

巨量 XML 真實程度模型之建構

Veracity Model of Big XML Data

陳世穎

國立臺中科技大學資訊工程系

sychen@nutc.edu.tw

陳宇威

國立臺中科技大學資訊工程系

willismax.com@gmail.com

摘要

近年來，巨量資料處理成為學界與產業界極重要的議題，眾多人力、物力投入資料分析的領域，XML (eXtensible Markup Language) 作為廣泛使用的交換資料格式，匯聚成巨量 XML 資料，具有巨量資料的 5V 特性 (Volume、Velocity、Variety、Veracity、與 Value)，其中真實性 (Veracity) 與資料分析結果的可靠與可信有關，然而巨量資料的大、快、疑、雜特性，資料真實性並非能在有限時間內清晰分辨，使用者也缺乏對理解資料真實性的處理方式。

本論文中針對 XML 資料是否具有參考價值的程度，建立巨量 XML 真實程度模型 (Veracity Model of Big XML Data)，提供分析資料參考，亦將透過資料視覺化，有效呈現 XML 文件的全域及各真實維度分析結果的視覺化模型，有助使用者理解 XML 資料的真實程度，該模型可適用在巨量資料環境，提供使用者參考。

關鍵詞：巨量資料、XML、真實模型、真實程度

Abstract

Big data processing has been of considerable importance with massive resources and effort invested in data analyzing throughout academic and industrial fields in recent years. As a widely used exchange data format, huge eXtensible Markup Language (XML) data has 5Vs of big data character, volume, velocity, variety, veracity, and value, among which the reliability and credibility of data analysis results are related to veracity. However, with massive volume, fast processing, doubtful and diversely origin, veracity of huge data is hardly achieved in a limited time; neither

have users the proper tools for distinguishing the authenticity of data.

In this study, Veracity Model of Big XML Data is built to provide reference for the analysis of data. Visualization of data is also used to present the global and real world of XML documents. A visual model of dimension analysis results helps users understand the quality of authenticity of XML data.

Keywords: Big Data, XML, Data Veracity

1. 前言

隨著科技進步，數位資料成長迅速，大量、多樣來源的數據無處不在，IBM[10]2015 年所推估，每天有超過 2.5 quintillion bytes 資料被創造，而且超過 90% 的資料是近 2 年內所產生，資料來源舉凡感測器、交通與航班訊息、金融交易資訊、社群媒體網站等，資料類型包含網站、文字、圖片或影音、GPS 訊號等。這些龐大、變動迅速且多樣的數據匯流成所謂大數據 (巨量資料)，巨量資料處理以成為學界與產業界極重要的議題，眾多人力物力投入巨量資料分析的領域，隨著資訊科技及數位生活的發展趨勢，越來越多應用以 XML 格式表示資料，而這些網路上的 XML 資料來源彙整起來將成為巨量 XML 資料。

XML 資料係指可擴展標記語言 (eXtensible Markup Language, 簡稱 XML 標記語言)，是用來建立描述結構化資料標示語言的語言，通過標記，電腦之間可以交換資料並處理資料內容，XML 標記語言由 SGML 標記語言精簡並延續 SGML 的優點，於 1995 年開始發展，於 1998 年由 W3C (World Wide Web Consortium, <http://www.w3.org/>) 發佈為標準 (XML 1.0)。XML 融合 SGML 的資料相容性與 HTML 的簡單標記法，提供一套簡單、標準並可擴展的語法，將各種資訊如文字、表格或

圖形等以原始的方式儲存；並於儲存的過程中加入可供辨識的標籤，讓伺服器或客戶端應用程式可藉此供辨識的標籤，將資訊內容做進一步處理得到所需的資訊。

巨量資料有大量(Volume)、快速(Velocity)、多樣(Variety)、真實性(Veracity)、價值(Value)等稱為「5V」的特性。在 XML 巨量資料上，XML 資料量大，節點、路徑、結構數量多(Volume)。其次，XML 文件產生速度快速(Velocity)，來自於各種關聯式資料庫、NoSQL 資料庫、網路交換文件、開放資料(open data)等都可能採用 XML 格式建立。而且，XML 文件結構多樣(Variety)，半結構化文件特性可針對各種應用服務定義、擴展、傳輸及儲存資訊。此外，由於資料來源多元、文件結構變化多、無 DTD 或 XML schema 規範的文件可能有同名異義或異名同義的元素等原因，因此 XML 文件資料上 veracity 的特性可以加以程度的區別，以鑑別文件的真實程度，提供使用者參考。對於 XML 文件的價值(Value)，XML 文件之原始內容與結構，經過處理分析後，產生的價值則由應用上之處理邏輯而異。

其中真實性表示巨量資料的真實程度，與資料是否可靠和可信有關。不可靠或不可信的資料將對資料分析、儲存與運用造成偏誤。有鑑於此，對於散佈於複雜網路上的各類文件，應有協助判斷巨量 XML 資料真實性、資料信賴與否的參考指標。

本論文目的旨在建構 XML 資料真實性模型，幫助使用者理解資料真實性。本論文以資料使用者的「資料理解性」(data understandability)角度描述 XML 文件的真實程度，建構適合描述資料真實程度(veracity degree)的「真實模型」(Veracity Model)。使用者可透過本論文所建置的真實維度、真實屬性及關鍵指標所呈現的真實值，作為判斷該 XML 文件真實性之參考。

