以本體論及語意分析為基礎的二階段藝文資訊整合技術

應鳴雄[1]　　鄧光宏[2]
[1]中華大學資訊管理學系　mhying@chu.edu.tw
[2]中華大學資訊管理學系　lalaethan@hotmail.com

摘要

本研究嘗試以本體論、語意分析及網頁探勘技術為基礎，提出二階段藝文資訊整合技術的藝文資訊整合系統(AIIS)。第一階段將利用代理人技術至多個藝文檢索網站自動取得藝文資訊，並建置藝文資訊本體架構以將各網站之資訊進行格式的統一與整合;第二階段使用語意分析技術對非結構性的藝文新聞內容進行資訊萃取，以改善藝文活動資訊不足的問題。本研究的成效結果顯示，在第一階段對結構性藝文內容的資訊擷取率可達 100%，並能將來自不同網站的資訊內容，有效地進行資訊內容及格式的整合;而在第二階段的實證結果則顯示， AIIS 不僅在藝文資訊的豐富度高於其它藝文資訊檢索網站，亦能提供即時完整的藝文資訊讓民眾進行查詢。

關鍵詞：網頁探勘、本體論、語意分析、資訊整合、代理人

1. 緒論

近年來，隨著生活水準與人文素養的提升，參與藝文活動已成為民眾假日的主要休閒活動。根據全國各縣市藝文展演活動的統計資料顯示，自 2004 年到 2010 年，全國各地的藝文活動場次已從 24,702 場成長至 57,289 場，成長率約為 232%，而觀賞人次則由 95,819 人增加至 172,832 人，增多了 7 萬 7 千多人(Council for Cultural Affairs, 2010); 2009 年社會指標統計資料則表顯示，我國平均每人出席藝文活動之次數從 2004 年的 4.2 次已增加至 2009 年的 7.1 次(Directorate General of Budget, 2010)，這些數據不僅顯示出民眾對藝文活動需求與日俱增，更突顯藝文資訊對於藝文參與者的重要性。

隨著網際網路興起之後，網路上的藝文活動檢索網站也四處林立，因此導致藝文活動資訊分散在網路各處，雖然民眾可選擇的藝文資訊查詢管道增加，但確也造成民眾需要花費更多的時間於查詢所需的完整藝文活動。目前網路上的藝文資訊檢索網站，大多採用結構性的方式來呈現藝文資訊，但各網站用字遣詞與排版的呈現方式皆不盡相同，因此想了解藝文訊息的民眾便需要去適應各個資訊檢索網站的資料檢索與資料呈現方式。而不同的藝文資訊檢索網站，所發佈的藝文活動資訊來源管道可能不同，因此容易導致不一致資訊的情形發生。例如藝文資訊網站 A 與藝文資訊網站 B 所提供的藝文事件，指的是同一則藝文事件，但兩者提供的活動地點卻不同，或者這二個網站針對相同的藝文資訊，卻各別擁有對方網站沒有的資訊。此外，許多藝文活動資訊檢索網站只提供最初的藝文活動資訊，許多藝文活動若因天氣或其他因素而停止演出，或發生活動時間、地點、展演內容…等訊息變更，，卻未提供這些訊息的即時更新，而這些活動變動的訊息，卻經常由即時的網路新聞提供，因此如何結合即時新聞資訊來提升藝文活動資訊的正確性，便成了值得研究的議題。

為了解決上述問題，本研究嘗試以本體論、語意分析及網頁探勘技術等概念，提出一個以二階段藝文資訊整合技術為基礎的藝文資訊整合系統(Arts Information Integration System, AIIS)。本研究共分為二個階段，第一個階段將採用網頁探勘技術建置一個資訊代理人在網際網路上自動搜集藝文資訊，且依據各網站呈現方式的差異，設計一個可整合藝文活動資訊之藝文資訊本體架構，以確保 AIIS 系統在進行各家藝文檢索網站的異質資料整合時，可將不同來源的藝文資訊轉換成標準化的格式，並透過資訊衝突決策函數以解決各式資訊衝突(Information Conflict)，再針對各種不一致資訊的情況進行分類及整合;第二階段，本研究嘗試使用語意分析技術(Semantic Analysis)針對非結構性的網路新聞進行資訊的萃取，以提升 AIIS 藝文資訊的完整性與豐富性,以達至系統的精確性、簡潔性、完整性與一致性的資訊品質標準(Loiacono, 2000; Wixom & Todd, 2005)。

2. 文獻探討

資訊整合通常用於企業上，例如 ERP 系統。由於 XML 可將資料重新定義，使資料擁有具結構性的形式(Erik, 2001)，而使用 XML 除了可以更容易將資料結構性外，也能有效幫助資料管理，並使成本降低(Kalakota & Whinston, 1997)。因此，藉由 XML 的方便性，可對異質資料庫甚至是異質平台上的資料，進行資訊整合。

2.1 本體論

本體論原起源於哲學領域，後來被積極使用於電腦科學領域中，在資訊整合、資訊表達、知識工程、知識分享以及自然語言處理等諸多應用上皆可看到本體論的蹤跡。許多學者曾對本體論進行定義，內容彙整如表一。本體論最早被用來定義真實世界的基本特性,它可用以解釋真實世界中真實的存在及關係,並可進行有系統的說明(Bunge, 1977; Smith & Welty, 2001)。本體論可被視為是一種階層式構架組合，因此可將真實世界中各領域或任務中所蘊含的知識加以分類，並明確描述出概念之間的關連性。而藉由本體論的特性，將想法概念轉換成實體概念，便可達到知識分享的效果以及目的。

表一: 本體論相關定義

本體論定義	學者
本體論為基本的術語與關係，用於組合術語(terms)和關係(relations)以定義詞彙延伸的一種規則。	Neches et al.(1991)
本體論可說明特定任務或領域之知識分類描述。	Alberts(1993)
本體論是一種概念化且明確之描述，可將某領域或實體現象抽象化，並且詳細明確的描述概念之間的關連關係。	Gruber(1993)
本體論為一種將知識代理人的想法表達成實體概念的理論。	Wielinga& Schreiber(1993)
本體論為一種具有概念性分享的正式規範形式	Uschold & druninger(1996); Borst et al.(1997); Studer et al.(1998)
本體論是一種基於階層式架構之詞彙組合，且會受到其所應用的特定領域及任務所影響	Swarout et al.(1997); Van Heijst et al.(1997)
本體論可視為一種邏輯理論，用來說明一系列詞彙的特定意涵。	Guarino(1998)

