基於雲端基礎之通用個人輪廓

A New Concept of Cloud-based Universal User Profile

李麗華	李富民	歐仁德	張仲毅
育達商業科技大學	朝陽科技大學資訊	澎湖縣政府警察局	朝陽科技大學資訊
資訊管理系教授	管理系副教授	資訊科技士	管理系研究生
lhli@ydu.edu.tw	fmlee@cyut.edu.tw	rendeou@gmail.com	zhungei@gmail.com

摘要

現今已有許多網路服務會利用挖掘客戶偏好來提供較佳的服務。也有許多網站建立使用者個人輪廓(User Profile)來分析使用者偏好並提升推薦準確度。不過這些個人輪廓資訊都被儲存於不同的網站，因此也面臨了個人資料之隱私權及管理一致性等問題。因而有學者提出客戶端之通用個人輪廓(Universal User Profile)，負責收集資訊並提供單一隱私權入口來達成保護效用。但因目前使用者設備多樣化，以往透過單一客戶端進行服務已不合適。

有鑑於此，本研究提出透過客戶端代理人進行個人資料之取用與 P3P 授權服務，再以雲端運算平台進行資料儲存與偏好分析運算。其中透過本體論、網路習性探勘等偏好分析方法建立雲端通用個人輪廓，最後則建立應用程式介面(API)供第三方網路廠商取用。最終建立一個較佳且具個人資料保護之雲端通用個人輪廓。

關鍵詞：通用個人輪廓、雲端、本體論、個人化。

Abstract

There are many internet services using mining technique for customer preference so that better service can be predicted. Many internet companies usually rely on user profile for analyzing user's web usage behavior and for better service. However, the privacy and the consistency of user data are not guaranteed in secure status. Therefore, some researchers suggested that a sel-protected universal user profile can resolve the above problems. However, as the device of web usage moving from PC to mobile devices, the traditional universal user profile is, hence, not enough for user.

This study intends to propose a new concept of cloud-based universal user profile. The user profile is stored in cloud storage but maintained by user itself. A user-profile agent is designed for handling the accessing request of personal information from internet merchant. The P3P authorizing service is conducted and the session identification is designed to protect the information. Therefore, the internet merchant can access the user information and analyzing user preference without maintaining the user information, because the user profile is maintained by the user itself. In this way, a user profile can be secured and the information can be shared based on user's permission.

Keywords: Universal user profile, Cloud Storage, Ontology, Personalization.

1. 緒論

1.1 研究背景

在這個資訊發達的時代，大多數網路使用者對於線上資訊幾乎都有非常高的需求，也因此造成了所謂的資訊過載(Information Overload)[1]。尤其目前無線行動設備的發展相當蓬勃，部分使用者除了個人電腦設備(PC)外，甚至可能會有諸如智慧型手機(Smart Phone)、平板電腦(Tablet)等行動裝置。因此如何在如此大量的資訊檢索過程中，進行資訊擷取(Information Retrieval)與資訊過濾(Information Filtering)等動作，以協助使用者在較短的時間內，特別是行動裝置的即時性上，找出較符合使用者需求的資料是一個值得討

論的議題。而資訊過濾中的個人化(Personalization)應用即是其中一個解決方法。

個人化是在系統與使用者的互動中所取得的使用者相關資料，如特定偏好或需求等等。再透過這些個人資料進行相關應用以符合個人化需求[22]。同時這些資料亦稱作使用者之個人輪廓(Profile)。使用者輪廓可有效地描述使用者之整體偏好與需求[6]。而如何正確並完整的描述使用者輪廓則是資訊系統的優劣評比標準。

1.2 研究動機

一般企業進行個人化的服務時常會遇到冷起始(Cold Star)問題。冷起始即是系統在首次面對使用者的情況下，因事前無任何學習或紀錄等行為，所造成的毫無任何資訊可供進行個人化服務應用的窘境。

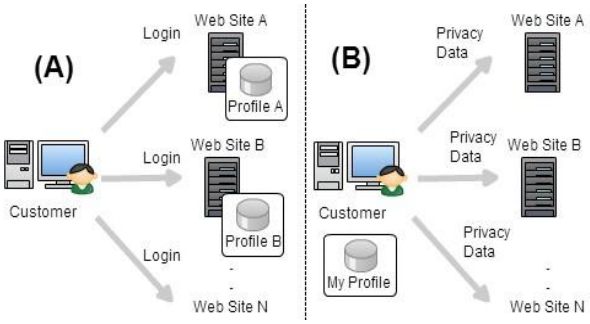

圖 1、(a)傳統個人輪廓 (b)通用個人輪廓

其次，傳統的網路服務提供者多依個別需求進行使用者輪廓的收集與儲存動作，即如圖1(a)所示。使用者無法掌握其個人資料是否有被妥善保存及利用，也就是會面對個人資料散落在各服務提供者間的「隱私權」及「風險」問題。

接著，同一個使用者在每一個網路服務間，個人資料仍需個別進行管理。故其資料會遇到「一致性」問題。同時以企業的角度而言，也需要面對使用者輪廓之儲存、保管、分析等營業維護成本。

如圖 1(b)所示，雖然有部分網路提供者[5][14][21]採取策略聯盟之方式，將聯盟內之個人輪廓檔進行整合並彼此分享，藉此解決上述問題。但此方法只能適用特定領域，且還是無法提供給聯盟外的服務進行取用。

此外，亦有以客戶端為主之通用個人輪廓(Universal User Profile)應用。此應用主要由客戶端之代理人程式(Agent)進行資料收集與分析，再透過如 P3P(Platform for Privacy Preferences)的隱私偏好平台授權後，最終提供給網路服務提供者進行資料取用動作。此種應用擁有較佳之隱私授權機制，對服務提供者而言亦可節省資料儲存管理及分析等成本。

但由於目前使用者客戶端裝置漸趨多樣化，不同裝置平台間之個人資料收集與維護亦為待解決的問題。且位於客戶端的個人輪廓檔案仍有領域概念(Domain Concept)問題。即個人輪廓需藉由專家對特定領域背景與專有名詞等知識進行定義後，通用個人輪廓才具有判斷使用者行為之依據。