2. 相關研究

2.1 XML similarity

XML 文件作為資料交換的儲存及交換的媒介，[14,6,15]討論比對 XML 相似度(XML similarity)的方法，約可分為三類。

第一類以原始樹編輯距離(edit distance)評估，[14]比對兩顆 XML 樹之間需要透過幾次

新增、修改與刪除節點之動作，從中尋求修改次數最少可達兩棵 XML 相同個數編輯距離；第二類僅考慮兩棵有序 XML 樹之間在新增及刪除節點兩種編輯方式之相似程度[6]。第三類的研究[15]則是建構樹編碼機制，計算兩棵樹彼此之間的最小樹編輯距離。

本論文與探討 XML similarity 之相關研究差異有下列二點。其一為巨量資料的大量與快速特性下，計算 XML 樹編輯距離將導致效能不佳。二為 similarity 的概念無法完整描述巨量資料之真實性。

2.2 Data veracity

巨量資料會因為資料收集方法、資料處理或分析技術而造成資訊偏差(biased)、模糊、與不準確。[11]以資料品質的角度定義資料真實性；而[7]以資料準確度的角度定義資料真實性。

[11]指出目前很少有討論資料真實性的研究，並提出一套關於資料真實性的理論和定義，以 objectivity/ subjectivity、truthfulness/deception 與 credibility/ implausibility 三個維度整合為「巨量資料真實指數(big data veracity index)」。此指數能有效運用於分析文章內容的資料品質(data quality)。而[7]以對於資料的信賴度(confidence)描述資料真實性，[7]指出資料真實性指的是資料準確度以及他如何相關於商業價值(business value)。

[11]與本論文的差異有下列三點。第一，[11]以文章內容的資料為主，而本論文以 XML 資料為應用對象。第二，[11]以資料品質(data quality)的觀點描述資料真實性，而本論文以資料理解性(data understandability)的觀點描述 XML 資料的真實程度。第三、[11]建立的三維模型無法反映以此資料理解性的觀點描述 XML 資料的真實程度，而我們將建構真實模型(veracity model)，可彈性的依據不同應用服務上的 XML 資料建立真實維度(veracity dimensions)，估算資料的真實程度。而且此模型可以彈性的擴展至其他應用，非僅限於巨量 XML 資料。因此，[11]與[7]分別以資料品質與資料準確度定義資料真實性；而事實上這些定義是可以包含在本論文建立的真實維度中。

3. 巨量 XML 真實程度模型

真實程度模型以針對不同的 XML 文件特性進行評價，且評價的方法應保留使用者自定義的彈性，以及適用於各種 XML 文件真實程度的泛用性為目標。

在此 XML 真實模型中，有一份基準 XML 文件 B 與被量測 XML 文件 M，文件 M 將與文件 B 相比較，估算文件 M 的真實程度。

XML 真實模型模型(veracity model)V 有 n 個真實維度 D_i ，亦即 XML 文件的真實模型為各真實維度之集合。V 之表示式為：

$$V = \{D_i\}, i=1..n。$$

在 XML 真實模型的真實維度 D_i 中，包含與此維度相關的「真實屬性 (veracity attributes)」。真實屬性可依據使用者對資料邏輯的理解進行定義。我們設計維度 D_i 將會有 d_i 個真實屬性， $P_{i,j}$ 表示真實維度 D_i 的第 j 個屬性，其表示式為：

$$D_i = \{P_{i,j}\}, j=1..d_i。$$

真實維度 D_i 與真實屬性 $P_{i,j}$ 提供描述資料真實性的面向，而為了具體量化做為可供評估真實程度的數值，模型 V 中建立各真實屬性 $P_{i,j}$ 的「量化因子」。每個真實屬性可有數個可作為量化評估的量化因子， $Q_{i,j,k}$ 表示第 i 個真實維度、第 j 個真實屬性的第 p_{ij} 個量化因子，表示式如下：

$$P_{i,j} = \{Q_{i,j,k}\}, k=1..p_{ij}$$

真實屬性 $P_{i,j}$ 可藉由各該量化因子 $Q_{i,j,k}$ 附屬之量化函數 $f(Q_{i,j,k})$ 計算真實屬性 $P_{i,j}$ 真實程度 $|P_{i,j}|$ 。而各真實屬性的真實程度因為度量的差異，而有可能相對過大或過小。因此，計算 D_i 的值時，以 $w_{i,j}$ 做為正規化各真實程度 $|P_{i,j}|$ 之係數。

$$\begin{aligned} |P_{i,j}| &= \sum_k f(Q_{i,j,k}) \\ |D_i| &= \sum_j (w_{i,j} \times |P_{i,j}|) \\ &= \sum_j \left(w_{i,j} \times \sum_k f(Q_{i,j,k}) \right) \end{aligned}$$

據此，本論文所提出的真實模型具有真實維度、真實屬性、量化因子，以計算真實程度的階層特性，如下圖 1。

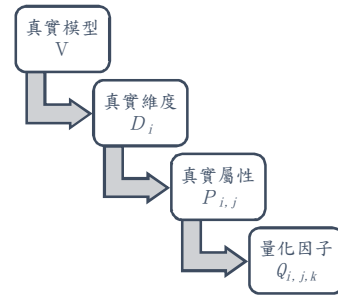


圖 1、真實模型之階層性表示

巨量 XML 真實程度模型提供使用者自定義使用方式，使用者對運用本模型可設計如何評估真實程度的標準與方法。為能使本模型使用者理解實際運作方式，本論文依據巨量 XML 文件特性，歸納出部分具有理解真實維度的設定方式，其次，我們必須決定真實維度 D_i 相關的真實屬性 $P_{i,j}$ 及量化因子 $Q_{i,j,k}$ 。在上述列舉的真實維度中，我們分別找出每個維度相關的真實屬性。表 1 為舉列說明在 XML 文件上真實維度、真實屬性及量化因子對應表。