資料來源:本研究

2.2 網頁探勘

網頁探勘(Web Mining)即是將資料探勘(Data Mining)的技術結合 Web 技術後，應用於網際網路上，以發現有用的資訊並加以分析(Caverlee et al., 2004; Nie & Kambhampati, 2004; Cooley et al., 1997)。而原本資料探勘技術大多是應用在資料庫上(Agrawal & Srikant, 1995; Ceri et al., 2002; Chen et al., 1996)，它主要是從資料庫的所有資料裡，萃取出具有潛力以及資料之間的關係，以供研究者做進一步的分析(Hebel, 1998)。Web Mining 大致分為三類，網站內容探勘(Web Content Mining)、網站使用探勘(Web Usage Mining)及網站結構探勘(Web Structure Mining)(Cooley et al.,1997;Kosala & Blockeel,2000)。

2.3 語意網

Berners-Lee 於 1999 年在網路上放置第一個網頁瀏覽器以及網頁伺服器提供人們下載與使用後，不僅引發了瀏覽器的爭奪戰，更掀起了一場網路使用習慣的革命，也改變了網際網路上資訊的傳遞以及取得的方式(Berners-Lee, 1999)。由於目前的全球資訊網的構成要素為統一資源識別碼（Universal Resource Identifier)、超文字傳輸協定（Hypertext Transform Protocol）以及超文字標記語言（Hypertext Markup Language)，人們必需經由這三種要素所組成的資訊載體，再透過超連結將資訊進行串連，才可以看到想要之訊息。然而這些資訊僅有人類可以看懂，機器並無法了解其中之

涵意。因此，語意網的概念便是讓電腦藉由超連結了解關鍵詞的定義，再將該關鍵字做推理，以取得其語意資料之意義，讓人機之間能有更好的互動合作，使機器能進一步處理並「理解」資料(Berners-Lee, 1999)。語意網其實就是能提供電腦閱讀及理解網頁中之涵義的一種網路內容形式，它有助於概念的溝通與知識體系的整合(Huang, 2003)。

2.4 中文斷詞

中文文句結構與英文文句結構的組成方式並不同，英文文句的詞彙間通常會以空格來區隔，而中文文句詞彙間則是緊緊相鄰，並無任何區隔標記，所以中文斷詞並不像大多數歐美語系國家的語文斷詞，可輕易的判斷出詞彙間的不同(Chen et al., 2000)。此外，中文的句法(Syntactic)和語意(Semantic)的基本單位是「詞」而非「字」，單獨的中國字未必是語意分析的最小單位(許菱祥，1986)，因此中文語意分析技術相較於歐美語系國家的語意分析技術，則有較高的困難度。而以非結構性的藝文活動內容，若要能夠分析其所闡述的意圖，就需要先對內容進行斷詞處理，以判斷出使用者的查詢意圖。

3. 研究方法與設計

3.1 藝文活動資訊本體概念設計

本研究為了使 AIIS 系統能瞭解藝文活動的資訊內容，本研究先針對許多藝文資訊檢索網站中所呈現的藝文資訊格式進行了解，並分析各網站藝文資訊所包含的結構及元素，最後提出一個藝文資訊的知識結構本體架構，以期能儲存藝文活動資訊的內容及結構，並將藝文活動內容進行格式統一及資訊整合(如圖 1)。本研究提出的藝文活動資訊本體，包括了活動名稱、梯次、時間、活動地點、辦理單位、表演者、類別、展覽說明、聯絡方式、參加對象限制、入場方式以及交通等元素，藉以描述完整的藝文活動資訊。

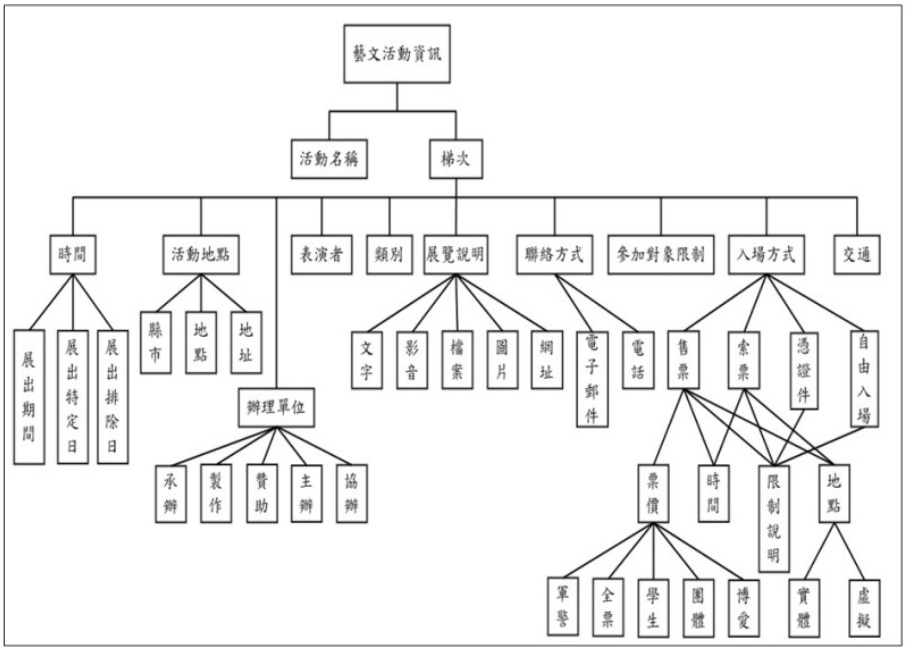


圖 1 藝文活動資訊本體

3.2 藝文資訊代理人流程設計

一般民眾在查詢藝文活動時，需要花費很多的時間四處查詢藝文活動的相關資訊，為了減輕民眾的資訊查詢負擔，本研究使用資料探勘技術，設計一支代理人程式(Agent)，並透過電腦的高速運算，不斷重覆執行網頁探勘的資料抓取動作，其流程說明如下(圖 2):

(1) Agent 會先至資料庫讀取藝文網站起始網頁地址。
(2) 下載新刊登的藝文活動之網頁內容。
(3) 判斷目前 URL 中的藝文網站中，是否已取得所有新刊登的藝文活動網頁內容。若尚有新刊登的藝文活動網頁內容未完全取得，則繼續下載該 URL 中新刊登的網頁內容，待全部的藝文活動網頁內容都取得完成，則繼續下一步驟。
(4) 判斷是否已經將資料庫中所有藝文活動網站起始網頁地址讀取完畢，若還沒讀取完畢，則繼續讀取資料庫中尚未讀取之起始網頁地址，若讀取完畢則將所有新刊登之網頁內容(HTML)存入資料庫，進行資訊解析。