1.3 研究目的

本研究旨在提供一個較佳的線上通用個人輪廓解決方案。以雲端服務的方式，將使用者個人資料進行線上上傳、雲端儲存、偏好分析運算、隱私授權等動作。

當雲端通用個人輪廓成形後，針對符合使用者需求之企業網路服務或產品，即可在有隱私授權的狀況下，進行網路服務最佳化或產品推薦服務。因此本研究之目的說明如下：

(1) 建立一個基於網際網路瀏覽行為之雲端通用個人輪廓。
(2) 透過雲端服務讓使用者端進行資料上傳、隱私設定；於雲端伺服器上進行儲存、偏好分析運算、隱私授權。
(3) 建立多平台上之軟體代理人(Agent)，作為客戶端資料擷取、上傳、管理隱私授權等動作。
(4) 建立分析方法，以驗證其通用個人輪廓是否與使用者偏好相符。

2. 文獻探討

2.1 個人化與個人輪廓

資訊時代之個人化，是企業進行商業行銷的重要功能。在以往尚具有網際網路時，個人化的定義，是藉由分析消費者需求及偏好來提供給消費者更好的服務所進行的互動行為[22]。而在網際網路盛行時期，個人化目標則為使網站能夠回應使用者之個別需求[17]。過去的網站個人化，主要透過統計大眾消費行為之特徵，但目前型態已漸漸變為一對一型式之服務。

個人輪廓(User Profile)主要為用來描述使用者特徵、外型等足以代表個人特性之資料。有學者指出所有資訊過濾系統主要目的均為準確的描繪出使用者，將其需求以個人輪廓型式表現[11]。其次，能否充分的表達使用者興趣則是系統的關鍵因素。而在此過程內，系統須先瞭解使用者偏好與需求。因此使用者勢必得先透露個人偏好與特徵，再交由個人化服務進行個人輪廓的建立。

在此透露的過程中，可分為顯性(Explicit)及隱性(Implicit)兩種收集模式。顯示模式是由使用者直接給予指示，例如直接透過偏好選擇或輸入特定資料等方式。但其缺點即為造成使用者之操作負擔，且可能因隱私顧慮而留下不真實之資料；而隱性模式則是透過系統收集使用者之點擊行為或瀏覽路徑等動作。因此對於使用者而言，其不需負擔額外的操作動作。但因收集成果需要花費額外分析整理等動作，故效果並不如顯示模式來得直接有效[12]。

因此本研究認為，隱性資料雖具有分析上的一定難度，但相較於顯性模式下對使用者所造成的額外負擔而言，隱性模式仍具有相當的優勢。惟兩者均還是屬於使用者隱私資料，故仍需注意到個人輪廓隱私權的授權。

2.2 個人輪廓隱私權(Privacy)

為達成個人化目的，服務提供者往往會要求使用者提供個人身分資料、消費習慣、網頁瀏覽習慣及興趣等；或是在閱讀並同意隱私權策略的狀況下進行隱性資料收集。系統即可透過這些資料進行購物型態或行為等分析。個人輪廓分析雖可降低企業之廣告成本，但缺點為若未透過使用者同意而進行紀錄描繪等動作，則可能變相成為干擾或監視行為，且資料亦可能遭到濫用或洩密。因此造成使用者可能不願在網路上揭露個人資訊，造成個人輪廓的代表性不足。因此企業得額外花費成本進行管理維護。

而以資料變化程度而言，有學者提出兩種分類[24]：(1)靜態私人資訊(Static Private Information)，如姓名、生日、家庭狀況等不會隨時間產生激烈變化之資訊；(2)動態私人資訊(Dynamic Private Information)，即會隨時間而有變化的資訊，如財務狀況、嗜好習慣等。而大部分使用者願意以部分個人隱私資訊與企業進行利益交換。但仍不願透露高度敏感資訊，如財務狀況等[13]。即為越能與使用者發生直接關聯的資料通常越不願意進行透露[10]。

而在透露之後，網站如何使用個人資料亦為使用者關切的重點。使用者輪廓除了幫助精確的描繪使用者外，其隱私權的保護與處理更為需要重視的特點。

2.3 通用個人輪廓系統

通用個人輪廓系統主要為提供一個多方共享之個人資料儲存空間，以提供個人化服務使用[2]。即不同領域的個人化服務，卻是使用同一組個人輪廓，且具備隱私權保護之機制，其好處為：(1)全面的了解使用者需求。因其整合了多種服務之輪廓，可從多角度對使用者進行瞭解。且於相同領域中，其輪廓可直接進行分享；在不同領域中可透過關聯等方式進行合作運用；(2)個人資料一致性。因個人資料統一管理，可減少分散各網站之管理成本及避免一致性問題發生；(3)減少冷啟始。透過共享之方式，可有效降低此問題發生；(4)隱私保護。以往資料為網站各自儲存，若有不當使用等情形將難以掌控。

透過通用個人輪廓，網路服務只能得到最終之抽象使用者偏好，而無法直接得到最原始之詳細資料，例如點閱行為(Click Stream)等。故可有效避免資料濫用，提升隱私權之保護。

2.3.1 隱私偏好平台(Platform for Privacy Preferences, P3P)

由全球資訊網路協會(World Wide Web Consortium, W3C)於1997年所提出之隱私偏好平台(Platform for Privacy Preferences, P3P)計劃[23]，較常使用之版本為P3P 1.0，2005已推出P3P 1.1版。

其目的為提供使用者更完善的個人資訊控制權，而非被動的接受個人隱私規則，並透過P3P機制與網站溝通協商如：使用者願意提供什麼資料、如何被運用等。且P3P提供一個網站將如何進行資料收集的標準化步驟，讓使用者擁有完全且主動的隱私控制權。

在P3P 1.0架構中，P3P Policy及P3P User Agent為兩個重要原件。

P3P Policy以XML語言儲存網站內各個網頁所使用的隱私資訊內容，如該網頁隱私權規

範、網站如何收集與如何被使用等。且依各個網頁之性質可分別給予不同的隱私資料開放度。此規範內容是參考儲存在公正第三方(W3C 或 Trust-e 等受信賴的隱私保護單位)的隱私政策參考檔(Policy Reference File)中。藉此提供標準與一致的隱私規範外，亦可確保使用 P3P 的網站會依規定進行隱私資料上的運用。

P3P User Agent 為提供使用者一個自動化的隱私資料代理程式，協助使用者設定願意公開的隱私資料，且自動化的與網站進行協商跟交換。當使用者瀏覽有支援 P3P 的網站時，它會將網站的 P3P Policy 與使用者自訂之偏好進行比對。若超過使用者定義的可接受範圍，使用者即會收到警示訊息。目前現行主流瀏覽器中，微軟 Internet Explorer 即有支援[20]。

