

國科會國際數位圖書館合作研究計畫
NSC's International Digital Library Collaborative Research

國科會國際數位圖書館合作研究計畫(IDLP)II-子計畫一
英中雙語資訊系統相關語言處理技術和資源整合之研究

NSC90-2750-H-002-731

主持人

陳信希 教授

台灣大學資訊工程學系

共同主持人

陳光華 教授

台灣大學圖書資訊學系

研究人員

林偉豪 林紋正 黃聖傑 蔡真慧

民國 91 年 7 月

新一代資訊傳播的特色是：網際網路突破空間距離，打造一個不分國界的資訊地球村。尤其透過全球資訊網，各地的資訊皆唾手可得，不但豐富且即時。在網際網路上流通的資訊除了數量非常龐大之外，所使用的語言種類也非常多。依據 1996 年 ETHNOLOGUE 目錄上的統計，全世界語言數目高達 6703 種(Grimes, 1996)。由語言使用情況來說，以 OCLC(Online Computer Library Center)為例，服務的對象超過 17,000 個圖書館，收錄的資訊歸屬於 370 個以上的語言；以全球資訊網為例，大約 80%的網站是英文網站，而將近 40%的網際網路使用者不會說英文(Hershman, 1998)。因此，如何一方面將數位圖書館這寶貴的資源，介紹給不同語言的使用者；另一方面如何廣為吸收其他語言所呈現的知識，都是資訊國際化不能忽略的重要課題。

如上所述，網際網路的重要特徵之一是無國界，所有資訊透過網際網路，超越地理位置的藩籬，語言文化的差異，送到網際網路使用者手上。而數位圖書館所擁有的大量文化資源，更是扮演網際網路內容提供者的重要角色，發揮文化傳播、教育、陶冶性情等多重功能。數位圖書館是內容與技術的整合，基本上有下列三項特色與前景：

- (a) 多媒體(multi-media)：透過不同媒體所呈現的內涵，更能引導不同層面的使用者，吸收數位圖書館的精華。
- (b) 多語言(multi-linguality)：網際網路所帶來的無國界特質，如何降低語言的障礙，呈現數位圖書館的內涵是重要課題。
- (c) 多文化(multi-culture)：由於網際網路的特殊資訊傳播功能，各個數位圖書館典藏的交流，會越來越密切。重要內涵彼此的觀摩，更帶動多重文化的比較，促進文化的融合。

因此，數位圖書館技術必須兼顧到上面的特徵。本報告分三部份，說明跨語言檢索相關議題：

- (1) 跨語言資料檢索的專有名詞翻譯
- (2) 反向異文字音譯相似度評量方法與跨語言資訊檢索

(3) 以學習語音相似度為基礎之反向機器音譯

第一部份：

跨語言資料檢索的專有名詞翻譯

摘要

近來，對於想在全球資訊網路上搜尋及檢索各種不同語言文件的人而言，語言障礙已成為主要的問題。這篇文章探討跨語言資訊檢索中查詢翻譯(query translation)的議題，特別對於專有名詞多所著墨，並且提出專有名詞確認、翻譯與搜尋的模式。在所搜尋的資料當中，確認中文組織名、人名、地名的召回率(recall rates)及準確率分別為中文組織名的76.67%，79.33%；中文人名87.33%，82.33%及中文地名的77.00%，82.00%。關於專有名詞的翻譯，只有0.79%的英文人名和1.11%的英文地名要薦。專有名詞搜尋的技術運用在機器翻譯上使網路資訊檢索的功能更加便捷。在此系統之下，使用者可以輸入查詢(query)，而且用他熟悉的語言閱讀文件。

介紹

全球資訊網是在網際網路上最有用也是最有力的資料散播系統。然而由於語言的多元化，對於想在全球資訊網路上搜尋及檢索各種不同語言文件的人而言，語言障礙儼然已成為主要的問題，這個現象在某種程度上削弱了網路傳播資訊的力量。跨語言資料檢索研究的縮寫為 CLIR(cross-language information retrieval) (Oard 和 Dorr, 1996;Oard 1997) 主要致力於超越資訊檢索的語言藩籬。CLIR 當中重要的議題包括：