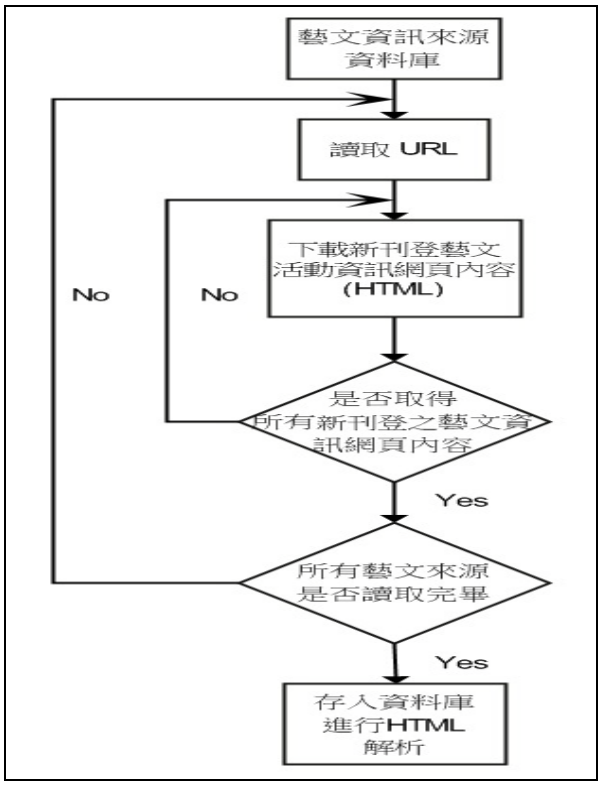


圖 2 代理人流程

3.3 網頁資料處理與實例解析

(1) 非結構性資訊解析

本研究以雅虎網路新聞藝文版中的網頁為例，說明如何解析(Parse)非結構性藝文描述的網頁內容。從圖 3 可以看到框起來的地方才是 HTML 中所需要比對的內容，這段內容置放於何處，是否有固定擺設於 HTML 的網頁語法中。從圖 4 可以看到兩者的新聞內容皆被放置於 以及 中，也就是圖中所標示的代理人解析位置，本研究所撰寫之 Agent 會將處於該位置之新聞內容抓取下來進行分析後，再將擷取出來的資訊儲存至資料庫中。


教部舉辦「再造家庭價值」世界咖啡館活動

更新日期:2011/07/08 16:36　徐詠絮

教育部今天（8）在新竹市舉辦「愛家515－再造家庭價值」世界咖啡館活動，邀請專家學者、國中小校長教師、家庭教育專業人員及家庭教育中心代表等共130人，針對婚姻價值、生育價值、家務價值、親職價值及祖孫互動價值等5項議題進行4回合小組討論，希望透過對話交流凝聚共識，研擬再造家庭價值的教育策略，作為未來家庭教育工作推動的重要參考。

(2011-07-08 16:23:54 徐詠絮)

圖 3 非結構性資訊原始網頁

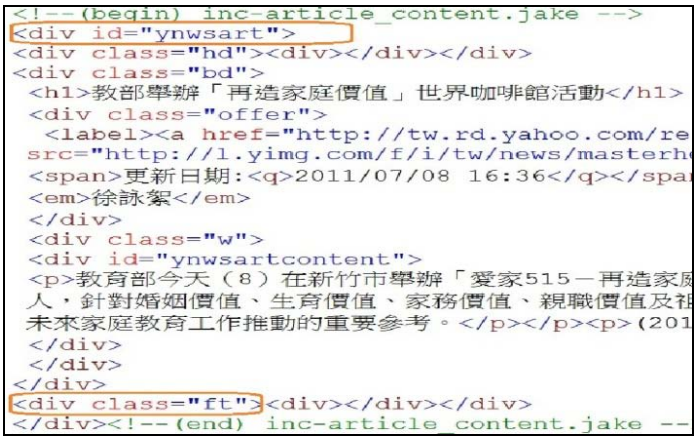


圖 4 非結構性資訊網頁之 HTML

(2)結構性資訊解析

本研究以文建會之藝文資訊網頁為例，說明如何解析具備結構性的藝文描述網頁內容。首先，從圖 5 可以看到各個藝文資訊皆被該網站先行使用標籤進行資訊分類。左邊的標籤名稱則是宣告右邊內容所屬的屬性。左邊的「活動名稱」便是其標籤名稱，而右邊的內容「上屋下屋過家聊，新竹縣地方文化館交流策展活動」，便是對應到左側的標籤名稱。接著，從圖 6 可以看到各個標籤皆被<tr valign="top" bgcolor="#eeeef3">及</tr>包住，然而其中的標籤名稱和標籤內容又各自被<td>與</td>分隔開來，因此，本研究便依此特徵進行標籤式藝文描述類型藝文資訊抓取。

圖 5 結構性資訊原始網頁

圖 6 結構性資訊網頁之 HTML

3.4 語意規則建置與對應

本研究採用中央研究院開發之系統(CKIP)做為斷詞的依據，然而其回傳的結果無法完全適用於特定領域。因此，本研究進行藝文活動內容分析時，所有藝文活動內容都會先經過領域詞彙修補程序，再結合本研究自訂詞類標記(表二)進行規則建置，以分析內容之資訊。

表二 本研究自訂詞類標記

詞類標記	相關詞彙	說明
Year	1999 年、2000 年...	年份相關詞彙
Month	一月、二月…	月份相關詞彙
Day	1 日、二日、廿號、卅號、今天、今日、即日…	月份相關詞彙
Time	早上、下午、晚上…	時間相關詞彙
ArtS	弦樂團、交響樂團、歌仔戲劇團…	表演者相關詞彙
Show	攝影展、舞台劇、獨奏會、音樂會…	表演類型相關詞彙
TiktType	索票入場、自由入場、購票…	票卷類別相關詞彙
Week	周一、週二、禮拜三、星期日…	星期相關詞彙
Holiday	國定假日、民俗假日、民俗節日…	假日相關詞彙

舉例而言， 一篇藝文事件之內容為「長榮交響樂團將於六月十一日於中正紀念堂演出，免費索票入場」，將這段內容傳送至 CKIP 中文斷詞系統處理後，輸出的斷詞結果內容如圖 7 所示。而 CKIP 系統處理結果會將「長榮交響樂團」及「索票入場」等特定領域詞彙斷成「長榮[N] 交響樂團[N]」及「索[Vt] 票[N] 入場[Vi]」，將使本研究之代理人程式不易得知這段內容的真實意涵，因此本研究會針對 CKIP 系統所輸出的斷詞結果內容進行領域詞彙之修補，同時亦會將 CKIP 原先斷詞之詞性，轉換修改為本研究所訂定的詞類標記，如「 交響樂團 (ArtS)」及「索票入場 (TiktType) 」，經過領域詞彙修補模組修正後，處理結果如圖 8 所示。