表 1:真實維度、真實屬性及關鍵指標

真實維度	真實屬性	量化因子
D_1 =文件來源	$P_{1,1}$ =文檔宣告	$Q_{1,1,1}$ = XML 版本、 $Q_{1,1,2}$ =語言
	$P_{1,2}$ =URI	$Q_{1,2,1}$ =網域層級、 $Q_{1,2,2}$ =文檔系統
D_2 =文件結構	$P_{2,1}$ =結構規範	$Q_{2,1,1}$ = DTD、 $Q_{2,1,2}$ = XSD、 $Q_{2,1,3}$ = DataGuide
	$P_{2,2}$ =結構比對	$Q_{2,2,1}$ =文件結構階層（深度）、 $Q_{2,2,2}$ =樹寬、 $Q_{2,2,3}$ =數量、 $Q_{2,2,4}$ =節點型態
D_3 =文件內容	$P_{3,1}$ =元素	$Q_{3,1,1}$ =元素名稱、 $Q_{3,1,2}$ =元素數量
	$P_{3,2}$ =路徑	$Q_{3,2,1}$ =指定路徑、 $Q_{3,2,2}$ =路徑數量
	$P_{3,3}$ =相依性	$Q_{3,3,1}$ =父子元素、 $Q_{3,3,2}$ =兄弟元素
D_4 =時間	$P_{4,1}$ =檔案時間	$Q_{4,1,1}$ =建立時間、 $Q_{4,1,2}$ =修改時間、 $Q_{4,1,3}$ =存取時間
	$P_{4,2}$ =順序性	$Q_{4,2,1}$ =文檔時間間隔、 $Q_{4,2,2}$ =產生順序

在使用者依照 XML 真實模型定義真實維度、真實屬性及用以計算程度的關鍵指標後，所顯示，所計算出的真實值可在 n 個真實維度所建立的雷達圖表示 V_i 的座標。因此，文件 M 對於文件 B 的「真實程度」可設計為所有真實值 V_i 在其座標 D_i 所圍出的區域面積，表示為：

$$V_{M|B}(D_1, D_2, \dots, D_n)$$

如圖 2 所示，假設有六個真實維度 D_1 、 D_2 、...、 D_6 ，目前量測文件 M 對於基準文件 B 在此六個維度的值分別為 66、70、50、76、80、95，則雷達圖上這六個值所為出來的面積即為此文件 M 對文件 B 之真實程度 $V_{M|B}$ 。

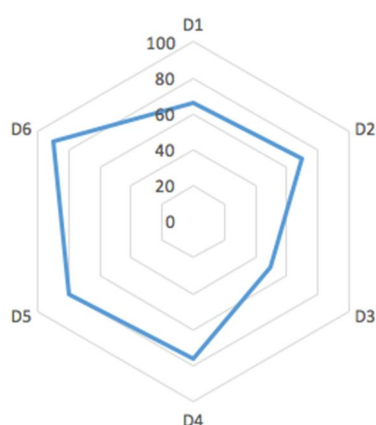


圖 2、真實程度(Veracity Value)

4. 應用範例

本論文以臺灣自動氣象站-氣象觀測資料[3]做為範例資料，透過本論文提出的真實模型，計算 XML 文件的真實程度。

4.1 範例文件

本範例以臺灣氣象資訊作為基準文件 B [3]，另一方面，以美國國家海洋暨大氣總署 (NOAA) 氣象資料[12]、台灣政府開放交通資料[16]分別做為被量測文件 M_1 與 M_2 ，進行真實程度計算，以說明本論文建構之以資料理解性觀點建立的真實模型與其真實程度。

基準文件 B 為「臺灣自動氣象站-氣象觀測資料」，該資料為臺灣資料開放平台[20]熱門引用數據。文件 B 每 10 分鐘更新一次，標籤包含測站 ID、觀測時間、溫度、濕度、風向及風速、觀測時間前推 15 分鐘內發生最大風速等資訊，該文件結構計有 9,471 行、40 個欄位，文件最深 5 層，係一般常見的 XML 資料設計方式，表 2 為自動氣象站資料標籤名稱說明文

件[5]、表 3 為該 XML 資料文件內容摘要[4]、圖 3 為該 XML 樹狀內容。

表 2、自動氣象站 XML 標籤

標籤名稱	中文說明
stationId	觀測站 ID
STNM	觀測站編號
OBS_TIME	觀測資料時間
TIME	未使用
LAT	緯度 (座標系統採 TWD67)
LON	經度 (座標系統採 TWD67)
ELEV	高度，單位：公尺
WDIR	風向，單位：度，風向 0 表示無風
WSD	風速，單位：公尺/秒位 攝氏
TEMP	溫度，單位：攝氏
HUMD	相對濕度，單位 百分比率，此處以實數 0-1.0 記錄
PRES	觀測站氣壓，單位：百帕
H_24R	日累積雨量，單位：毫米
WS15M	觀測時間前推十五分鐘內發生最大風速，單位：公尺/秒
WD15M	觀測時間前推十五分鐘內發生最大風風速，單位：度
WS15T	觀測時間前推十五分鐘內發生最大風的發生時間，hhmm (小時分鐘)
CITY	縣市
CITY_SN	縣市編號
TOWN	鄉鎮
TOWN_SN	鄉鎮編號

