助理型軟體系統之研製與應用
子計畫三：助理型軟體之製造及個人化技術
Implementation and Personalization of Software Agents

計畫編號：NSC 88-2213-E-002-007
執行期限：87/08/01~88/07/31
主持人：許永真 副教授 (yjhsu@csie.ntu.edu.tw)
執行機構：國立台灣大學資訊工程研究所

一、中文摘要

隨著電腦軟硬體的進步，人們對於軟體的期望不再只是進行傳統的資料處理，而是能主動蒐集資料以解決問題，這也就是研究助理型軟體（software agents**）的動機。本計畫將重點放在發展下列技術：**

1. 多重助理軟體系統：

利用 CORBA **分散式物件架構的模式，賦予助理軟體在分散、開放的環境中生存的基本能力，以及相互溝通的標準介面。並以** KQML **為助理軟體之間的溝通語言。此外，我們定義了多重助理型軟體合作模式，提供了多重助理軟體系統動態增減助理軟體的能力；使得藉此基礎建設建構之多重助理軟體系統，可同時確保系統的開放性及助理軟體之間的合作關係。**

2. 網際網路的資訊擷取：

針對網路文件的特色，我們發展出了以樹狀結構樣版（template tree**）為基礎的資訊擷取的方法，可以用來作為網路文件資訊自動擷取的工具。**

3. 助理型軟體的個人化：

以電子郵件助理軟體為實驗對象。以文件分類的貝式分類元（Bayesian Classifier**）技術為基礎，學習使用者對電子郵件的偏好，並對新信件提供分類及排序的建議。**

關鍵詞：助理型軟體、個人化、多重助理軟體系統、資訊擷取、貝式分類元、CORBA、HTML **文件**

Abstract

With the rapid growth in computer technology, software is no longer limited by the traditional view of “programs”. A new concept of “software agents”, which proactively collect information and solve problems on behalf of the users, has emerged. This project aims to develop the following techniques for software agents:

1. Multi-agent infrastructure:

The infrastructure for multi-agent system is designed based on CORBA. It provides a convenient platform for agents that communicate with each other in order to accomplish a task. A subset of KQML is adopted as the communication protocol. The architecture is flexible in that agents can be added or removed dynamically, and it facilitates cooperation among agents in multi-agent systems.

2. Information extraction:

In this project, we analyzed the structure of standard HTML documents, and defined the idea of a “template tree” that enables automatic extraction of target information from semi-structured HTML documents.

3. Personalization of software agents:

We implemented an email agent, and experimented using Bayesian classifier in learning a user’s preferences in classifying and prioritizing emails.

Keywords: software agents, personalization, multi-agent system, information extraction, Bayesian Classifier, CORBA, HTML

二、緣由與目的

隨著電腦硬體計算能力的提昇，人們對於軟體的期望不再只是傳統的程式，而是能解決真實世界問題的助理，也就是助理型軟體（software agent）。助理型軟體並非僅是一時髦的觀念，它會對軟體系統之設計方法造成極大的衝擊與改革。最常見的助理型軟體工作包括個人資訊之處理，如電子郵件、新聞、與電子討論區之篩選；又如個人開會、行程安排、電子購物比價、股票買賣、網路資訊搜尋、列印工作處理等，均可見到助理型軟體之蹤跡。

然而，真實世界問題的複雜度，往往不是上述的單一助理型軟體所能處理的。舉例來說，我們如果想列印某日的頭條新聞。我們可以在 news agent 上增加 printer agent 所有的功能，不過這樣會使 news agent更難實作。如果我們能將不同的助理軟體集合成一個多重助理軟體系統（multi-agent system），定義好助理軟體之間的平台及溝通語言，那麼這件工作就可以由 news agent 和 printer agent 合作完成。

其次，建立一個新的助理軟體也不是件簡單的事，因為它往往會牽涉到多方面的專業知識。例如，我們可能希望助理軟體能了解使用者輸入的自然語言查詢，然後去相關的網站找尋可能的資料。這樣的助理軟體就需要自然語言處理和資訊擷取方面的知識。所以，除了需要多重助理軟體環境，藉由助理軟體彼此的合作來增加服務的水準，我們還需要建立一些發展助理軟體常用的工具，使得專業知識所造成的阻礙能夠減小。