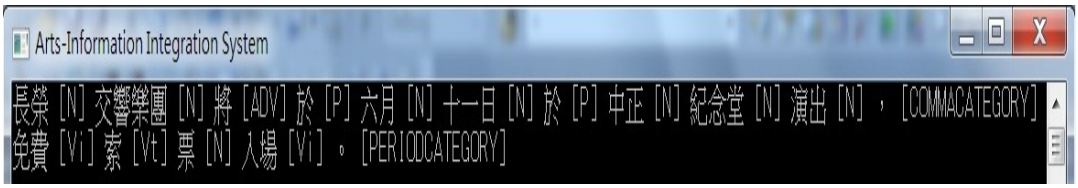


圖 7 CKIP 回傳結果

圖 8 領域詞彙修補後之斷詞結果

本研究參閱藝文事件中常見的敘述結構並依據自行建置的藝文資訊本體架構來設計語法規則，由於本研究之研究目的是解析非結構性的藝文資訊，以使 AIIS 中能包含一般藝文活動資訊網站缺乏或遺漏的藝文資訊，進而增加 AIIS 之資訊完整度與豐富度。為了降低非結構性資訊頗析的複雜度，本研究提出的雛形系統，僅先針對藝文資訊中活動名稱、活動時間以及活動地點等三個重要元素概念進行資訊的剖析與擷取。

限於篇幅，以下僅針對活動名稱之文法結構進行規則的解釋說明。「活動名稱」結構規則：此類規則用以辨識藝文活動的名稱，活動名稱常見的敘述結構範例如表三所示。例如有一則藝文活動的內容為『將在中正紀念堂演出「薛丁山傳奇」』，因為內容結構符合其中編號 1 之規則『演出[Vt] + 「[...] + …[...] +」[...] 』規則，所以該內容將解析出節目名稱為薛丁山傳奇。(註：詞彙規則的中括號代表限制之詞彙，...代表不限制詞彙; 詞性規則的中括號代表限制詞性，...代表不限制詞性)。

表三 活動名稱結構規則

規則	活動名稱結構規則內容
1	演出[...] +「[...] + ...[...] + 」[...]
2	「[...] + ...[...] + 」[...] + 為 + 主題
3	「[...] + ...[...] + 」[...]
⋮	⋮

3.5 語意網建置

中文詞彙的辨別皆以字意為主，而非字形。若兩組字詞意義相同，則視為相同詞彙; 反之若意義不同，則視為相異。由於藝文資訊大多採用同義字來敘述藝文活動的內容，因此本研究在語意關係中主要在辨別 Huang(2003)所提九種詞義關係中的同義關係(Synonymy)。

假設藝文之敘述內容與語意規則意思相同，但所用之詞彙不同如表四所示。若依照語法規則，顯然無法從敘述內容中分析出任何資訊。但建置語意網後，本研究代理人程式若發現詞彙不符合便會自行進入資料庫搜尋其它的同義字，以判斷「於」是否為「在」的同義字、「演出」是否和「展出」是同義字，若相同則視同符合規則。

表四 語意規則範例

規則	... [Month]	... [Day]	在 [...]	... [...]	展出 [...]
敘述 內容	5 月 [Month]	31 日 [Day]	於 [P]	美術館	演出[Vt]
同義詞			{在、 於...}		{演出、 展出...}

3.6 資訊衝突多目標訊息可信函數

由於藝文活動眾多資訊中，除了活動日期、活動時間、活動地點、活動地址等四項資訊產生衝突會造成參與民眾的困擾外，其餘資訊若重覆或衝突對民眾來說影響並不大，因此本研究僅對上述四項屬性進行資訊衝突的決策。然而，為了防止劣幣驅逐良幣的情形產生而提出資訊衝突演算法，公式(1)為本研究設計之資訊衝突多目標訊息可信函數 ：

$$D_{(i)} = \frac{W_p * P_{(i)} + W_t * T_{(i)} + W_c * C_{(i)}}{(W_p + W_t + W_c)} * 100\% \quad (1)$$

i ： 表示衝突群組數, i =1, 2, 3, …, n, 其中 n 為群組總數.

$P_{(i)}$ ： 表示第 i 個衝突群組的公信力指標

$T_{(i)}$ ： 表示第 i 個衝突群組的時間接近指標

$C_{(i)}$ ： 表示第 i 個衝突群組的訊息一致指標

W_p、W_t、W_c為上述三項指標之加權值

$D_{(i)}$:表示第 i 個衝突群組的可信分數

- 公信力指標

$$P_{(i)} = \max_{x \in \{1,2,\ldots,j\}} PW_{ix} \quad (2)$$

j ： 表示第 i 個衝突群組內的資訊來源數

PW_{ix} ： 表示第 i 衝突群組內, 有 j 個資訊來源，其中第 x 個來源的網站公信力

- 時間接近指標

$$T_{(i)} = \max_{x \in \{1,2,\ldots,j\}} \frac{SD-(TD-PD_{ix})}{SD} \quad (3)$$

j ： 表示第 i 個衝突群組內的資訊來源數

SD :表示接近天數標準,例如管理者認為 7 天前的資訊不予採信，則此處 SD 即為 7

TD ： 代理人執行解析時的系統日期

PD_{ix} ： 表示第 i 個衝突群組內， 有 j 個資訊來源，其第 x 個資訊來源的發佈日期

- 訊息一致指標

$$C_{(i)} = \frac{j}{CT} \tag{4}$$

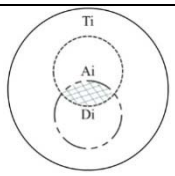
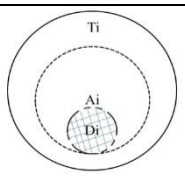
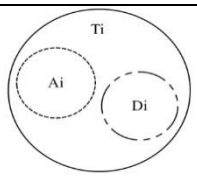
j ： 表示第 i 個衝突群組內的資訊來源數

CT ： 表示在 n 組的藝文衝突群組中，所有的藝文資訊數

3.7 藝文不一致資訊之整合步驟

本研究之代理人程式在進行所有不一致資訊類型資訊的整合之前，會先確認其交集之資訊元素是否完全相同，若有資訊衝突，則會先進行資訊衝突決策，然而才執行資訊不對稱之整合。針對各種不一致資訊類型之內容整合步驟，說明如表五。其中 Ti 表示資料庫第 i 筆藝文資訊在真實世界中的完整資訊,Di 為資料庫第 i 筆藝文資訊的實際儲存資料，Ai 為代理人程式正在解析的藝文事件，其符合資料庫中第 i 筆藝文資訊。

表五 不一致資訊類型之整合及解說

不一致資訊類型	整合步驟及說明
 聯集型-不一致資訊	(1) 確定 Di 與 Ai 為相同的藝文事件，並且 Di 與 Ai 擁有相同的資訊 Di ∩ Ai。 (2) 接著本研究開始處理 Di 與 Ai 的差集元素之整合。也就是將 Ai－Di 之元素加至 Di 中。
 子集型-資訊不對稱	(1) 確定 Di 與 Ai 為相同的藝文事件，並且 Di 為 Ai 之子集,Di ⊂ Ai。 (2) 開始處理 Ai 與 Di 的差集，也就是將 Ai 中所有屬於 Di 的元素去掉(Ai － Di)所形成的資訊集合加以整合。
 互斥型-資訊不對稱	(1) 確定 Di 與 Ai 為相同的藝文事件，並且 Di 為 Ai 為互斥 Di ∩ Ai = Ø (2) 本研究接著將 Ai 所解析出之元素，加至 Di 中。

