字格式)與意義，標準訂定的項目包括：

Table 4: 第二階層之品質因子資料表格

Source Δ	Evaluator	Entry_date
B001 Δ	Monica	1999-03-21
B002 Δ	Bill	1999-03-21
⋮	⋮	⋮

- "時效性"品質標準的定義：資料具有時效性或非揮發性，用以表示資料的時效性限制。資料欄位中儲存資料建立的日期時間，其格式爲(日/月/西元年)或(西元年/月/日)。
- 可接受的誤差範圍：假設資料的時效性定義爲具時效性，則表示資料的取得時間或建立時間對使用者或系統應用上有著重大的意義，時間差距的定義可用來識別出過期的資料。例如：自資料建立日期始，資料的有效期限爲三十天，則在三十天範圍外的資料便不再適用。
- 其它相關規則：規則的訂定視作業需求而定。以銀行利率來舉例，利率的起伏不得高於或小於最大利率與最小利率。

2. **對各項品質問題加以排定優先權：** 訂定出各項資料品質因子的標準之後，便可識別出不適用的資料，而部分機構組織未必有修正所有不適用資料的需要與必要，因此對於未符合各項資料品質標準的資料，應就品質因子的分類排定其重要性(Ranking)，主要目的在於使分析人員可以很快的發現，到底是那些致命性不良資料降低系統使用的決策效率，進而決定要修正改善的不良因子項目與數量。重要性的排定可依主題不同來做數項排定，如資料錯誤率、各項品質因子不良對系

統所造成的損失以及提昇各項資料品質因子所需付出的成本等觀點，可綜合各項結果來決定哪些資料問題是一定要修改，而哪些資料不需要改善。

這二項程序的處理是爲下階段的成本/效益評估做事前準備，評估的結果是依據使用者的需求與目標來決定，先將不良資料加以識別並排定優先權，有助於成本/效益的評估，以便於最後決定應該要修正那些不良的資料。設定資料品質的目的，就是希望使資料倉儲中的資料品質，在嚴格控制下，達到資料品質保證的目的，未附有品質保證的資料，將大幅降低分析結果的可靠性。因此確定資料之良窳，予以修正改善其資料品質，可提高分析結果的可信度，並加强使用者對資料品質的可信度與資料的可用性，達到使資料適於使用的目標。

4.2 成本/效益評估模型

資料品質成本/效益評估模型的一個關鍵點在於如何將資料的品質予以量化，量化的標準又取決於何項因素。面對衆多使用者各自相異的資料品質需求與標準，如何使資料品質的量化同時兼具主觀性與客觀性是很重要的。模型建構中的一項重要假設，是建立在時間點上的基礎，所有更進一步的分析與探討，都是衍生自時間點上的關係。我們將資料品質惡化的程度定義爲時間的函數$Q(t)$，用以表示在時刻t時，資料倉儲環境中整體資料品質的惡化程度。其中資料的整體品質又可由各項資料品質因子的惡化程度決定

$$Q(t) = \sum_{i=1}^{n} Q_i(t) \qquad (1)$$

其中$i \in N$，$Q_i(t)$爲n項資料品質因子中之第i項資料品質因子的資料品質惡化程度。式子(1)中所定義之函數，無法得知各項資料品質因子間優先處理關係的資訊，每個資料品質因子的重要性與成本支出基本上是視爲相同的。如同前述部分機構組織在經濟效益考量的觀點上，未必有將資料品質提昇到完美零缺點境界的需要，在此前題下，分析人員可視應用需求對各項資料品質問題的重要性，使用Pareto Chart的統計圖表依不同觀點(資料的錯誤率或損失)進行分析與重要性分配[1]，將有助於決定改善資料品質問題時資料量與資料項的選擇。由於$Q(t)$爲在時刻t時資料品質的惡化程度，惡化的程度若以不良資料在資料倉儲中所佔的比例來表示，可將$Q(t)$再量化成$0 \sim 1$之間的數值，同時也能夠較客觀的由$Q(t)$的數值表示出資料倉儲中資料品質的實際狀況，分析人員對量化的結果具有解讀其資料品質對資料倉儲所代表的意義。故(1)可改成如下之式子：