因此透過 P3P 之隱私資料授權，其資料不僅可儲存於網站中，亦可將其個人輪廓儲存於個人電腦之中，讓使用者充分的掌握其隱私資料控制權。

2.3.2 通用個人輪廓系統模式

此類型系統依過去研究[2]可分為兩大類：(1)伺服器共享個人輪廓模式與(2)客戶端主導之通用個人輪廓模式。而在本研究中則再加入一類為：(3)第三方分享個人輪廓模式。

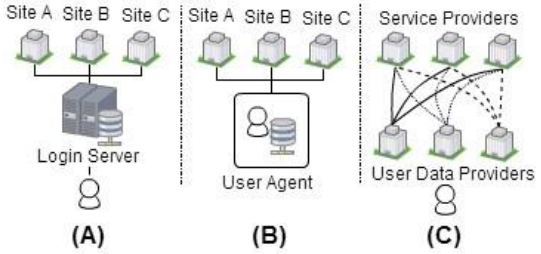

圖 2、(a)伺服器共享個人輪廓 (b)客戶端主導之通用個人輪廓 (c)第三方分享個人輪廓

(1)伺服器共享個人輪廓模式是由多個網路服務提供者間彼此組成聯盟，並分享使用者資訊及一致的登入入口。讓使用者可方便的利用其聯盟內的所有資源。知名聯盟案例為 Microsoft Passport[14]及 Liberty Alliance[21]。如圖 2(a)所示。

(2)客戶端主導之通用個人輪廓模式，此種模式下使用者能完全主導掌握其資料被利用的程度，多半會以 P3P 作為授權之標準。於客戶端同時可以管理統一的個人輪廓，且進行全面的資料收集並分析。有學者[2]曾提出適用於網際網路環境中的個人資料管理系統，並稱其為個人資料平台(Personal Data Backbone, PDB)。此系統在客戶端建立一個監控代理程式，並收集瀏覽網頁資料進行分類儲存。當有網路服務提供者需要用資料時即可產生當時之偏好資料。如圖 2(b)所示。

(3)第三方共享個人輪廓模式。此種類服務，將由一個具有較多使用者之主流網路服務提供者做為個人輪廓資料第三方提供者。此提供者具有一個應用程式介面(Application programming interface, API)，供其他服務商以特定隱私授權方式進行使用者之資料取用與提供服務。較為知名案例即為 Facebook 與 Google。如圖 2(c)所示。

本研究認為，雖伺服器共享與第三方共享個人輪廓等模式對使用者而言具有較高的方便性，但因其隱私資料仍散佈於各主流網路服務提供者，仍會遭遇一致性等問題。客戶端主導模式則亦受限於多平台下之管理維護問題。

2.4 本體論(Ontology)

客戶端的通用個人輪廓全面的收集使用者行為，其結果應能呈現使用者之個人偏好。但如何在客戶端資料內，找出適用於各個領域的分類知識是非常不容易的。因此本研究針對客戶端輪廓進行改良並結合本體論，再以本體論相關技術來解釋使用者偏好。

2.4.1 定義

本體論是一種可用來描述與表示各種領域內知識的技術。由 W3C 之 Web-Ontology 工作小組所規劃的標準化方法[25]。

在本體論內包含定義特定領域內的各種基本概念(Concept)，以及每一個概念間的關聯性(Relation)。其可用來表達某個專業領域內的術語、語意與資訊等等。同時其特色為機器可讀與可理解的。因此可系統可透過本體論之知識架構進行溝通與分享。本體論能將詞彙用階層式架構呈現，藉此描述方式即可得出領域知識之骨幹。有學者[18]認為本體論具有以下優點：(1)可重複使用的領域架構；(2)詳細與明確的描述領域中的假設；(3)領域知識與運算上的知識之間可有所分隔；(4)利於分析領域知識架構。

2.4.2 本體論與網站分類服務(Web Directory Service)

因通用個人輪廓之領域概念需由專家建立，才能解釋使用者行為[2]。因此能否取得一個合適的網際網路服務本體論以做為通用個人輪廓使用是相當重要的[16]。而網站分類服務即是相當適合套用的架構。

網站分類服務類似黃頁服務，為由入口網站透過專家，將網站依性質進行分類。曾有學者研究[13]指出網站分類服務為一個適用於網路環境的本體論。較為知名的網站分類服務為Yahoo! Category，以及另一個廣泛被採用的開放分類專案(Open Directory Project)[19]。

2.4.3 概念推論

過去有研究者[3][15]以特定領域知識結合使用者瀏覽紀錄後，建立個人化系統。可找出使用者之過去偏好，並可透過本體論階層架構，找出未瀏覽過的潛在偏好。

由 Middleton[15]等人提出之 Quickstep 論文推薦系統。此系統先以問卷收集來建立使用者之初始領域偏好，再透過系統推薦合適論文給予使用者。其中藉由將子類別權重分享給父類別的分享動作，可發掘使用者未曾看過的領域。由曾信誠[3]提出之技術，透過使用者查詢之關鍵字與專家定義之領域本體進行比對後，以經常查詢的關鍵字為基礎來建立使用者專屬的個人本體論。即可提供個人化的資料檢索成果。

2.5 網路習性探勘(Web Usage Mining)

除了透過通用個人輪廓與本體論進行分析外，仍需透過客戶端資料進行分析。也就是將瀏覽器紀錄(Web Logs)進行使用者行為識別後進行推薦應用。而在使用者進行瀏覽行為中，由於網頁層次目錄間的點擊將會造成瀏覽紀錄膨脹，因此需要透過探勘動作來減少無意義資料的存在。

2.5.1 瀏覽序列識別(Session Identification)

欲進行探勘動作，需先找出使用者網站上的瀏覽序列(Session)[8][9]。序列是指單一使用者為了特定目的而發生的連續瀏覽行為。以伺服器記錄檔為例，若欲找出特定使用者之瀏覽序列則需透過以下步驟：

(1) 資料清理(Data Cleaning)：清除對使用者無意義之資料，如圖檔(jpg, gif)、樣式檔(css)、程式檔(dll, js)等；

(2) 使用者識別(User Identification)：透過 IP Address 或 Agent 等資訊來識別是否為相同使用者。而單一使用者在瀏覽過程中必定是連續的連結；

(3) 瀏覽序列識別(Session Identification)：為了解使用者是否會在一段時間後、或基於不同目的重複瀏覽相同網站等；

(4) 路徑完成(Path Completion)：最後透過拓樸的方式即可完成使用者瀏覽之序列路徑。

2.6 雲端同步儲存(Cloud Synchronous Storage)