3.8 AIIS 系統架構設計

為了使 AIIS 系統能夠透過語意分析的技術，讓電腦瞭解網頁內容中所蘊含的藝文資訊，本研究提出一個語意分析藝文整合系統之雛形系統架構(圖 9)。系統架構包含本

研究自行設計之藝文資訊 Agent 以及相關之處理模組與其資料庫。模組包括了 HTML 處理模組、領域斷詞修補模組、語意分析模組及最後的資訊衝突及整合處理模組，資料庫則有藝文來源網站資料庫、斷詞內容暫存資料庫、領域詞彙資料庫、斷詞內容暫存資料庫、語意規則資料庫以及藝文事件資料庫等。管理者可執行藝文來源網站的管理、領域詞彙管理、語意規則管理如藝文事件的管理。

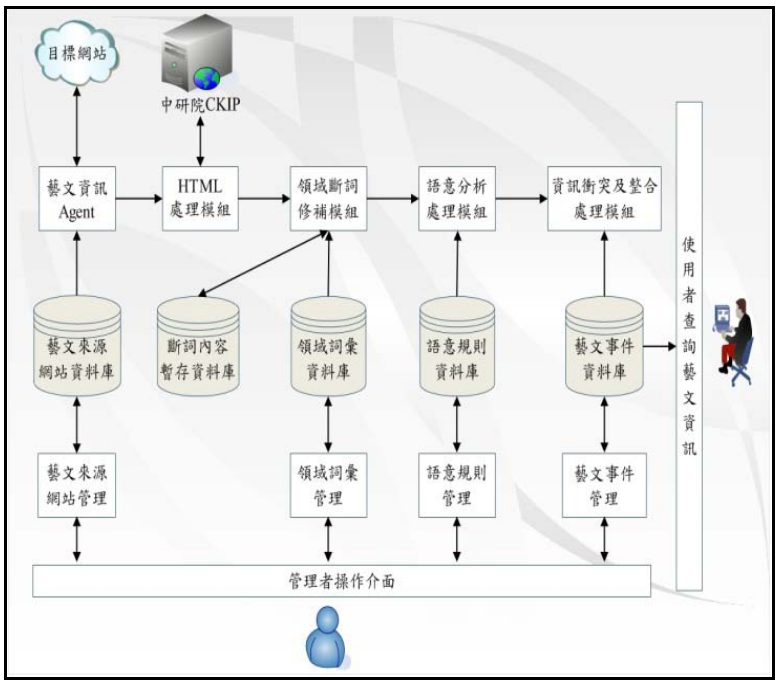


圖 9 AIIS 系統架構

4. 系統實作與展示

4.1 第一階段成效驗證

(1)結構性藝文內容之資訊擷取成效驗證

本研究為了瞭解 AIIS 在結構性藝文檢索網站之資訊擷取成效為何，故從 AIIS 由 2011 年 6 月 1 至 7 月 15 號為止，於各來源網站所搜集之 170 則結構式藝文資訊，並從中隨機抽取 15 則藝文事件進行擷取成效驗證。在驗證上我們使用資訊檢索領域中常用的精確率(P)，召回率(R)及 F 度量(F)等三個指標來進行評估。其公式如下所示，其中 Ac 表示擷取正確資訊數，Cc 表示真實資訊個數，Ag 表示擷取資訊個數 ：

$$R = \frac{Ac}{Cc} \tag{5}$$

$$P = \frac{Ac}{Ag} \tag{6}$$

$$F = \frac{2 * P * R}{P + R} \tag{7}$$

圖 10 的結果顯示，AIIS 在結構式藝文資訊網站的資訊擷取召回率、精確率及 F-measure 等三個成效指標分別為 98%、100%及 0.99%。未達 100%的原因是由於其中

有幾筆資藝文資訊之作者在進行藝文事件之撰寫時，本身內容就有誤，因此導致本研究在擷取率上未達至 100%。

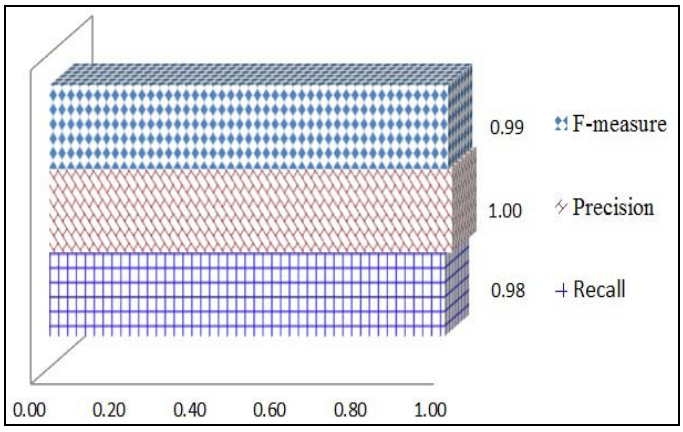


圖 10 結構式藝文敘述之資訊擷取統計表

(2)不同網站內容格式整合

民眾在進行藝文查詢時，除了遇到不同網站對於相同資訊稱呼上的差異外，還可能需要適應資訊表達方式的不一致性，例如不同的作者在時間格式，有人寫早上 9:00~18:00（圖 11），也有人寫早上 9:00~下午 6:00，甚至於用 AM 及 PM 的方式來描述時間;在電話的格式上，作者有些會加區域號碼有些卻不會，在分機的呈現，有些使用*字號，有些用-來表示，使得民眾在瀏覽各網站的藝文事件時容易造成不舒服的感覺，顯示出藝文資訊整合的迫切性及重要性。因此，本研究在取得各網站之藝文資訊後，會針對原網站所使用的格式轉換為統一的格式， 例如時間格式，AIIS 即會將之轉為 24 小時制後，再將之存入資料庫，以提供一個統一格式的藝文資訊介面及環境給民眾，如圖 12 所示。

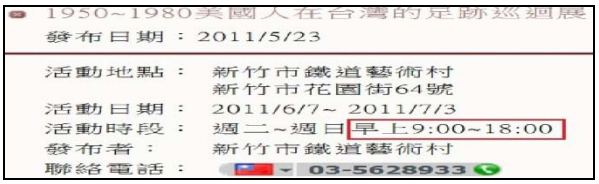


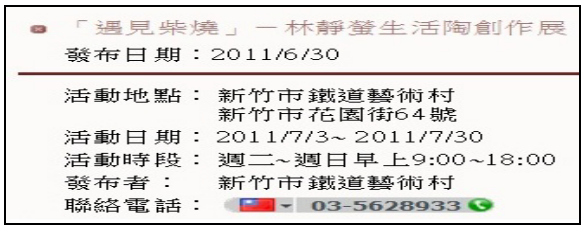
圖 11 藝文資訊時間格式

藝文名稱	1950~1980美國人在台灣的足跡巡迴展
活動日期	2011-06-07 至 2011-07-03
開放參觀	週二到週日
活動時間	9:00 到 18:00 (順位1)
活動時間	9:00 (順位2)
活動縣市	新竹市
活動地點	新竹市鐵道藝術村
活動地址	新竹市花園街64號
聯絡電話	03-5628933