個人化是智慧型助理軟體研究的重點之一。個人化的助理軟體會記錄使用者的個人偏好，並且在使用者的使用過程中，對於偏好記錄作進一步的調整，使得助理軟體的偏好記錄能更趨近使用者真正的偏好。有別於傳統的應用程式，個人化的助理軟體記錄與學習使用者偏好的能力，使得程式可以對於不同的使用者進行調整，而非以往全然是使用者去適應程式。

目前在個人化的研究方面，比較有進展的是建立偏好記錄，來判斷使用者喜歡閱讀哪些文章。這類技術有被應用在網頁的瀏覽上。使用者在瀏覽網頁時，助理軟體會根據使用者的偏好，來決定目前網頁中的哪些超鍊結所指到的文件會讓使用者感到有興趣；同時，助理軟體也會觀察使用者的使用情形，來修改使用者偏好記錄。使用者偏好記錄是以關鍵字詞為基礎，助理軟體將使用者喜好的文章轉換成關鍵字詞的集合，去除了彼此之間的相關性，然後計算和使用者偏好的相似程度。相似程度愈大表示該網頁愈為使用者所喜愛。

本計畫依照前敘的方向進行。在多重助理軟體系統方面，我們提出一個架構，來作為多重助理軟體系統中，各個助理軟體之間賴以溝通並進而合作的基礎建設。我們利用了分散式物件架構中的 CORBA ，來賦予助理軟體在分散、開放的環境中生存的基本能力，以及相互溝通的標準介面，並且以 KQML 為助理軟體之間的溝通語言。此外，我們定義了多重助理軟體合作模式，提供了多重助理軟體系統動態增減助理軟體的能力；使得藉此基礎建設建構之多重助理軟體系統，可同時確保系統的開放性及助理軟體之間的合作關係。

在發展助理軟體的工具上，我們著重於網際網路上的資訊擷取。網際網路的快速成長已經改變了人們處理日常生活資訊的方法及習慣。有愈來愈豐富的資料是以 HTML 文件的格式呈現在網路上，因此以全球資訊網（WWW, World Wide Web）為工作平台的資訊處理系統已逐漸取代一般的資訊處理系統。為了能夠有效地利用已存在的大量線上資料，需要各式各樣的資訊擷取（information extraction）系統。針對網路文件的特色，我們發展出了以樹狀結構樣版（template tree）為基礎的資訊擷取的方法，可以用來作為網路文件資訊擷取的工具。

在個人化方面，我們則是以大家使用頻繁的電子郵件為實驗對象。隨著網際網路的普級，電子郵件常被用來通訊，甚至是作為廣告宣傳用途。如果時間有限，信件又多時，我們會希望挑選出有重要、或是有興趣的信先看。在看完信後，有些人會有將信件

歸類的習慣。我們可能得瀏覽信件標題、送信人等資訊，甚至有時要看一小段信件的內容，來判斷它是否重要；看完之後可能還要幫信件分類，免得以後它被大量的其他郵件淹沒。針對這個問題，我們認為應該發展電子郵件助理軟體，來學習電子郵件使用者不同的習慣，進而幫助使用者挑選重要信件，以及對信件分類，來節省使用者花在電子郵件閱讀與管理的時間。我們以文件分類方面的貝式分類元（Bayesian Classifier）技術為基礎，來進行助理軟體對個人信件的分類習慣以及重要性看法的學習效果的實驗。

助理型軟體系統將會對人們的生活有很大的助益。本研究對於助理型軟體的製造技術方面，提出了多重助理系統的架構，以及可以作為助理型軟體發展工具的網頁資訊擷取技術；在個人化技術方面，則是對貝式分類元技術應用於學習個人分類電子郵件和對郵件重要性的判斷方面進行了實驗。這兩個方向都是未來助理型軟體發展的重點。本計畫的研究結果，除了可作為相關研究的參考，我們也會逐步將助理型軟體賦予這些技術，進而增強助理軟體對於人們的的幫助。