近幾年間雲端運算(Cloud Computing)及雲端儲存(Cloud Storage)相當盛行。因個人通訊裝置逐漸多樣化的影響，透過公有雲的開放特性，將不同來源檔案進行線上資料儲存具有良好的特點。

隨時透過任何網路裝置進行資料取用(Broad Network Access)之雲端特性，對於擁有較多裝置的使用者而言，於雲端建置「軟體即服務」(Software as a Service, SaaS)平台可讓使用者輕易透過軟體安裝的方式，簡單的建立客戶端入口。並可透過此應用程式入口進行應用服務。

此外考量企業成本問題，透過「平台即服務」(Platform as a Service)進行伺服器租用服務，並於其上進行儲存、分析與推論等動作，將可降低線路與硬體維護成本。

本研究考慮到現今使用者，大多數人多可能擁有數種行動裝置(PC, Notebook, Smart Phone, Tablet)，其瀏覽資料亦同時散佈於各種裝置媒體上。因此本研究將透過雲端作為個人化資料共同儲存媒介。並透過雲端伺服器進行運算等服務。

3. 系統架構

本研究將以雲端服務為基礎進行相關應用。首先透過一個安裝於客戶端之軟體代理人

(Agent)，進行使用者瀏覽紀錄收集、上傳與隱私設定。使用者瀏覽紀錄將上傳至雲端伺服器進行儲存。接著於雲端伺服器上，透過數種偏好推論法建立通用個人輪廓(Universal User Profile)，並以此輪廓進行推薦服務應用。最後則建立一 API 介面，提供作為第三方服務應用使用。圖 3 為本研究之系統架構圖，其架構可概略分為客戶端與雲端之環境。以下將簡單介紹各元件之用途。

客戶端環境中，其主要功能為收集使用者之網路瀏覽紀錄(Common Web Log)，接著透過設定 P3P 隱私協定後將資料上傳至雲端儲存空間，以供偏好推論運算使用。

雲端環境中分為兩個主要部分：雲端儲存部分，主要負責儲存由客戶端傳送過來之使用者網路瀏覽紀錄。而原始記錄經過偏好分析後即成為通用個人輪廓，亦儲存於此環境中。在雲端運算部分，透過 PaaS 服務(如 Amazon EC2、Google Cloud Platform)所提供之運算服務，負責進行網路分類檢索、偏好推論與推薦服務等工作。並提供 API 介面供企業應用。

接著為各元件之詳細介紹。

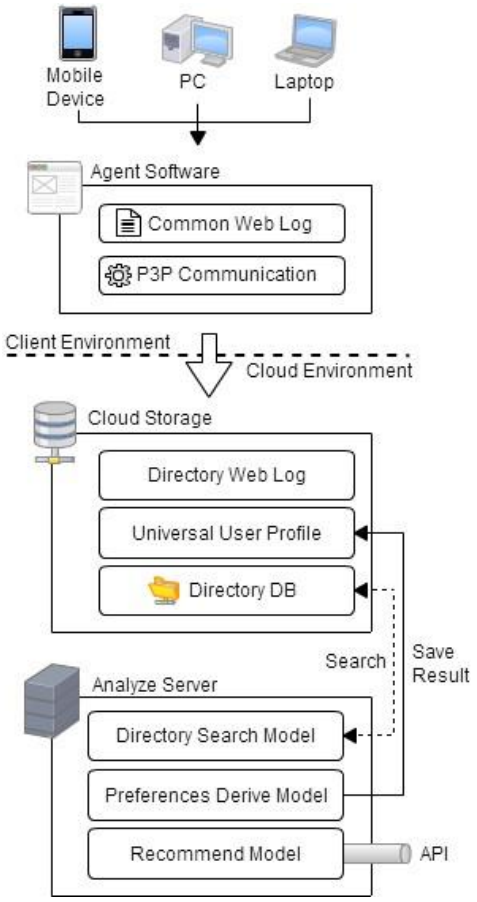

圖 3、系統架構圖

3.1 代理人軟體(Client Agent)

代理人軟體的設計如圖 4 所示，此部分為一個安裝於客戶端之代理人程式。主要先從使用者之各種裝置間，透過每一種裝置之網頁瀏覽程式(Web Browser)，取得使用者之歷史瀏覽資料(Bowering History)。此資料包含使用者瀏覽不同網站之網路瀏覽行為。收集完成之紀錄包含瀏覽紀錄編號(Browsing ID)、存取時間(Time)及瀏覽網址(URL)等，稱為通用瀏覽紀錄(Common Web Log)。

使用者接著可進行隱私瀏覽紀錄的偏好設定，即如同進行私密瀏覽(Private Browsing)之動作。避免將如特定頁面、信用卡、購物車等個人重要隱私網頁紀錄上傳。

在確認完隱私設定後，即可接收該上傳目的地伺服器之 P3P Policy 宣告。再與 P3P 信任服務(Trust Service)取得上傳授權，即可開始將使用者瀏覽資料上傳到雲端。

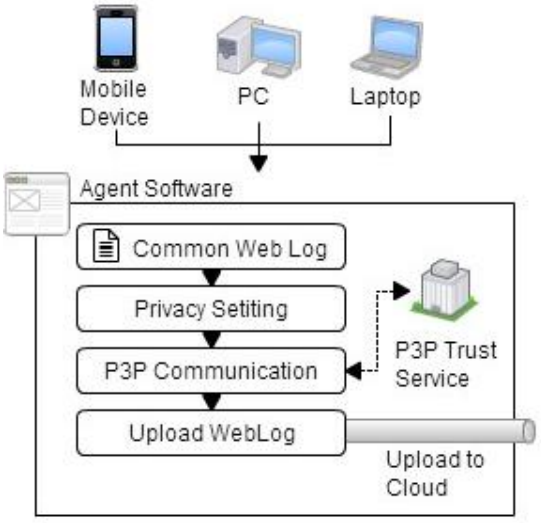

圖 4、代理人軟體架構圖

3.2 分析伺服器(Analyze Servers)

分析伺服器的設計如如圖 5 所示，分析伺服器主要為建置於雲端上之運算服務。並配合雲端儲存(Cloud Storage)進行使用者個人資料之儲存。主要功能為將使用者之瀏覽紀錄(Common Web Log)，透過分類檢索、偏好推論等動作後，再產生個人化的推薦內容。以下小節將介紹詳細內容。