圖 12 AIIS 藝文資訊呈現樣式

(3)藝文內容的資訊整合及資訊衝突實例展示

資訊整合成功與否所遭遇的最大挑戰，無異是資訊衝突及不一致資訊的判斷正確與否。因此本研究提供了資訊衝突多目標訊息可信函數及不一致資訊的整合方式。圖 13 及圖 14 便是 AIIS 在進行整合時遇到資訊衝突的範例。本研究代理人在經過藝文相似性判斷後，發覺這兩則藝文資訊為相同事件，便依流程進行資訊衝突的決策及資訊整合。AIIS 將兩則不同來源的藝文資訊整合後的藝文內容如圖 15 所示，其中日期、地點及地址各有兩個欄位，這是因為兩則藝文資訊各別提供的資訊產生衝突的原故，然而經由本研究可信函數處理後，將可信分數較高的資訊放至順位 1，次之則為順位 2，依序放置下去。該則藝文僅擁有兩則不同來源之藝文，因此僅至順位 2。在圖 15 亦可看到本研究除了解決兩來源資訊的資訊衝突外，更將兩則藝文之差集資料進行整合，證明本研究所提之資訊衝突目標訊息可信函數及整合策略為有效且可行。


「遇見柴燒」－林靜螢生活陶創作展
發布日期：2011/6/30

活動地點：新竹市鐵道藝術村
新竹市花園街64號
活動日期：2011/7/3~2011/7/30
活動時段：週二~週日早上9:00~18:00
發布者：新竹市鐵道藝術村
聯絡電話：03-5628933

圖 13 相同藝文事件之來源 A

活動名稱：	「遇見柴燒」－林靜螢生活陶創作展
所在縣市：	新竹市
活動型態：	展覽
活動類別：	工藝－陶瓷－不分細類
活動展演者：	(中華民國)林靜螢 林靜螢/中華民國
活動時間：	2011/07/07 09:00 ~ 2011/07/30 18:00
詳細日期時間：	
活動場地：	新竹市鐵道藝術村 展覽場
場地地址：	新竹市東區公園里花園街64號
票價：	無售票資訊
參與單位：	主辦：新竹市文化局 承辦：皓基國際實業有限公司
活動網址：	http://www.hcccb.gov.tw/
簡介：	將陶藝與生活情境充分融合，創造出生活週遭敏銳的感觸，林靜螢將生活創作出一件件賦予生命力的創作品，的學員，共同創作及尋找生活中感動創作品約10件，件件精彩，歡迎有興美成果。

圖 14 相同藝文事件之來源 B

藝文名稱	「遇見柴燒」－林靜螢生活陶創作展
活動日期	2011-07-03 至 2011-07-30 (展位1)
活動日期	2011-07-07 至 2011-07-30 (展位2)
開放參觀	週二到週日
活動時間	9:00 到 18:00
活動縣市	新竹市
活動地點	新竹市鐵道藝術村 (展位1)
活動地點	新竹市鐵道藝術村展覽場 (展位2)
活動地址	新竹市北園街64號 (展位1)
活動地址	新竹市東區公園里北園街64號 (展位2)
活動類別	展覽
活動演出者	林靜螢
主辦單位	新竹市文化局
承辦單位	皓基國際實業有限公司
聯絡電話	03-5628933
活動介紹	將陶藝與生活情境充分融合，創造出自我獨特的生活，這是林靜螢接觸陶藝近20年來創作的理念，會接觸陶藝純屬興趣，藉由對生活過遭敏銳的感觸，林靜螢將生活與情感充份融合於陶藝中，運用手捏、泥條、陶板、拉坯...等方式，創作出一件件賦予生命力的創作品，近年來林靜螢更將理念延伸，帶領著多位情緒障礙及對陶藝充滿熱忱的學員，共同創作及尋找生活中感動的力量。本檔展覽除展出林靜螢個人創作品約25件外，亦包含其學員創作品約10件，件件精彩，歡迎有興趣民眾前往欣賞，共同感受林靜螢創作生活陶記真的過程及燒成的甜美展出同時分別於7月23日(六)下午3:00及7月30日(六)下午3:00免費辦理生活陶藝DIY活動，每檔次限額20名，需預約報名，報名專線:03-5628933。 目前簡介來源為：「鐵道藝術村」，內容豐富度為100% 其它參考： 「文建會藝文活動查詢系統」，內容豐富度為73.54%
活動連結	連結1 連結2
活動圖片	

圖 15 AIIS 整合後之呈現

4.2 第二階段成效驗證

(1)資訊豐富度驗證

本研究建立一個驗證的假設情境，若一個使用者欲查詢 2011 年 7 月 16 日至 7 月 31 日期間在新竹縣、市展出的藝文活動，他分別前往新竹縣政府文化局、新竹市政府文化局、文建會藝文活動資訊系統及本研究之 AIIS 等藝文資訊網站進行藝文活動的查詢。

其情境的藝文查詢結果如圖 16 所示，文建會藝文活動資訊系統一直以來都是一般民眾在搜尋藝文資訊的指標性網站，所擁有的藝文數也是各網站最多的，而本研究所建置之 AIIS 藝文資訊整合系統在搜集各網站之藝文資訊後，所提供的藝文資訊豐富度不僅遠高於其它網站，更超越文建會的 65 筆藝文資訊，總筆數為 76 筆。結果顯示 AIIS 在第二階段的成效上，確實有效提高資訊豐富度。

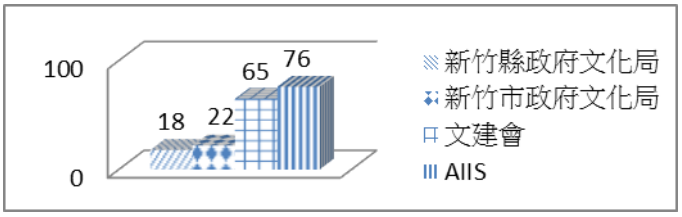


圖 16 驗證情境下的藝文活動數量

(2)AIIS 資訊覆蓋率分析

在經過資訊豐富度的驗證後，本研究繼續深入了解 AIIS 對於各來源網站之覆蓋率為何，此處所謂的資訊覆蓋率為本研究所設計之 AIIS 與所抓取藝文來源資訊重疊率，值越高則表示藝文資訊包含率越高。表六可以看到本研究在各來源的覆蓋率皆未達 100%，這是由於.NET 環境及技術上的限制，使資訊代理人常於下載網頁原始碼時遇到轉碼錯誤(亂碼)的問題，而為了避免網頁內容呈現出亂碼狀況，因此本研究代理人若遇到亂碼，則會對該則藝文資訊進行刪除，導致 AIIS 未能於各來源取得全部的藝文資訊，若扣除上述本研究所不能控制因素而刪除之藝文資訊，AIIS 對各網站覆蓋率可達 100%。此外，本研究 AIIS 所包含的資訊量也遠高於各網站，從圖 17 可以看到，AIIS 所擁有的資訊數相較於縣政府所提供的藝文資訊，高達 4.22 倍以上，而在文建會方面亦有 1.17 倍之水準，證明本研究 AIIS 不僅覆蓋率可達 100%，且所提供的資訊量也最豐富。