表 3、自動氣象站 XML 內容摘要

<pre><?xml version="1.0" encoding="UTF-8"?> <cwboappendata xmlns="urn:cwb.gov.tw:cwbcommon:0.1"> <identifier>ca5efb48-b33a-4ea9-ac83-48c3b5524 179</identifier> <sender>weather@cwb.gov.tw</sender> <sent>2016-11-12T09:35:52+08:00</sent> <status>Actual</status> <msgType>Issue</msgType> <dataid>CWB_A0001</dataid> <scope>Public</scope></pre>
--

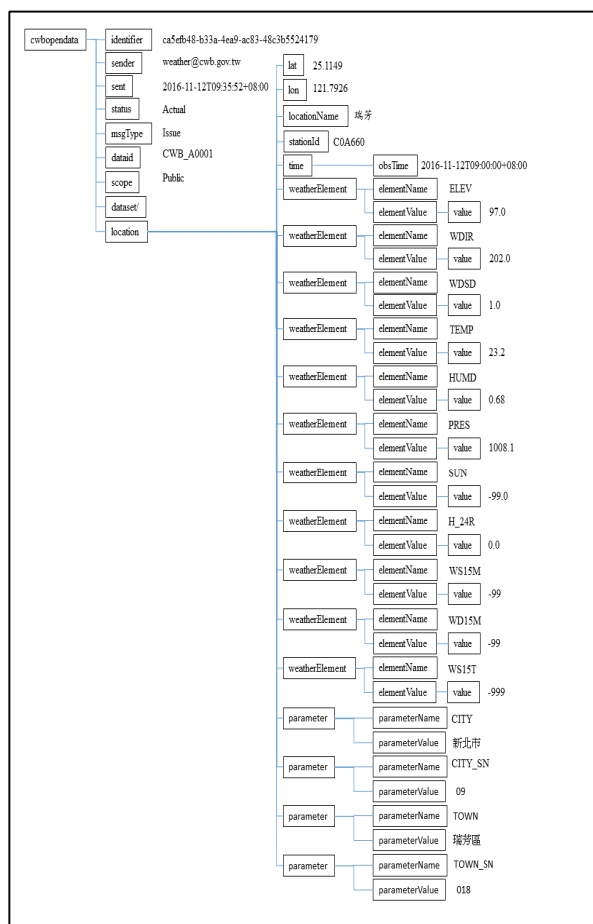


圖 3、自動氣象站 XML 樹狀內容

測量文件 M_1 作採用美國國家海洋暨大氣總署（NOAA）氣後觀測資訊[21]。NOAA 為了為預測環境變化、保護與管理海洋資源，具有專業氣象預測資訊，提供氣候監測、預先警示數據。在目前天氣狀況 XML 資料中，提供美國 1,800 個地點的氣候資訊，並提供 XML 及 RSS 以幫助訊息的發布，在同一時間點下載 NOAA 分享的 XML 文件壓縮檔，解壓縮後 XML 文件計有 4,672 份，由於每份文件結構相同，我們將以其中一份 XML 文件資訊進行真實程度估算。表 4 為 NOAA 氣後觀測資料標籤名稱說明文件[8]，表 5 為該 XML 資料文件內容摘要[13]、圖 4 為其 XML 樹狀內容。

表 4、NOAA 氣候觀測 XML 標籤

標籤名稱	說明
Credit	固定宣告為 NOAA's National Weather Service
credit_URL	固定宣告為 http://weather.gov/
url	固定宣告為 http://weather.gov/images/xml_1_0go.gif

Title	固定宣告為 NOAA's National Weather Service
Link	固定宣告為 http://weather.gov
suggested_pickup	固定宣告為 15 minutes after the hour
suggested_pickup_period	固定宣告為 60
Location	觀測站地點
station_id	觀測站 ID
Latitude	緯度
longitude	經度
observation_time	最近觀測時間，如：Last Updated on Dec 26 2016, 12:00 am AST
observation_time_rfc822	最近觀測時間，如：Mon, 26 Dec 2016 00:00:00 -0400
temperature_string	溫度
temp_f	華氏溫度，如：21.0
temp_c	攝氏溫度，如：-5.9
wind_string	風速及風向資訊字串，如：from the North at 18.4 gusting to 27.6 MPH (16 gusting to 24 KT)

表 5、NOAA 氣候觀測 XML 內容摘要

```

<?xml version="1.0" encoding="ISO-8859-1"?>
<?xml-stylesheet href="latest_ob.xml"
type="text/xsl"?>
<current_observation version="1.0"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://www.weather.gov/view/current_observation.xsd">
  <credit>NOAA's National Weather Service</credit>
  <credit_URL>http://weather.gov/</credit_URL>
  ...
</current_observation>