$$Q(t) = \sum_{i=1}^{n} \frac{\rho_i}{W} \qquad (2)$$

其中W表示所有將進行資料品質評估處理的資料屬性量；ρ爲資料品質評估處理後識別出的不良資料之數量；ρ_i爲資料品質評估處理後識別出第i項品質因子的不良資料數量。很明顯的，當式子(2)中量化後的$Q(t)$趨近於0時(即$\frac{\rho}{W} \cong 0$)，意味著資料倉儲中不良資料所佔的比例非常少，當W的值非常大時，ρ幾乎可視爲不存在；相對的當$Q(t)$的值趨近於1時(即$\frac{\rho}{W} \cong 1$)，不良資料的數量與資料字集的數量很相近，資料倉儲中的資料絕大部分都不適用。基於成本效益均衡的考量，以上之討論應該綜合考慮資料品質不良所導致的損失費、改善資料品質所需付出的提昇費與提昇資料品質因子項目個數間等因素的關係，以最低的總成本支出來決定應該提昇的資料品質因子項目。總成本(total cost, TC)支出的計算包含了損失費(lost cost, LC)與提昇費(improvement cost IC)，這決定了應自識別出的不良資料中選擇若干項目予以修正。爲求進一步的分析，我們將對損失費與提昇費做如下之定義：

損失費： 一般而言，由不良資料所造成的損失與資料品質的惡化程度成正比，資料品質的惡化程度則與資料倉儲中不良資料所佔的比例有關。

提昇費： 提昇的資料項目愈多，要付出的成本也會愈多，因此提昇成本除了與資料品質因子的提昇項數有關外，也和每項資料品質因子的提昇成本有關。

損失費與資料品質的惡化程度有關，而資料品質的惡化程度與不良資料存在及被修正的時間有關，修正時間的長短又取決於欲修正的資料品質因子項數，修正的項目愈多，資料的品質就愈好。提昇費除了與修正的項數有關外，也與修正時間的長短有關。設不良資料開始存在的時刻爲$t = t_0$，開始修正不良資料的時刻爲$t = t_1$，修正完成的時間爲$t = t_2$。則不良資料品質對資料倉儲所造成的損失爲$Q(t_2)$；提昇資料品質所需之成本爲$Q(t_2 - t_1) + C$，C爲與時間無關之材料成本。其次，我們發現研究$\frac{dQ}{dt}$會比$Q(t)$更爲直接和方便。因爲$\frac{dQ}{dt}$所代表的意義是單位時間內不良資料所佔的比例，表示不良資料惡化的程度。模型的構成要對$\frac{dQ}{dt}$、不良資料所造成的損失及改善資料品質所需付出的提昇成本做出合理的簡單假設，亦即對$\frac{dQ}{dt} \sim t$的關係分佈建立模型：

1. 資料品質惡化程度$\frac{dQ}{dt}$定義爲單位時間內資料倉儲中不良資料所佔的比例。
2. 損失費與資料品質的惡化程度成正比，比例係數爲C_1, C_2，表示單位資料品質惡化的損失費。
3. 從錯誤資料存在到開始修正錯誤資料的這段時間內(即$t_0 \le t \le t_1$)，資料品質惡化速度$\frac{dQ}{dt}$與時間t的關係成正比，比例係數爲β，表示資料品質惡化的速度。
4. 修正資料品質因子錯誤性資料項，開使修正後(亦即時刻$t \ge t_1$)，資料品質惡化速度爲$\beta - \lambda x$，其中λ可視爲每項資料品質因子的

平均修正速度，在理想狀況下(使用正確的方式成功修正資料)顯然應有$\beta < \lambda x$。

5. 資料品質因子的修正費用可依與時間的關係分爲兩類：
   - 與時間相關：單位時間的修正費用爲C_2，因此每項品質因子的修正費用是$C_2(t_2 - t_1)$。(如：人力的投入)
   - 與時間無關：因此每項品質因子的修正費用是C_3。(如：物力的投注)

6. 每項資料品質因子用來爲不良資料做分類，資料品質的惡化程度$\frac{dQ}{dt}$取決於每項資料品質因子$\frac{dQ_i}{dt}(i = 1, 2, \ldots, n)$的總和，根據式(1)的公式可得：

$$\begin{aligned} \frac{dQ(t)}{dt} &= \frac{d\sum_{i=1}^{n} Q_i(t)}{dt} \\ &= \sum_{i=1}^{n} \frac{dQ_i}{dt} \end{aligned} \quad (3)$$

根據以上六項假設條件，模型構成如圖3所示之$\frac{dQ}{dt} \sim t$關係圖。在$t_0 \le t \le t_1$時段，不良資料的品質惡化速度會隨著時間的流逝成線性比例，其資料品質惡化速度爲β，同時在時刻$t = t_1$時，不良資料所佔的比例爲$\frac{dQ(t=t_1)}{dt} = q$；在$t_1 \le t \le t_2$時段，爲修正不良資料問題的時刻區段，自此資料品質的惡化程度受到控制，逐步改善，在理想情況下，資料品質的惡化速度會改善爲$\lambda x - \beta$，其中x爲欲修正的不良資料品質因子項數，每項資料品質因子的平均修正速度爲λ。在圖3中，我們假設發現不良資料開始存在的時刻爲$t_0 = 0$，時刻t_2爲完成修正資料品質問題的時間，使用者可依據實際應用需求自訂時間表，因此根據假設在時刻t_2時，將完全修正所有不良之資料($\frac{dQ(t=t_2)}{dt} = 0$)。