表六 本研究 AIIS 對各來源網站之資訊覆蓋率

	新竹市政府	新竹縣政府	文建會
AIIS 覆蓋率%	91	94	92

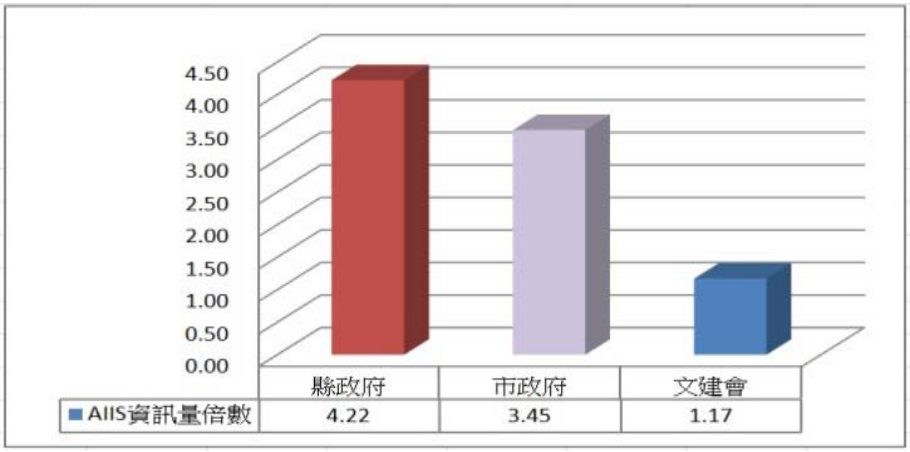


圖 17 AIIS 對各來源網站之資訊量比較

(3)非結構性資訊的擷取

一直以來，很少有藝文資訊檢索網站對非結構性的資料內容進行解析，因此本研究嘗試使用語意分析技術來針對即時性的非結構網路新聞進行解析，以提供具備即時性的藝文相關資訊給民眾。然而本研究主要目的是透過解析即時性的藝文新聞內容來增加藝文豐富度或提供與藝文事件相關的資訊，故在解析非結構的藝文內容時，僅擷取藝文名稱、活動地點以及活動時間等三個屬性以供民眾參考。

純文字敘述屬於非結構性文章的關係，在使用語法上較無約束，作者容易使用不同

的句法敘述來詮釋一段甚至於一整則藝文敘述，因此本研究在擷取非結構性的藝文資訊上，遇到許多困難。雖是如此，但 AIIS 仍於數百篇新聞稿中，成功的從 13 篇非結構性的藝文事件新聞資訊，成功擷取出 12 則藝文資訊。其擷取內容如圖 18 所示，圖 19 為其原始的藝文網頁內容。由此可知，本研究的整合系統不僅可提供一致性的呈現格式，更可以從非結構的資訊中擷取中藝文的重要資訊存入資料庫中，使 AIIS 具備即時性藝文資訊的提供。

藝文名稱	蕭如松2弟子 辦聯展憶師恩
活動日期	2011-07-03 至 2011-07-25
活動地點	蕭如松藝術園區
活動介紹	〔自由時報記者黃美珠／竹東報導〕新竹縣知名已故畫家蕭如松，其學生賴世明和范弘志受他啟蒙，數十年畫筆不輟，賴世明甚至在小學當志工教畫。兩人即日起到25日，在蕭如松藝術園區舉辦聯展，感念恩師的提攜，也和後進分享畫畫之樂。七十歲的賴世明和六十六歲的范弘志，彼此是竹東高中的學長和學弟，進入高中前從來不曾畫過畫，直到遇見了蕭如松老師，才開啟了他們藝術的生命。賴世明的作品，早年曾入選全省美展、台北市美展等競賽，表現不俗，後來進入當年的國立藝專深造，踏出社會後也一度以教畫為生，退休後改當志工。而范弘志高中畢業後雖投入軍旅，但仍難忘提筆作畫，因為畫畫讓他在軍人一板一眼的作息中，有個自由揮灑的喘息空間。賴世明說，恩師最愛寫生，透過寫生，他感到很放鬆，因此多年來，也始終喜愛用水彩畫景。這次和學弟范弘志聯展，端出的廿二幅全都是水彩畫的寫生。其中他最愛的就是一幅就是在馬祖的畫作。至於范弘志雖愛寫生，但也畫靜物，昨天展出的作品中，一幅「紫鉤花」，他認為如果恩師在場，一定也會喜歡。如果時光倒流，重回課堂上，定能獲得好成績。 目前簡介採自「雅虎藝文新聞-藝術展覽」
來源網站	雅虎藝文新聞-藝術展覽, 內容豐富度為100%
活動圖片	

圖 18 AIIS 擷取非結構藝文內容

圖 19 非結構性藝文原始網頁內容

5.結論與未來建議

本研究的第一階段驗證結果顯示，在結構性的藝文資訊擷取率上，其召回率、精確率、F 度量皆達 100%之水準，在藝文資訊的格式統一及資訊不對稱的整合上亦有相當的成果。在第二階段的驗證資料顯示，在資訊豐富度上，AIIS 所提供的資訊數量高於其它藝文資訊檢索網站，證實本研究之 AIIS 系統擁有高水準的資訊豐富度;在資訊覆蓋率之探討結果顯示 AIIS 達至 90%以上之涵蓋率，在使用語意分析對非結構性藝文資訊的解析

上，因為藝文領域的範圍太廣，所涉及的語言、語法過於複雜，使得本研究在資訊的擷取上遇到較大的困難，雖然 AIIS 未能將所有非結構的藝文資訊成功擷取，但仍可以將部份即時性的非結構性藝文以提供給民眾參考。

綜合上述成果，本研究不僅對結構性藝文資訊有優秀的擷取成效，還擁有格式一致性、高資訊豐富度及提供即時性藝文資訊能力等優點，更具備資訊整合及衝突決策的能力。因此，顯示出本研究所提出的二階段藝文整合技術確實可應用於藝文資訊領域，以協助進行藝文資訊的整合。