```

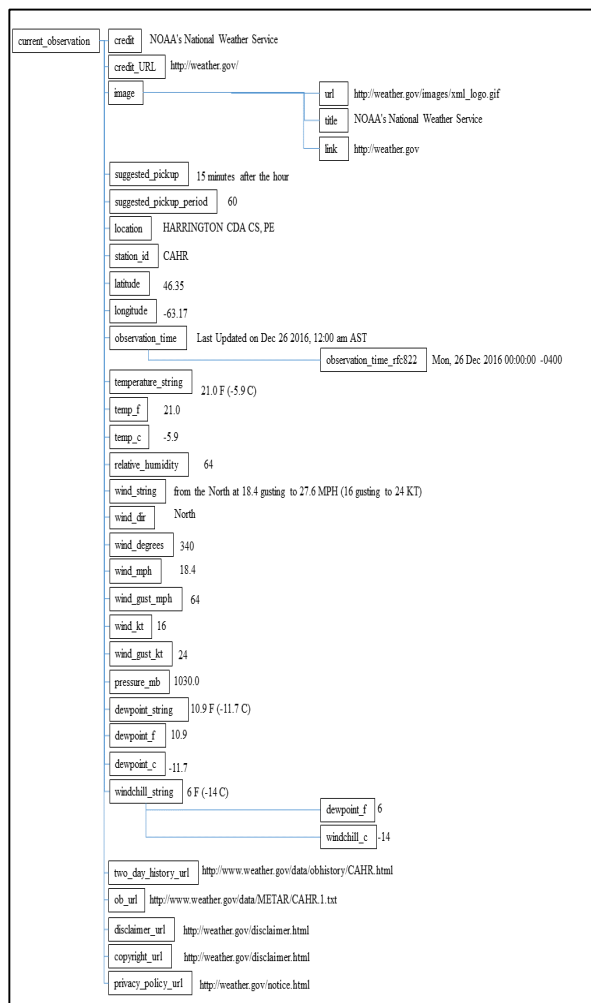



圖 4、NOAA 氣候觀測 XML 樹狀內容

測量文件 M_2 為交通部開放資料「VD 一分鐘動態資訊」[16]，該資料庫資料來源為「高快速公路整體路網交通管理系統」與「高速公路電子收費系統」，資料格式係依據交通部發布之「路側設施即時交通資訊發布標準格式」及「國道高速公路電子收費交通資料蒐集支援系統 (Traffic Data Collection System, TDCS) 使用手冊[22]。提供 VD 狀態及偵測值的標準格式，每分鐘提供各車道不同車種的速率、流量及佔有率，並標示 VD 運作狀況。表 6 為資料標籤名稱說明 [2]、表 7 為內容摘要[1]、圖 5 為 XML 樹狀內容。

表 6、交通部 VD 動態資訊 XML 標籤

標籤名稱	說明
vdid	設備代碼
status	狀態

vsrdir	車道方向，以 0 和 1 表示
vsrid	車道代碼
carid	車種代碼
speed	依車道逐一詳列 1 分鐘平均速率偵測值(單位：kph)，如：24。
volumn	依車道/車種逐一詳列 1 分鐘流量偵測值，如：66。
Lane-occupy	依車道逐一詳列 1 分鐘佔有率偵測值(單位：%)，如：50。
Data-collecttime	資料蒐集時間

表 7、交通部 VD 動態資訊 XML 內容摘要

```
<?xml version="1.0" encoding="utf-8"?>
<XML_Head version="1.1" listname="VD 一分鐘動態資訊" updatetime="2009/10/06 11:31:23" interval="60">
  <Infos>
    <Info vdid="63000VD-1" status="0" datacollecttime="2009/10/06 11:30:00">
      <lane vsrdir="0" vsrid="1" speed="24" laneoccupy="30">
        <cars carid="S" volumn="66" />
        14
        <cars carid="T" volumn="74" />
        <cars carid="L" volumn="60" />
      </lane>
      <lane vsrdir="0" vsrid="2" speed="26" laneoccupy="30">
        <cars carid="S" volumn="66" />
        <cars carid="T" volumn="74" />
        <cars carid="L" volumn="60" />
      </lane>
      ...
    </XML_Head>
```

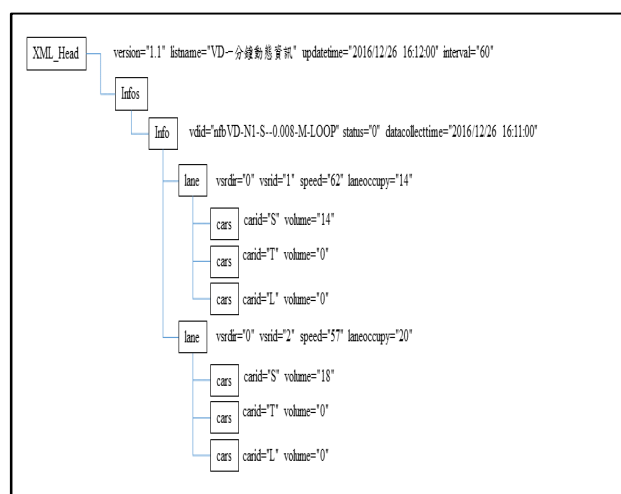


圖 5、VD 一分鐘動態資訊 XML 樹範例

4.2 XML 真實程度模型設計範例

本實驗舉例 XML 真實維度至分別為資料來源 D_1 、資料結構 D_2 、資料內容 D_3 、資料時間 D_4 等維度，分別就各維度所對應的真實屬性與關鍵指標如下。