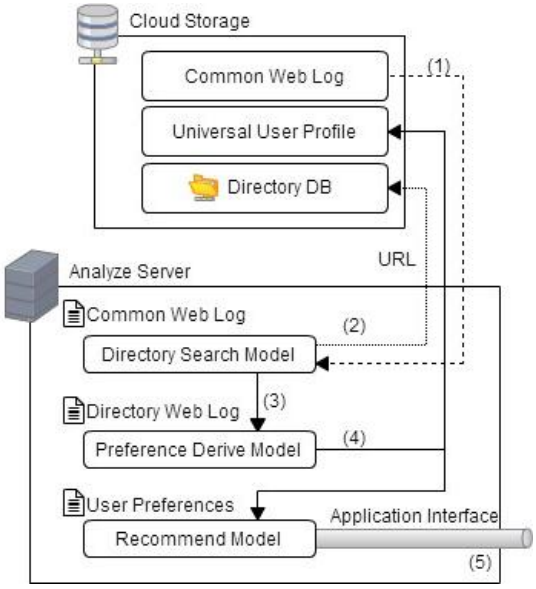

圖 5、雲端分析伺服器與儲存架構圖

3.2.1 網路分類檢索模組(Directory Search Model)

為了將一般瀏覽紀錄(Common Web Log)轉換為可使用的分類網路瀏覽紀錄(Directory Web Log)，因此需透過「前處理」與「分類檢索」兩步驟，將原始資料轉換為可用資料。

前處理其主要目的為過濾不易分析的網路資源，節省系統處理時間並減少雜訊對偏好分析的影響。常見之雜訊為網頁中之圖片、多媒體資訊(Flash, 音樂等)或各種程式檔等。因此透過副檔名之過濾分析，可有效降低需分析之紀錄量。

接著在分類檢索部分，是將使用者之一般瀏覽紀錄(Common Web Log)中的網址(URL)欄位，跟網路分類資料庫(Directory Database)中的網路分類(Directory)，兩者間進行比對(Matching)動作。最後將取得之「分類網路瀏覽紀錄」(Directory Web Log)傳送至偏好推論模組進行相關應用。

3.2.2 偏好推論模組(Preference Derive Model)

在取得分類網路瀏覽紀錄(Directory Web Log)後，即可進行偏好推論的動作。本研究以三個方法[4]進行使用者偏好推論基準：

(1) 權重累積法(Weighted Sum Method)；

(2)序列識別法(Session Identification Method)；

(3)本體推論法(Ontology Inference Method)。

藉此三種方法，可得知通用個人輪廓及本體推論技術是否可提升系統了解使用者偏好的能力。以下小節將詳細介紹各方法。

3.3 權重累積法(Weighted Sum Method)

藉由分類網路瀏覽紀錄，可得知使用者瀏覽過哪一些網頁，因此若對記錄中各個分類進行權重加總，即可知道使用者喜歡什麼樣的網路類別。如公式(1)[4]所示：

$$W\_Count_j = \sum_{i=1}^{n} W_{i,j} \quad (1)$$

W_Count_j 為第 j 個網路分類之權重值；i 值為各瀏覽紀錄之編號，j 值為各個網路分類。此方法直接從分類網路瀏覽紀錄計算使用者瀏覽各個分類之頻率，透過簡單的計次法可直接了解通用個人輪廓描述偏好的能力。

3.3 序列識別法(Session Identification Method)

在分類網路瀏覽紀錄中，每一筆均記載了瀏覽順序、瀏覽時間、網址及分類名稱。透過這些資訊與網路習性探勘技術，對分類網路瀏覽紀錄進行分析，將使用者偏好單位，由每個分類的「瀏覽次數」轉化成為「瀏覽序列」。

以往的網路習性探勘需識別在瀏覽序列中的目標使用者資訊，接著再以瀏覽時間及網站拓樸(Web Topology)，來分析使用者的動作是否是連續行為。但因本研究之使用者紀錄均為來自同一人之瀏覽紀錄，故不需進行前述第一步驟。但由於本研究中網路瀏覽紀錄包含許多網站記錄，因此無法進行拓樸結構。故應以另一種瀏覽序列識別為辨識依據。

本研究認為使用者未完成特定目的，故在瀏覽過程中必會連續讀取相同分類的網頁。若在過程中出現其他領域分類，即為另一個目的瀏覽動作。此變化視為一個瀏覽序列終點。

而雖然連續讀取同領域網頁，但若網頁讀取間隔時間超過特定門檻值，雖然其目的未變更但應仍視為瀏覽序列的一個終結，重新賦予權重。因此本研究定義兩條網路瀏覽分割原則：

(1)若前後兩筆資料包含網路分類不同，則

進行瀏覽序列分割；

(2)若前後兩筆紀錄的網址讀取時間超過門檻值，則進行瀏覽序列分割。

圖 6 為序列識別法之流程圖，Threshold 為瀏覽序列的時間門檻值：Record 為網路瀏覽紀錄中之一筆紀錄；Session 為瀏覽序號。每一個瀏覽序列均由幾個連續的 Record 組成。

分割後的瀏覽序列(Session)其權重內容計算方式如公式(2)[4]：

$$W_{i,j} = 1/S\_Length_k \quad (2)$$

透過「網路分類」及「時間門檻值」兩者進行序列切割後，雖然可能不了解各網站拓樸架構，但仍然能夠切割成適當的瀏覽序列，有效降低瀏覽行為過多的狀況與降低雜訊等。

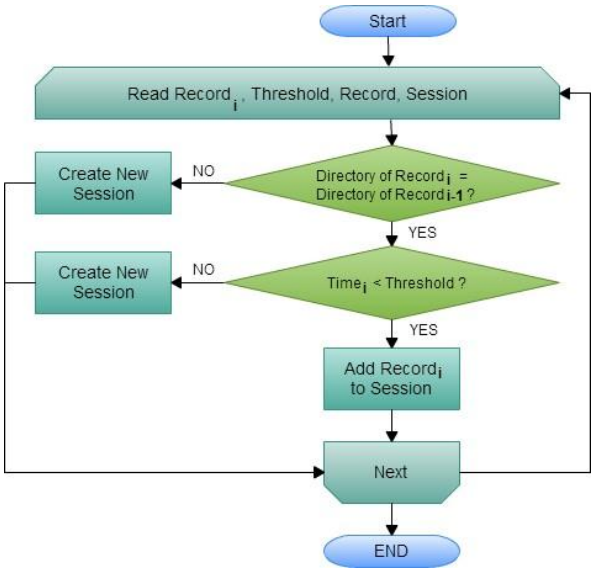

圖 6、序列識別法流程圖

3.4 本體論推論法(Ontology Inference Method)