本研究在非結構性的藝文資訊擷取上，僅對藝文名稱、展出日期及展出地點進行解析，後續研究者可自行建置其餘藝文資訊本體架構中的其它屬性進行資訊的擷取分析。

參考文獻

1. 許菱祥，1986「中文文法」，大中國圖書公司。
2. Agrawal, R. and Srikant, R. “Mining Sequential Patterns” Proceedings of IEEE International Conference on Data Engineering 1955, pp.3-14.
3. Alberts, L.K. “YMIR: An Ontology for Engineering Design” Ph.D., Thesis, University of Twente 1993.
4. Berners Lee, T. “Weaving the Web: The Original Design and Ulitmate Desitiny of the World Wide Web” London: Verso 1999.
5. Borst, P., Akkermans H., and Top J. “Engineering Ontologies, International” Journal of Human-Computer Studies (46)1997, pp.365-406.
6. Bunge, M. “Ontology I: The Furniture of the World. Treaties on Basic Philosophy” Boston, MA: Reidel (3)
7. Caverlee, J., Liu, L., and Buttler, D. “Probe, Cluster, and Discover: Focused Extraction of QA-Pagelets from the Deep Web” Proceedings of IEEE International Conference on Data Engineering 2004, pp.103-114.
8. Ceri, S., Klemettinen, M., Lanzi, P., & Milano, P. “A Tool for Extracting XML Association Rules from XML Documents” Proceedings of the International Conference on Tools with Artificial Intelligence 2002.
9. Chen, M.S., Han, J., & Yu, P.S. “Data Mining: An Overview from a Database Perspective” IEEE Transactions on Knowledge and Data Engineering 1996, p866-883.
10. Chen, J.S., Hsieh C.L., & Hsu, F.C. “A Study on Chinese Word Segmentation: Genetic Algorithms Approach(以遺傳演算法為基礎的中文斷詞研究)” Journal of E-business, 2(2) 2000, pp. 27-44.
11. Cooley, R., Mobasher, B., Srivastava, J. “Web Mining：Information and Pattern Discovery on the World Wide Web” Proceedings of Ninth IEEE International Conference of Tools with Artificial Intelligence 1997, pp.558-567.
12. Council for Cultural Affairs, R.O.C, Retrieved March 29, 2010, from the World Wide Web: http://www.cca.gov.tw/public.do?method=list&categoryId=3

13. Directorate General of Budget, Accounting and Statistics, Executive Yuan, R.O.C,Retrieved March 29,2010,from the World Wide Web: http://www.dgbas.gov.tw/ct.asp?xItem=27725&ctNode=3263
14. Erik, T. R. “Learning Xml”, Taiwan:Oreilly 2001.
15. Gruber, T. R. “A Translation Approach to Portable Ontology Specifications” Knowledge Acquisition(5:2) 1993, pp.199-200.
16. Guarino, N. “Formal Ontology and Information System” in Formal Ontology in Information Systems, Proc. Of the 1st International Conference, edited by Guarino, N.,Trentom Italy, IOS Press (6:8) 1998, pp.3-15.
17. Hebel, D. “Data Mining and Knowledge Discovery Gaining Competitive Advantage with Example in the Area of Finance” 1998, Retrieved January 25, 2010, from the World Wide Web: ttp://andrew.cmu.edu/user/dhebel/telecom/paper/DataMiningPaper.html
18. Huang, C.R. “Semantic web, WordNet and Ontology: A talk on knowledge management on future’s web (語意網、詞網與知識本體：淺談未來網路上的知識運籌)” Information Management for Buddhist Libraries(33) 2003, pp.6-21.
19. Kalakota R., & Whinston A. B. “Electronic Commerce: A Manager Guide” Addison-Wesley 1997.
20. Kosala, R., Blockeel, H. “Web Mining Research: A Survey” SIGKDD Explorations (2) 2000, pp.1-15.
21. Loiacono, E. T. “Web Quality: A Web Site Quality Instrument, Doctor Dissertation”, Georgia: The university of Georgia 2000, pp.88-147.
22. Neches, R. Fikes, R. Finin, T.Gruber, T., Patil, R., Senator, T. Swartout, W. R. “Enabling Technology for Knowledge Sharing” AI Magazine (12:3) 1991, pp.36-56.
23. Nie, Z. & Kambhampati, S. “A Frequency-based Approach for Mining Coverage Statistics in Data Integration” Proceedings of International Conference on Data Engineering 2004, pp.387-398.
24. Smith, B. & Welty, C. “Ontology: Toward a New Synthesis” Proceedings of the international conference on Formal Ontology in Information Systems, Ogunquit, Maine, USA 2001.
25. Studer, R., Benjamins, V. R., Fensel, D. “Knowledge Engineering: Principles and Methods” Data and knowledge engineering (25) 1998, pp.161-197.
26. Swarout, B., Ramesh, P., Knight, K. & Russ, T. “Toward Distributed Use of Large-scale Ontology” In Farquhar, A., Gruninger, M., Gomez-Perez, A., Uschool, M. and Vet P, V. D. (Eds.) AAAI'97 Spring Symposium on Ontological Engineering 1997, pp.138-148.
27. Uschold, M. & Gruninger, M. “Ontologies: Principles, Methods and Applications” , The Knowledge Engineering Review (11:2) 1996, pp.93-136.
28. Van Heijst, G., Schreiber, A. T., & Wielinga, B. J. “Using explicit Ontologies in KBS development” International Journal of Human-Computer Studies(46) 1997, pp.183-292.

29. Wielinga, B. J., & Schreiber, A. Th. “Reusable and Shareable Knowledge Bases: A European Perspective” Proceedings of International Conference on Building and Sharing of Very Large-Scaled Knowledge Bases ’93, Tokyo, Japan 1993, pp.103-115.
30. Wixom, B. H., & Todd, P. A. “A Theoretical Integration of User Satisfaction and Technology Acceptance” Information Systems Research, (16:1) 2005, pp.85-102.

A Two-Phase Technique for Arts Information Integration Based on Ontology and Semantic Analysis

Ming-Hsiung Ying [1] Guang-Hong Deng[2]
[1] Department of Information Management, Chung Hua University
[2] Department of Information Management, Chung Hua University lalaethan@hotmail.com

Abstract

In recent years, the living standard of human is making a big progress, with its following increased demand for “Culture and Arts” which has strongly highlighted the importance of the information on “Culture and Arts” for those people who doing this activities. This study used ontology, web mining and Semantic analysis technique as a basic approach, and proposed a two-phase arts information integration system (AIIS). The first phase of AIIS system is that it automatically gathering the needed arts information and integrating them to the same format inside a framework by agent technique. The second phase of AIIS system is using semantic analysis technique to refine and extract the useful information from those non-structuration arts information taken from the first phase. The results indicate that our AIIS system could reach up to 100% in gathering available arts information. And the result from the second phase shows that our AIIS system not only provide abundant in art activity information more than other websites but also can provide the updating newest arts activity information for searching.

Keywords: Web Mining, Ontology, Semantic Analysis, Information Integration, Agent