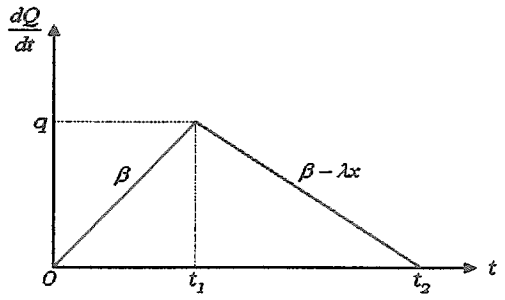

Figure 3: $\frac{dQ}{dt} \sim t$關係圖

在$t_0 \le t \le t_2$時段，爲不良資料所存在的時段，因此資料品質惡化程度爲$Q(t_2) = \int_0^{t_2} \frac{dQ}{dt} dt$，恰好是圖3中$0 \le t \le t_2$的三角形面積。根據假設條件2可計算出損失費，損失費與資料品質惡化程度成正比，比例係數爲C_1，故可得損失費(LC)如下：

$$\mathrm{LC} = C_1 Q(t_2) = C_1 \int_0^{t_2} \frac{dQ}{dt} dt = \frac{1}{2} C_1 q t_2 \quad (4)$$

設$\frac{q}{t_2 - t_1} = \lambda x - \beta$，我們得

$$\mathrm{LC} = \frac{1}{2} C_1 q t_1 + \frac{C_1 q^2}{2(\lambda x - \beta)} \quad (5)$$

時段$t_1 \le t \le t_2$係修正不良資料的時段，資料品質惡化程度$Q(t_2 - t_1) = \int_{t_1}^{t_2} \frac{dQ}{dt} dt$爲圖3中$t_1 \le t \le t_2$的三角形面積，依假設條件5，算出提昇費用如下：

$$\mathrm{IC} = C_2 x(t_2 - t_1) + C_3 x = \frac{C_2 q x}{\lambda x - \beta} + C_3 x \quad (6)$$

於是總成本支出便可由損失費與提昇費的總合計算出來：

$$\mathrm{TC} = \frac{1}{2} C_1 q t_1 + C_3 x + \frac{C_1 q^2}{2(\lambda x - \beta)} + \frac{C_2 q x}{\lambda x - \beta} \quad (7)$$

爲使總成本支出降至最低以決定要修正不良資料品質因子的項數，因此根據式子(7)，設$\frac{dC}{dx} = 0$，計算總成本C之最小值即爲計算C之一階導函數：

$$\mathrm{x} = \sqrt{\frac{C_1 \lambda q^2 + 2 C_2 \beta q}{2 C_3 \lambda^2}} + \frac{\beta}{\lambda} \quad (8)$$

由式子8所求出之x代表意義爲：以最低之提昇資料品質總成本來決定所應修正的項不良資料品質因子問題。x所得出之結果由兩部分組成，其中一部分$\frac{\beta}{\lambda}$是爲了修正不良資料所必須的最低限度，因爲β代表資料品質的惡化速度，而λ是每項資料品質因子的平均修正速度，所以這項結果是很明顯的。由圖3中也可看出，只有當$x > \frac{\beta}{\lambda}$時，斜率爲$\beta - \lambda x$的直線才會小於0，有與$t$軸相交的機會。其次，修正不良資料項數的另一部分，亦即在最低限度之上的的項數，則與實際應用時的各個參數有關。當每項資料品質因子的平均修正速度λ和提昇費用係數C_3增大時，修正項數減少；當資料品質惡化速度β、開始修正資料時之品質惡化程度q及損失費用係數C_1增加時，修正項數增加。模型的建立必須根據實際情況而定，實際應用這個評估模型時，C_1, C_2, C_3是已知常數，β與q可由檢查測試得出，λ則由資訊技術人員的素質所決定。模型的實際分佈情形則可依據實際應用與參數的調整訂定。

5 結論與未來研究方向

本文詳述如何擴展實體關係模型，用以將資料區分爲一般性資料與品質性資料，並對品質資料的表示方法與存取處理做深入的探討。其次，依據併入的品質資料與各項品質定義計算出目前資料倉儲中的資料品質惡化程度，再根據資料品質的惡化程度與成本效益評估，計算出應先加以修正的不良資料項目。綜上述，我們認爲本文可以呈現以下之成果與貢獻：