根據過去研究發現，本體論之階層式架構有助於找出使用者未曾點擊過，但卻有可能喜歡之分類，稱之為「潛在偏好」。在進行此動作前，須先將使用者瀏覽過之網路分類建立「個人偏好樹(User Preference Tree)」。

個人偏好樹為網路分類資料庫的子集合。它將使用者曾經直接瀏覽過的網頁與網站分類資料庫比對，此種偏好也稱為「直接偏好(Direct Preference)」。而與直接偏好有連結者則稱為「間接偏好(Indirect Preference)」。

例如使用者查詢「朝陽科技大學」後，所得到之查詢結果為「教育學習/大學校院/朝陽科技大學」，其結果如圖 7 所示。

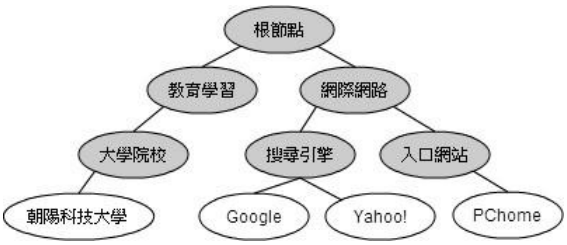

圖 7、簡單個人偏好樹

則使用者直接偏好為朝陽科技大學，對於其父節點大學院校、教育學習則具有間接偏好。而若在搜尋過程中，使用者曾經瀏覽過與朝陽科技大學有相關之網站，則亦會建立相關概念，最終建立成如圖 7 之個人偏好樹。淺色概念為使用者直接點擊過，為直接偏好；深色部分未直接點過，但為淺色概念之間接偏好。

在長期收集完成後偏好樹將越趨完整。而透過本體推論法，將可從結構進行推論使用者的「潛在偏好」。即某一概念未曾被點過，但由於數個子概念都收到使用者喜愛，則這個偏好就可能成為潛在偏好。

通常位於越底層之概念越為具體，應該使用者偏好來說較為重要，應保留下來。而潛在偏好的推論程序為(1)權重分享、(2)分類聚合與(3)設定聚合上限。

在權重分享階段，將網路分類原始權重分享一部分給其父分類。經分享後而增加權重的網路分類稱為「候選網路分類」。

在分類聚合階段，主要是從候選網路分類中產生使用者「潛在偏好」。此階段將候選網路分類與直接偏好進行比較：當前者大於後者時，將候選網路視為新的直接偏好。並將原本的直接偏好從偏好樹中移除。而潛在偏好在剛被發現時，它必須與現在的分類進行競爭。優勝者才可成為推薦項目。此階段可避免類似的分類概念重複出現，並留下重要子分類。

在設定聚合上限部分，主要目的為避免個人偏好樹的傾斜。即可能其中一子樹權重過大，其他子樹因重要性不及導致消失。最終推薦結果只能得到顯著偏好，而忽略其他部分。其層級多寡需視偏好樹分支度而定。

權重分享(Weight Sharing)計算公式如公式(3)[4]所示：

$$W\_Candidate_j = W_j + \sum_{m=1}^{n} W\_Son_{j,m} * S \quad (3)$$

S 為權重分享強度(Share Strength)，

$W\_candidate_j$ 為權重分享後的候選權重值，$W\_son_{j,m}$ 為第 j 個網路分類的第 m 個子分類。在權重分享過程中，較多子分類的項目其權重會大幅提升。

概念聚合(Concept Aggregate)則透過公式(4)[4]進行判斷：

If $W\_candidate_j > W\_Son_{j,m}$，then
Add $W\_candidate_j$ to *R_list*
From *R_list* delete $W\_Son_{j,m}$ (4)

比較 W_candid_j 與 $W\_son_{j,m}$ 權重值，若 $W\_candidate_j$ 較大，則取代 $W\_son_{j,m}$ 成為推薦列表的一員。本體推論法之流程圖如圖 8 所示。

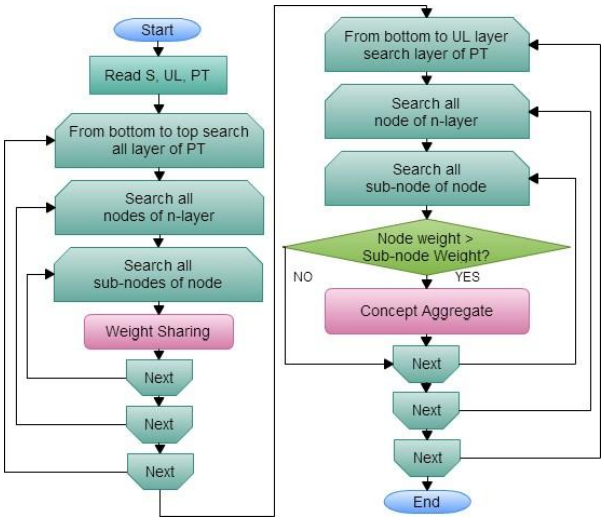

圖 8、本體論推論流程圖

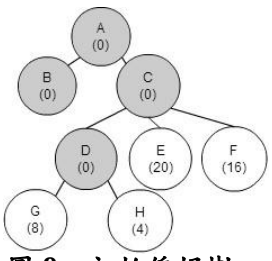

圖 9、初始偏好樹

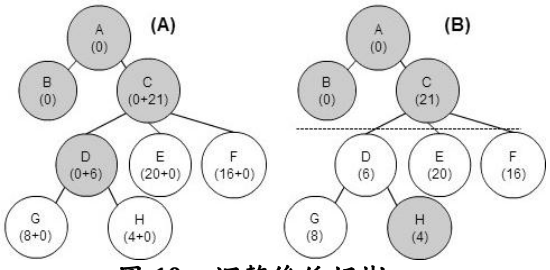

圖 10、調整後偏好樹

舉例而言，若權重分享強度設 0.5，聚合上限設為 2，其初始偏好樹如圖 9 所示。使用者曾經瀏覽過的網頁共有 E、F、G、H，因此剩餘節點即是間接偏好。括弧內之數值為瀏覽次數，亦稱為分類權重 W。故原始權重強度排名為：E、F、G、H。接著透過公式(3)進行計算，則 D 分類之新權重值計算公式如下：

$$W\_CandidateD = 0 + WG * S + WH * S = 0 + 8 * 0.5 + 4 * 0.5 = 6$$

經過迴圈計算後，新的偏好樹如圖 10(a)所示。因聚合上限設定為 2，故第三層後即可進行聚合動作。由於 H 的權重小於 D，故以 D 取代 H。雖然 C 分類大於其所有子分類，但因其層級大於聚合上限，故不列入考量。最終之偏好樹如圖 10(b)所示。