表 8:真實維度、真實屬性及關鍵指標

真實維度	真實屬性	量化因子
D_1 =文件來源	$P_{1,1}$ =文檔宣告	$Q_{1,1,1}$ =XML 版本 $Q_{1,1,2}$ =語言
	$P_{1,2}$ =URI	$Q_{1,2,1}$ =網域層級
D_2 =文件結構	$P_{2,1}$ =結構規範	$Q_{2,1,1}$ =結構規範
	$P_{2,2}$ =結構比對	$Q_{2,2,1}$ =階層（深度） $Q_{2,2,2}$ =不重複樹寬
D_3 =文件內容	$P_{3,1}$ =元素	$Q_{3,1,1}$ =關建元素名稱
	$P_{3,2}$ =路徑	$Q_{3,2,1}$ =指定路徑
	$P_{3,3}$ =相依性	$Q_{3,3,1}$ =父子元素
D_4 =時間	$P_{4,1}$ =檔案時間	$Q_{4,1,1}$ =建立時間
	$P_{4,2}$ =順序性	$Q_{4,2,1}$ =文檔時間間隔

(1)文件來源 D_1

在真實維度-文件來源 D_1 中，我們設定真實屬性為「文檔宣告」 $P_{1,1}$ 及「URI」 $P_{1,2}$ 。文檔宣告係指 XML 文件第一行所標示的內容，如：`<?xml version="1.0" encoding="UTF-8" standalone="yes"?>`，其中版本宣告（version）為必要項，編碼（encoding）及獨立宣告（standalone）為選擇項，比對的文件是否有宣告，抑或是是否有相同宣告，即可做為評判真實程度參考。我們設計 XML 版本是否批配為關鍵指標 $K_{1,1,1}$ ，編碼是否相符為 $K_{1,1,2}$ 。

「URI（Uniform Resource Identifier）」是 XML 文件提供一種避免資料元素名互相稱衝突的方法，使用其獨特性用來識別某項資源，只要將 URI 的參引加到元素或屬性名稱的前面，就能使名稱具有單一性。通常，URI 以網域名稱（domain name）及具有路徑識別性的文件系統所組成，而網域名稱有階層性，頂級域名（Top-level domains, TLDs）及國家代碼頂級域名（country code top-level domain, ccTLDs），前者如：com、gov、org、edu；後者如：tw（台灣）、de（德國）、eu（歐盟）、jp（日本）、tw

（台灣）等，除了頂級域名，還有二級域名（SLD，second-level domain）。我們可以設計關鍵指標 $K_{1,2,1}$ 網域層級依照 URI 網域的階層性進行比對，層級由高至低比對結果作為評估方法。

(2)文件結構 D_2

在真實維度-文件結構 D_2 中，可以運用文件 B 與 M_1 、B 與 M_2 之間兩者是否依循相同「結構規範」，如是否皆遵循相同的要遵守 DTD 或 XML Schema（XSD），或者即便參照不同的 DTD 或 XSD，透過結構映射仍有相似的結構；「結構比對」方面，我們採用 DataGuide[9]進行文件結構摘要，DataGuides 僅表示 XML 文件所有路徑的摘要結構，相同標籤名稱序列構成的所有路徑在該 XML 文件的 DataGuide 中僅以該序列表示路徑摘要，不會受到如 DTD 及 XSD 為事前宣告之文件結構所限制。圖 6 至 8 分別為臺灣自動氣象站、NOAA 氣候觀測、VD 一分鐘動態資訊的 DataGuide，使用者可透過路徑摘要進行文建結構分析及比對。其中，在圖 6 臺灣自動氣象站的 DataGuide 比原始 XML 樹簡化許多，原因是許多不同資料使用 weatherElement 標籤，以其子標籤 elementName 內容紀錄溫度、相對濕度、大氣壓力等資訊，及以子標籤 value 內容紀錄其值；圖 7 NOAA 氣候觀測 DataGuide 未簡化應有資訊、圖 8 VD 一分鐘動態資訊 DataGuide 大幅度簡化重複的標籤。我們以結構規範是否屬同樣的 DTD 或 XSD 作為關鍵指標 $K_{2,1,1}$ ，以文件最大階層（深度）為 $K_{2,2,1}$ ，文件結構不重複同階層最大寬度（即以 DataGuide 觀察）為 $K_{2,2,2}$ 樹寬，進行文件結構的真實程度計算。