三、結果與討論

(一) 多重助理軟體系統

由於多重助理軟體所面臨的環境具有分散式、開放式的特性，而分散式物件架構 CORBA 可以提供助理軟體在這樣的環境中生存所需要的基本能力，所以最終，我們以 CORBA 為基礎，利用 CORBA 定義中的 IDL (interface definition language) 定義出助理軟體的外觀，使得助理軟體之間的相互呼叫有一個標準的方式。利用 IOR (interoperable object reference) 來作為分散式環境中，助理軟體的位址。並以 KQML 作為助理軟體之間相互溝通的共同語言，定義出多重助理軟體系統的基礎建設－agent request broker，使得助理軟體可賴以在分散式的環境中生存、溝通。

再配合我們所定義之一簡單的多重助理軟體合作模式。在此一合作模式中，有著 butler agent、professional 助理軟體以及 tool 三種角色。其中的運作方式是：對於某一問題，butler agent 依照所牽涉到的領域不同，將問題切割成不同的次問題，並分別找到適當的 professional agent 來交付次問題。而對於一個次問題，professional agent 利用各種不同的工具來加以解決。這樣的合作模式就像是人類合作模式中的秘書、專家以及工具的關係。並在此一合作模式中提供 professional agents trading service 以及 tools trading service 兩種轉介服務，除了分別可供 butler agent 找到適當的 professional agents、以及 professional agent 找到適當的 tools 之外，並藉此達成此多重助理軟體系統的擴張性，使得此系統具有相當的開放性。[2][6][7]

(二) ***資訊擷取助理型軟體***

為了要達到有效地利用已存在的大量線上資料，其最重要的關鍵技術即在於是否能夠成功地將各個新聞網頁上的時事條例擷取出來。針對網路文件的特色，我們發展了一個資訊擷取助理型軟體。

由於文件可以被視為由三種組成分子構成：內容（content）、格式（format）和結構（structure）。內容是一份文件所要表達的真正含意，而格式則是文件呈現的方式，至於結構則表示了一份文章之內容順序性上的關聯性。對於同一類的文件而言，它們的內容可能不盡相同，格式也不見得會一樣，但是卻會有相似、甚至完全相同的結構，而這個結構會把整份文件區分為一個個獨立而有意義的資料區塊。這樣的特性給了我們一個啟示，即對於同一類的文件，結構的特性可以幫助我們去擷取出文件中有意義的資料部份。

針對網路文件的特性，一份 HTML 文件可以根據它的標籤（tag）而被表達成一棵文件樹，並進而表達出該文件的結構資訊。而相似的文件通常具有相同的文件結構，因此我們利用一個樹狀結構樣板（template tree）來表達這個相同的文件結構特性。透過一個樹狀配對法（tree

matching），我們可以決定樣板和文件之間的對應關係，進而從文件中擷取出所要的資訊。

目前所發展出來的資訊擷取方法，已經被實際地應用於處理幾個常用的網路搜尋引擎知名網站（Yahoo!, Altavista, Openfind, MetaCrawler, CNET Shopper.com, Cora, etc.）及線上新聞網站（BBC News, CNET News.com, Infoworld, NBA, etc.）之網頁資料。很明顯地，網路搜尋引擎是網路應用程式的一個重要資訊來源；線上新聞網站是每個人每天資訊的重要來源之一。結果顯示了我們所提出的方法確實可以很有效地擷取出想要的資訊，也更加確認了這套方法的可行性及實用性。[1][3]

(三) 個人化技術

我們選擇了日常生活中常會使用的電子郵件軟體來作為實驗環境，藉以研究個人化服務所需之要素及技術，進而提供個人化的服務。

在電子郵件軟體所必須提供的服務中，我們從提供自動分類及管理郵件的功能出發，利用機器學習分類技術，在與使用者的互動過程中，學得使用者過去的習慣與偏好。具體來說，目前發展出來的技術，使電子郵件助理可以根據它對每個不同使用者的認識，自動管理電子郵件。而不僅僅是提供使用者管理郵件的工具。同時它必須能夠對新信件提供分類的建議，以及根據信件的重要性加以排序，使得使用者可以選擇最重要的電子郵件進行閱讀。[5]