1. 加强資料的詮釋性。品質資料的擴展與併入有助於資料語意的釐清，可解決資料品質問題中的語意衝突問題。避免資料倉儲因年代久遠、人事更迭造成資料意義流失或誤用等問題。

2. 根據各項品質定義的訂定，品質資料的存在不僅可做爲評估資料是否適於使用的標準，也可協助找出錯誤發生的原因與資訊源。

3. 偵測出資料倉儲目前的資料品質程度爲何，使資料倉儲的使用不必再建立於盲目的可信度上，藉由實際檢測出的結果就可判定資料是否適於使用或是有改善資料品質的必要。

4. 使用者可依據不同的品質需求或作業需要，過濾資料倉儲中所取得的資料以完成作業。在支出成本與時間考慮的限制下，可在合理之有限成本下，計算出應先修正之不良資料項目，以達到最大之效益。

隨著資料倉儲應用的日益普及，資料品質的評估與保證課題將成爲資料倉儲建構過程的一項重要工作。我們相信認爲未來在以下各方面仍有相當的空間可再加以深入探討研究：

1. 在成本效益評估階段中資料品質惡化程度與時間的關係是建立在比較簡化的線性比假設上，實際應用時可考慮再加入其他可能的影響因素。其次，每項不良資料項目的平均修正速度爲常數λ，也可再進一步探討λ與開始修正不良資料時資料品質惡化程度q的關係，例如：q愈大，λ就愈小。這時便須對函數$\lambda(q)$作出合理的分析以得到更進一步的結果。

2. 各項品質定義的訂定可做爲評估資料是否適於使用的標準。不在標準與可接受範圍內的資料便視爲不良資料。但是對於沒有定義的不良資料便無法識別出，因此對於品質定義與規則的訂定可在人工智慧的領域上再作進一步的討論，使系統具有學習(Learning)的功能，可加强識別不良資料的能力。

3. 本文係以成本與效益的觀點來評估資料的品質問題，爲使考慮的層面涵蓋更廣，未來可再整合其他的因素與觀點發展新的評估模型，用以精進及深化資料品質的評估作業。

References

[1] Ballou, D.P.,R.Y. Wang, H.L. Pazer, and G.K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science,* Vol. 44, No. 4, pp. 462-484, April 1998.

[2] Ballou, D.P., and H.L. Pazer, "Cost/Quality Tradeoffs for Control Procedures in Information Systems," *International Journal of Management Science,* 15(6), pp.509-521, 1987.

[3] Barquin, R., and H. Edelstein, *Building, Using, and Managing the Data Warehouse,* PTR Publishing, 1997.

[4] Brackett, M.H., *The Data Warehouse Challenge,* Wiley Publishing, 1996.

[5] Elmasre, R., and S.B. Navathe, *Fundamentals of Database,* Addison-Weslry Publishing, 1994.

[6] Hufford, D., "Quality and the Data Warehouse," http:// www.datawarehouse.com/ resources/ articles/

[7] Huh, Y.U. et al., "Data Quality," *Information and Software Technology,* Vol. 32, No. 8, pp. 559-565, 1990.

[8] IBM White Paper, *The IBM Information Warehouse Solution - A Data Warehouse Plus!,* IBM Corporation, 1996.

[9] Featuring Red Brick Warehouse, "Informix Decision Frontier Solution Suite for Business - Critical Data Marts," http://www.redbrick.com/

[10] Kesh, S., "Evaluating the quality of entity relationship models," *Information and Software Technology,* Vol. 37, No. 12, pp. 681-689, 1995.

[11] Oracle Corporation, "Oracle Data Warehouse," http:// www.oracle.com/ datawarehouse/ index.html

[12] Parsaye, K., and M. Chignell, *Intelligent Database Tools and Applications : Hyperinformation Access, Data Quality, Visualization, Automatic Discovery,* Wiley Publishing, 1993.

[13] Wang, R.Y., M.P. Reddy, and H. B. Kon, "Toward Quality Data : an Attribute-based Approach," *Decision Support Systems* Vol.13, pp.349-372, 1995

[14] 楊鍵樵，"異質性環境資訊整合與資訊交換技術應用研究，"行政院環境保護署科技研究計畫(EPA-88-U1L1-03-003)，八十八年六月

[15] 楊珊珊，"資料倉儲環境中資料品質評估方法之研究，"國立台灣科技大學碩士論文，八十八年六月