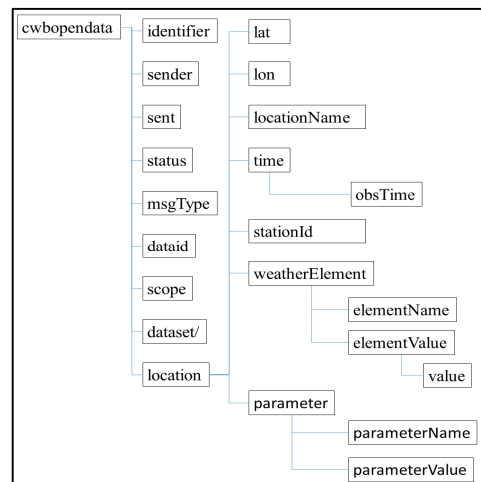


圖 6、臺灣自動氣象站 DataGuide

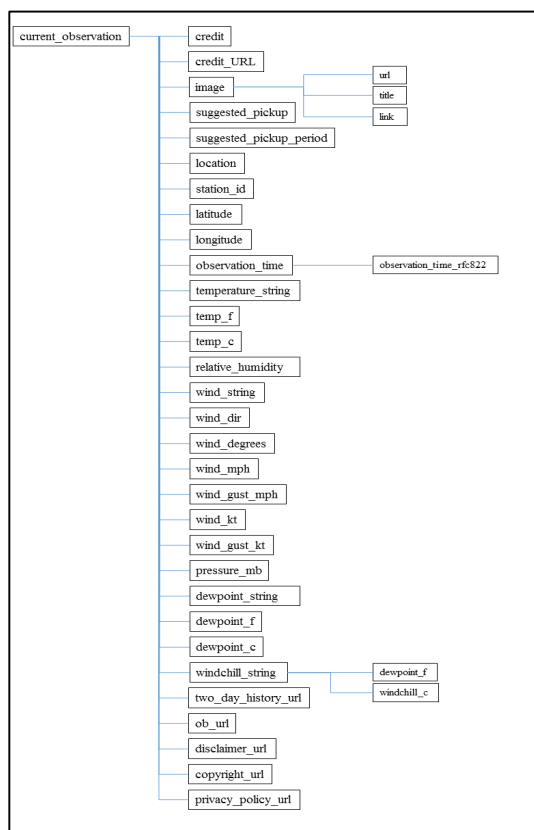


圖 7、NOAA 氣候觀測 DataGuide

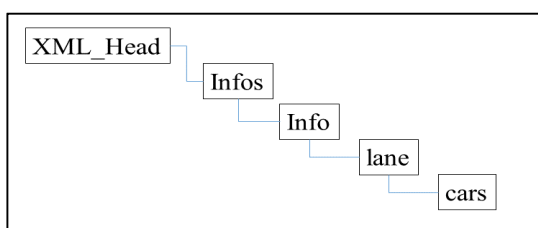


圖 8、VD 一分鐘動態資訊 DataGuide

(3)文件內容 D₃

接著我們依據 B 文件的資料內容，以 DataGuide 設計可突顯其真實性的關鍵元素，「元素」、「路徑」及「相依性」皆可作為比對真實性的屬性。關鍵元素名稱 $K_{3,1,1}$ 指定為表示溫度的 Temp，則 M_1 有相似的 temp_f 及 temp_c 標籤， M_2 則無類似標籤；而在指定路徑 $K_{3,2,1}$ 及父子元素 $K_{3,3,1}$ 透過 DataGuide 觀察指定任何元素及路徑皆無真實程度可言。

(4)文件時間 D₄

文件時間 D_4 我們採用可依建立檔案的「檔案時間」、「順序性」來判斷其文件特徵。使用者可以設計符合檔案建立時間的邏輯，例如擷取即時性資訊時，資料建立時間與擷取資料時

間相差 1 月，則缺乏真實性；另外如果以固定週期發佈的資料，是否間隔週期相同也可以作為真實評估標準。建立時間 $K_{4,1,1}$ 與資料時間 B，文件與 M_2 文件幾乎即時產生， M_1 文件則因美國各區站點回報時間不同而有差異；順序性 $K_{4,2,1}$ 我們以 B 文件每間隔 1 小時為準， M_1 則以每 20 分鐘更新、 M_2 以 1 分鐘進行更新，從中進行真實程度評估。

表 9、真實程度

量化因子	使用者定義量化函數功能	文件 B	文件 M_1	文件 M_2
$Q_{1,1,1}$ =XML 版本	Version 真實程度	1.0	1.0	1.0
$Q_{1,1,2}$ =語言	編碼真實程度	UTF-8	ISO-8859-1	UTF-8
$Q_{1,2,1}$ =網域層級	URI 真實程度	xmlns="urn:cwb:gov.tw:cwbcommon:0.1"	xmlns:xsd="http://www.w3.org/2001/XMLSchema"	無
$Q_{2,1,1}$ =DTD 或 XSD	同採 DTD 或 XSD	XSD	XSD	XSD
$Q_{2,2,1}$ =文件最大深度	文件最大深度	5	3	5
$K_{2,2,2}$ =文件結構不重複同階層最大寬度	文件結構不重複同階層最大寬度	9 (第 2 層)	31 (第 2 層)	1
$Q_{3,1,1}$ =關鍵元素名稱	關鍵元素真實程度：指定 temp	temp	有相似：temp_f temp_c	無
$Q_{3,2,1}$ =指定路徑	指定路徑：location 至 time 至 obsTime	無	無	無
$Q_{3,3,1}$ =父子元素	指定父子元素 weatherElement 與 elementName	無	無	無

$Q_{4,1,1}$ =建立時間	產生時間	即時	各區站點回報時間不同	即時
$Q_{4,2,1}$ =文檔時間間隔	文件產生時間間隔(頻率)	60 分鐘	20 分鐘	1 分鐘

$P_{4,2}$ =順序性	$Q_{4,2,1}$ =文檔時間間隔	40	2
----------------	---------------------	----	---

依據本論文真實模型的真實維度、真實屬性及量化因子，使用者即可依據資料特性自行設計不同量化函數，最後設定出真實程度數值。依據使用者自定義的量化函數計算出各個量化因子，並統一尺度範圍為 0 至 100，如表 10，接著使用者以模型的階層性計算各維度權重。我們舉例計算後量測文件 M_1 於基準文件 B 在此 4 個維度的值分別為 45、65.5、25、55， M_2 4 個維度的值分別為 75、77.8、0、51，則雷達圖測繪真實程度如圖 9，如以雷達圖各維度所圍之面積做為真實程度值， $V_{M1|B}$ 與 $V_{M2|B}$ 分別為 4,217.5 及 3,218.75。使用者即可依據所計算出的真實值做為資料真實程度參考。