在電子郵件助理的實作上，我們把焦點放在機器學習分類技術。針對這個問題，我們將在文件分類上表現良好的貝氏分類元（naive Bayes classifier）應用到郵件分類的問題上。給定一封電子郵件，電子郵件助理會先將之表述為具有四個特徵：寄信人，收信人，主題，以及內文。每一個特徵由關鍵字組成，然後再套用貝氏分類元分類之。每個使用者在使用過程中，對助理軟體建議的分類，可給予正確的回饋，藉以建立不同的偏好記錄，日後便能得到個人化的服務。此分類技術同時可應用在郵件自動分類或是重要性預測上。

經過實驗證實，無論是郵件自動分類或是重要性預測，一般郵件分類正確率可達70%以上，新聞論壇佈告更可達 80%-90%，因此本計劃之成果已具有實用性。[4]

本實驗中對個人化技術的實驗和研究也可以應用在其他類型的助理軟體中。

四、成果自評

本研究依原計畫進行，並已達成數項重要助理型軟體技術之開發。研究成果偏向於助理型軟體的基礎技術，但同時亦完成了印表助理、電子郵件助理、以及 FAQ 助理等雛形系統。本計畫致力於技術研究，提供良好的解決方案，對於現階段助理型軟體的研究有重要的貢獻。

1. 多重助理軟體系統勢必是助理型軟體的發展方向。我們利用 CORBA 作為系統的底層架構，使得將來發展助理軟體時，可以省去不少軟體之間溝通的問題。例如一個在 UNIX 環境上的助理型軟體可以方便地和一個 MS-Windows 上的助理型軟體溝通。同時，CORBA 有日益流行的趨勢，但是將它和多重助理軟體系統結合的研究還寥寥可數。所以這部份的研究是有不小的實用性和前瞻性的。
2. 資訊擷取技術方面，本研究著重在 HTML 文件上的資訊擷取，在開發系統資訊助理型軟體（system information agent）的過程中，面臨了一個問題：如何取得助理型軟體所需要之資訊？由於網路上已存在著大量的資料，因此著手於開發一有效的方法來迅速地將這些資料轉變成助理型軟體可用之資訊。目前已發展完成一個一般性的 HTML 文件之資訊擷取方法，因而可應用於各式各樣的線上網頁，比原先只期望可以處理 FAQ 文件更進步，也更具實用價值[8]。
3. 在個人化技術的研究方面，除了可以用

來發展個人化的電子郵件助理外，這項技術更可以直接應用在個人化新聞閱讀軟體，以及個人化文件過濾系統，實用性也是相當高的。而個人化技術的發展經驗，如互動式的介面，使用者的偏好紀錄等，亦可應用於其他助理系統的建構。

五、參考文獻

[1] 呂青鴻，樹狀配對法於 HTML 文件資訊萃取之研製與應用．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十七年六月

[2] 李步健，區域網路之代理型程式與服務共享架構．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十七年六月

[3] 易文韜，樹狀 HTML 文件之資訊擷取．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十六年六月

[4] 蔡宗翰，機器分類學習技術於個人化電子郵件助理之研究．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十八年六月

[5] 鄧宇雄，個人化電子郵件助理．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十七年六月

[6] 駱克明，建構於分散式物件上的多重助理程式合作架構．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十八年六月

[7] 簡銘維，應用於分散式網路環境之智慧型印表機代理程式．碩士論文，國立台灣大學資訊工程學研究所．中華民國八十六年六月

[8] Hsu, J. Y. J and Yih, W.-T. Template-based information mining from HTML documents. In Proceedings of AAAI-97, pages 256-262, 1997.

[9] Hsu,J. Y. J. A multi-agent framework for intranet service integration. In Toru Ishida, editor, Proceedings of Pacific Rim International Workshop on Multi-Agents, pages 26-37, November 1998.

[10] Cowie, J. and Lehnert, W. Information extraction. Communications of the ACM, 39（1）:80-91, 1996.

[11] Hao, X., Wang, J. T. L., and Ng, P. A. Information extraction from the structured part of office documents. Information Sciences, 91:245-274, 1996

[12] Kushmerick, N., Weld, D., and Doorenbos, R. Wrapper induction for information extraction. In Proceedings of Fifteenth International Joint Conference on Artificial Intelligence （IJCAI-97）, pages 729-735, 1997.

[13] Bradshaw, J. M. editor. Software Agents. AAAI/MIT Press, Menlo Park, CA, 1997.

[14] Sycara, K. P. Multiagent systems. AI Magazine, 19(2): 79-92,1998.