聚合演算法具有下列特性：(1)聚合分類使權重增加，提升潛在分類的重要性；(2)聚合上限可防止權重被取代次數過多，但還是可能發生高權重樹枝完全取代某子樹枝的狀況；(3)代表性子分類仍會被保留。

4. 實驗設計

為了測試本體論之通用個人輪廓的個人化效果，因此須將系統進行實作並分析之。首先是將取得之個人瀏覽紀錄來源透過代理人程式進行 P3P 授權與上傳。接著由雲端運算伺服器，進行分析處理。取得使用者所瀏覽之網站資料意義後，再使用不同偏好推論方法進行輪廓建立。最後在以特定指標來進行優劣分析。

4.1 資料來源(客戶端)

針對客戶端使用者資料，收集使用者端之各種裝置的瀏覽歷史紀錄(Browsing History)。

此部分資料為使用者近期的瀏覽行為，也就是透過瀏覽器瀏覽網路時，其將頁面下載至客戶端的行為(Temporary Internet Files)。當使用者再次瀏覽該網頁時，瀏覽器會根據該網頁是否有更新，若有更新才會再次重新下載。此瀏覽器機制目的是節省下載上的負荷量，提升瀏覽速度。因此此資料可作為客戶端的瀏覽紀錄(Web Logs)。

但因原始網路瀏覽資料內包含有太多對分析無益處之資料。為減少雲端伺服器上之系統分析負荷量，以及因頻寬考量勢必得減少資料上傳量。故藉由客戶端之代理人軟體，先將使

用者原始瀏覽紀錄進行篩選，以減少效能消耗成本。因此本研究預計篩選資料如以下說明。

(1)圖片資料，副檔名為 jpg、gif、png 等。由於現今網頁多會以圖片作為裝飾，故將圖片剃除以減少紀錄量；(2)程式，副檔名為 js、dll、css 等。此部分多為輔助主畫面之效果，因此將之排除；(3)多媒體檔案，副檔名為 wmv、mp3、wav 等。多媒體資料因難以判別其內容，故意將之排除在記錄清單外；(4)廣告、文件檔及其他，副檔名為 txt、doc、pdf 等。此部分包含有一些彈出式廣告視窗(Pop-up)等文件將會過濾掉。且因一般文件檔案亦難以進行判別，故也排除。

4.2 分類搜尋模組(伺服器端)

在此模組主要將使用者之網路瀏覽紀錄(Common Web Log)，透過雲端運算平台進行網路分類搜尋與使用者偏好本體的建立。

首先需要定義一個本體論，用來解釋與儲存使用者的瀏覽偏好，也就是須取得現行之網路分類資料庫。在網路分類服務的相關服務中，較為常見為 Yahoo! Directory 與 ODP(Open Directory Projecy)。因 ODP 屬於免費開放式目錄，可直接取得相關之目錄資料，亦可直接線上查詢。故本研究採用 ODP 作為主要的分類資料庫。

在定義完本體論與取得分類資料庫後，第二步驟則是透過分類資料庫，於雲端伺服器上進行網路資源的定義及比對。

在此為一比對範例：如圖 11 所示，若使用者瀏覽記錄網址為「www.cyut.edu.tw」，則透過 ODP 分類資料庫取得的分類便屬於「World: Chinese Traditional: 參考: 教育: 大專院校: 台灣: 台中縣: 私立朝陽科技大學 (1)」。得到朝陽科技大學之分類後，即可將分類瀏覽紀錄(Directory Web Log)進行儲存。

圖 11、線上 ODP 查詢範例

4.3 偏好推論模組處理

此部分將以第三章之「累積權重法」、「序列識別法」及「本體推論法」等演算法來產生推薦結果。此三種方法最終會得到一分類瀏覽紀錄清單，在進行分數之排序後，即可得到使用者對於網際網路瀏覽行為上的偏好。而此資料由雲端伺服器提供一介面(API)，供相關網路服務可運用此資料，進行推薦行為等動作。以下小節將各別介紹此三種方法。

4.3.1 累積權重法

此方法藉由使用者讀取分類之次數，決定使用者之偏好強度。此方法之推薦分數將所有跟該網路分類有關的權重值進行總和，其分數應多會低於讀取次數。

4.3.2 序列識別法

此方法為網路習性分析技術，可分析分類網路瀏覽紀錄。由於此方法首先需要決定適當的分割「時間門檻值」，其目的為判斷使用者是否花太多時間於同一網頁上，表示使用者可能已離開電腦或已完成特定事項故停留在該網頁。

第二項則必須判斷「分類連續瀏覽」原則。即使用者讀取的前後網頁具有相同分類，下一筆紀錄若變換分類，等同於已完成此次瀏覽目的，完成此一序列。

因此實驗必須在取得使用者瀏覽紀錄後先進行統計動作。觀察其連續瀏覽比率及瀏覽間隔時間之比率圖，再依使用者狀況取得較合適

之時間門檻值。最後以圖 8 之序列識別演算法進行運算後即可取得瀏覽序列資料。其每一筆序列之權重分配則如公式(2)。例如：若五筆紀錄被分割為同一序列，表示每一筆紀錄之權重值為 1/5。

4.3.3 個人偏好樹與本體推論法

此方法首先以使用者之網路分類瀏覽紀錄(Directory Web Log)建立個人偏好樹後，以瀏覽次數作為直接偏好的權重。接著透過公式(3)計算出潛在偏好的候選權重。即可推論哪一些概念是使用者之潛在偏好，並進行推薦動作。

而在設定聚合上限的部分，因其分類目錄前一至三層為概念較龐大之節點，故需判斷使用者之偏好樹狀況，再給予較合適之聚合上限。

藉著網路分類繼承關係，本研究可得知使用者喜歡的網路分類大多屬於哪一類型。因此可發現一些使用者未曾瀏覽，但卻有可能喜歡的潛在偏好。且透過概念聚合(Concept Aggregate)的步驟，亦可減少部分重複出現的喜好。

4.4 應用程式介面(API)

由前面所述之偏好推論模組產生的使用者之分類瀏覽紀錄清單進行排序後，即為使用者之偏好資料。而本研究之系統再提供應用程式介面，供第三方廠商在有限授權的狀況下進行資料取用，以進行推薦服務等應用。