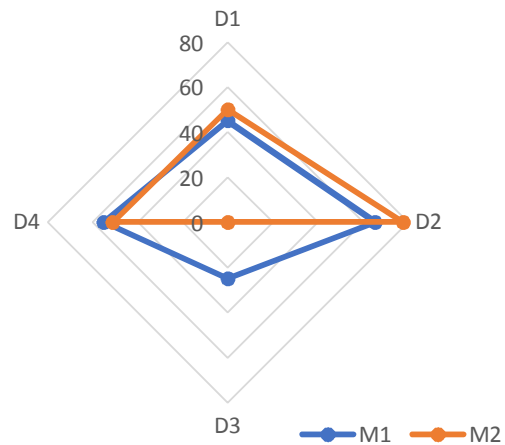


圖 9、文件 M_1 、 M_2 對文件 B 之真實程度視覺化呈現

表 10、真實程度

真實維度	真實屬性	量化因子	M_1 真實程度	M_2 真實程度
D_1 =文件來源	$P_{1,1}$ =文檔宣告	$Q_{1,1,1}$ =XML 版本	100	100
		$Q_{1,1,2}$ =語言	40	100
	$P_{1,2}$ =URI	$Q_{1,2,1}$ =網域層級	20	0
D_2 =文件結構	$P_{2,1}$ =結構規範	$Q_{2,1,1}$ =結構規範	100	100
	$P_{2,2}$ =結構比對	$Q_{2,2,1}$ =階層(深度)	33	100
		$Q_{2,2,2}$ =不重複樹寬	29	11
D_3 =文件內容	$P_{3,1}$ =元素	$Q_{3,1,1}$ =關鍵元素名稱	75	0
	$P_{3,2}$ =路徑	$Q_{3,2,1}$ =指定路徑	0	0
	$P_{3,3}$ =相依性	$Q_{3,3,1}$ =父子元素	0	0
D_4 =時間	$P_{4,1}$ =檔案時間	$Q_{4,1,1}$ =建立時間	70	100

5. 結論

本論文建立巨量 XML 真實程度模型，適用於巨量資料情境之下，協助資料使用者依照本論文模型架構自行定義量化函數，以資料視覺化協助判讀 XML 文件真實性，提供使用者評估真實與否的參考。後續研究者可就模型介面設計、視覺化應用、模型應用效能等面向繼續發展。

誌謝

本研究由科計部計畫補助支持，計畫編號 MOST 104-2221-E-025 -011，特此誌謝。

參考文獻

- [1] 交通部. 路側設施即時交通資訊. http://tisvcloud.freeway.gov.tw/vd_value.xml.gz. Accessed January 10, 2017.
- [2] 交通部. 路側設施即時交通資訊. 發布標準格式 v1.1. [http://tisvcloud.freeway.gov.tw/路側設施即時交通資訊發布標準格式\(含補充\).pdf](http://tisvcloud.freeway.gov.tw/路側設施即時交通資訊發布標準格式(含補充).pdf). Accessed January 10, 2017.
- [3] 國家發展委員會. 自動氣象站-氣象觀測資料. <http://data.gov.tw/node/9176>. Accessed January 10, 2017.
- [4] 國家發展委員會. 資料資源. <http://data.gov.tw/node/gov/resource/3537>. Accessed January 10, 2017.

- [5] 國家發展委員會. 自動氣象站資料集說明檔. http://opendata.cwb.gov.tw/opendatadoc/DI_V2/A0001-001.pdf. Accessed January 10, 2017.
- [6] Algergawy, A., Nayak, R., & Saake, G. (2010). Element similarity measures in XML schema matching. *Information Sciences*, 180(24), 4975-4998.
- [7] Big data means big opportunities for chemical companies. <https://home.kpmg.com/xx/en/home/insights/2016/07/big-data-means-big-opportunities-chemical-companies.html>. Accessed January 10, 2017.
- [8] Chery P. Product document description: NWS current observations using RSS and XML based formats 5/7/2004 part I -mission connection part II -technical description, (2008). http://products.weather.gov/PDD/NWS_Current_Observations_RSS_XML.pdf. Accessed January 10, 2017.
- [9] Goldman, R., & Widom, J. (1997). Dataguides: Enabling query formulation and optimization in semistructured databases.
- [10] IBM, Big Data and Analytics. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>. 2015.
- [11] Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible?. *Advances in Classification Research Online*, 24(1), 4-15.
- [12] Service NoNW, Team NIS. XML data feeds - current conditions - NOAA's national weather service. http://w1.weather.gov/xml/current_obs/. Accessed January 10, 2017.
- [13] Service NoNW, Team NIS. XML data feeds - current conditions - NOAA's national weather service. http://w1.weather.gov/xml/current_obs/all_rss.zip. Accessed January 10, 2017.
- [14] Tai, K. C. (1979). The tree-to-tree correction problem. *Journal of the ACM (JACM)*, 26(3), 422-433.
- [15] Tekli, J., & Chbeir, R. (2012). A novel XML document structure comparison framework based-on sub-tree commonalities and label semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11, 14-40.
- [16] TISV. <http://tisvcloud.freeway.gov.tw/>. Accessed January 10, 2017.