4.5 實驗評核指標

為了評估實驗準確性，本研究透過兩個評核指標「預測準確率」(Precision)及「平均絕對誤差」(Mean Absolute Error)做為個人化結果之成效標準。其次則建立個人化網路推薦問卷，以取得使用者對推薦效果之評價。

預測準確率常用來評估推薦系統之成效，因此可透過使用者之回饋為喜歡之推薦，如公式(5)所示。

$$precission = \frac{|correct \cap recommend|}{recommend} \quad (5)$$

因此為了解推薦列表與使用者期望之差距，故以平均絕對誤差做為評估指標。在系統產生推薦列表後，推薦模組會請使用者填寫項目在使用者心目中適合的排列順序。藉兩者之差距即可計算推薦系統是否發生效用。其計算方法如公式(6)所示。

$$Error = X_i - Y_i$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|X_i - Y_i| \quad (6)$$

而問卷部分，則將透過前述三種偏好計算方法所產生共 15 項推薦項目，並建立共有兩子題之問卷進行施測。第一子題詢問使用者是否喜歡該推薦項目，即可透過預測準確率進行計算。第二子題則依喜好對推薦項目進行排名，即可計算平均絕對誤差。

5. 結論

本研究建置一個雲端環境，並設計一個客戶端之代理人程式於各種不同平台裝置中，透過 P3P 隱私權限進行個人瀏覽資料之收集。而為降低客戶端之運算需求，因此將偏好運算與分類搜尋運算等部分移至雲端服務執行。

且除了將通用個人輪廓移至雲端空間託管外，亦可透過雲端平台建立應用程式介面(API)供其餘第三方服務進行資料取用。讓使用者之個人資料可永續使用，並可於各種網路服務中提供較佳的個人化效用。

對於企業而言，可節省資料儲存、保管、分析等營業維護成本，且其資料可供無數的網路服務進行取用；對於使用者而言，其資料一致性將可良好維持。且因輸出之推薦資料已經過分析，故非使用者之原始隱私資料，可有效避免個人隱私直接的洩漏出去。

參考文獻

[1] 邱永祥，”運用類神經網路與資料探勘技術於網路教學課程推薦之研究”，碩士論文，朝陽科技大學資訊管理系，2003。

[2] 查士朗(2003)，網路時代個人資料管理系統之設計與實作，博士論文，國立台灣大學資訊管理系，台北。

[3] 曾信誠(2004)，以本體論為基礎之使用者喜好萃取、隱私權控管與側寫建構，碩士論文，國立東華大學，花蓮。

[4] 歐仁德(2005)，結合本體論與通用個人輪廓於個人化推薦之研究，碩士論文，朝陽科技大學，台

中。

[5] Arlein, Robert M., Ben Jai, Markus Jakobsson, Fabian Monrose, and Michael Reiter(2000), "Privacy-Preserving Global Customization," ***The Second ACM conference on Electronic commerce***, pp. 176–184.

[6] Belkin, N. J. and Croft, W. B., "Information Filtering and Information Retrieval: Two Sides of the Same Coin?," ***Communications of the ACM***, Vol. 35, No. 12, pp. 29-38, 1992.

[7] Cingil, Ibrahim, Asuman Dogac, and Ayca Azgin(2000), "A Broader Approach to Personalization," ***Communications of the ACM***, Vol. 43 No. 8, pp. 136–141.

[8] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava(1997), "Web Mining: Information and Pattern Discovery on the World Wide Web," ***International Conference on Tools with Artificial Intelligence***, pp. 558-567.

[9] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava(1999), "Data Preparation for Mining World Wide Web Browsing Patterns," ***journal of Knowledge and Information Systems***, Vol. 1, No. 1, pp. 5-32.

[10] Felipe S.J, Joan F., "Usability of Browser-Based Tools for Web-Search Privacy," Technical rept, March 2010.

[11] Hanani, Uri, Bracha Shapira, and Peretz Shoval(2001), "Information Filtering: Overview of Issues, Research and Systems," ***User Modeling and User-Adapted Interaction(UMUAI)***, Vol. 11, Issue 3, pp. 203-259.

[12] Jingjing Liu (2010), "Personalizing information retrieval using task stage, topic knowledge, and task products," ***ACM SIGIR Forum***, Vol. 44, No. 2, pp. 87-87

[13] Labrou, Yannis and Tim Finin(1999), "Yahoo! as an Ontology- Using Yahoo! Categories to Describe Documents, " ***Proceedings of the Eighth International Conference on Information and Knowledge Management(CKIM99)***, pp. 180-187.

[14] Microsoft, Microsoft Passport, Technical White Paper, March 2001.

[15] Middleton, Stuart E., Nigel R. Shadbolt, and David C. De Roure(2004), "Ontological User Profiling in Recommender Systems," ***ACM Transactions on Information Systems***, Vol. 22, Issue 1, pp. 54-88

[16] Mariam D., Lynda T. L., Mohand B., Bilal C., (2009), "A session based personalized search using an ontological user profile," ***ACM symposium on Applied Computing***, pp. 1732-1736

[17] Nicolaas M., Filip R.(2011), "Personalizing web search using long term browsing history," ACM international conference on Web search and data mining, pp. 25-34.

[18] Noy, Natalya F. and Deborah L. McGuinness(2001), "Ontology Development 101: A Guide to Creating Your First Ontology," ***Stanford Medical Informatics Technical Report*** ,SMI-2001-0880.

[19] Open Directory Project (ODP), retrieve at: http://dmoz.org/.

[20] Platform for Privacy Preferences Project (P3P), Retrieve at:http://en.wikipedia.org/wiki/P3P

[21] Pfitzmann, Birgit (2003), "Privacy in Enterprise Identity Federation Policies for Liberty Single Sign-On," ***Proceedings of the 3rd Workshop on Privacy Enhancing Technologies***, pp. 26-28.

[22] Surprenant, C. F. and M. R. Solomon(1987), "Predictability and Personalization in the Service Encounter," Journal of Marketing, Vol. 51, pp. 86-89.

[23] The World Wide Web Consortium – Platform for Privacy Preferences (P3P), W3C Resource, Retrieve at: http://www.w3.org/P3P/.

[24] Wang, Huaiqing, Matthew K. O. Lee, and Chen Wang(1998), "Consumer Privacy Concerns about Marketing," ***Communications of the ACM***, Vol. 41, 96 Issue 3, pp. 63-70.

[25] Web-Ontology(WebOnt) Working Group, Retrieve at: http://www.w3.org/2001/sw/WebOnt/.