- (1) 查詢字串和搜尋文件是不同的語言，所以翻譯是必要的。
- (2) 查詢字串當中可能有語意含糊不清之處，因此歧義分析(disambiguation)是必要的。
- (3) 查詢字串通常很簡短，因此適當的擴充是必要的。
- (4) 有些語言的查詢字串的範圍並不清楚(Chen and Lee, 1996)，因此適當的分割(segmentation)是必要的。
- (5) 同一份文件可能不只包含一種語言，因此語種的確定是必要的。

本文的焦點專注於查詢翻譯的議題，特別是對專有名詞的翻譯。

使用者查詢字串內包含專有名詞的比率相當高，根據一項在 1995 年以數天為期的實驗報導(Thompson and Dozier, 1997)指出，在查詢華爾街日報、洛杉磯時報和華盛頓郵報時，分別有 67.8%，83.4% 及 38.8% 的查詢(query)是以專有名詞來搜尋。在資訊檢索研究當中，有三項重點：專有名詞的確認，專有名詞的翻譯與專有名詞搜尋。因為專有名詞通常是未知的字，在單一語言的字典中很難查詢得到，更遑論雙語字典了。涵蓋範圍是以字典為本之方法(Ballesteros and Croft, 1996; Davis, 1997; Hull and Grefenstette, 1996)的主要問題之一。而以語料庫為本的方法(Brown, 1996; Oard 1996; Sheridan and Ballerini, 1996)則是從大規模的語料庫當中建立詞典。他們呈現了語言窄而精確的覆蓋度，與字典廣而模糊的特性互補。然而，以語料庫為本的方法之主要的限制在於領域的改變及術語排列的準確性。此外，專有名詞相較於文件中的其他字詞在語料庫裡屬於罕見的字。在資料檢索中，出現頻率最高和罕見的字都會被視為不重要，因而極容易被忽略。

本文將提出方法提取並且分類中文查詢中的專有名詞（第 1 節）。然後，將中文專有名詞翻譯為英文（第 2 節）。最後，再將翻譯過後的查詢字串(query)傳送到機械翻譯的伺服器在全球資訊網上做資料檢索(Bian and Chen, 1997)。搜尋到的英文首頁可以英文或中文呈現。

1 名詞的提取和分類

人、事、物、時、地點是構成一份文件的五個基本實體(entities)。如果能捕捉到這些基本的構成實體，我們對一份文件就能有某種程度的了解。而這些構成實體也正是使用者感興趣的對象。也就是說，使用者時常會輸入查詢(issue query)去檢索這一類的構成實體。這些基本的構成實體在自然語言文本中是主要的未知字詞，而且時常是以專有名詞的方式出現。因此，名詞的提取(name extraction)對自然語言的理解與資料檢索來說都是不可或缺的。

在著名的訊息理解系統評估(message understanding system evaluation) 與訊息理解會議(message understanding conferences 簡稱 MUC)和相關的多語言實體工作

multilingual entity tasks (MET) 中，涵蓋範圍包括命名組織、人名、地名、時間日期、幣制、百分比表現方式的附名實體，都是評估技術的工作之一。在 MUC-6 附名實體工作中，SRA (Krupka, 1995) 和 BBN (Weischedel, 1995) 發展出來的系統在人名辨識的部分有非常高的準確率（超過 94%）。

在中文語言處理領域，陳和李(Chen and Lee 1996) 提出幾種不同的策略，用以識別並且分類專有名詞的三大類型—中文人名、中文譯名、以及組織名稱。在幾項大規模的實驗的當中，平均的準確率是 88.04%，中文人名確認方面的平均召回率更高達 92.56%。

上述的方法可以用來收集全球資訊網(大規模的語料庫) 上的中英文專有名詞組。專有名詞的確認在查詢(queries)中與在大規模文本中是不同的，主要的差異在於詢問(query)通常很簡短。因此，它的內容勢必比全文短的多，用來處理長篇內文的技術就顯得無用武之地了。以下將描述我們在中文專有名詞確認上所採取的方法。

一個中文人名是由姓氏和名字兩部份所組成。大部分的中國姓氏都只有一個字，少數包含兩個字，已婚婦女可能在婚後冠上夫姓。如此一來就有三種可能的姓氏類型，即單姓，複姓和冠夫姓。大多數的名字有二個字，少數的名字只有一個字。理論上每個字都可能被視為是名字而非一個固定的組合(a fixed set)。因此中文人名的長度範圍為 2 到 6 個字。以下為採行的三種確認方式：

- (1) 公式化命名統計(name-formulation statistics)
- (2) 上下文提示。舉例來說，頭銜、職稱、語言行為動詞(speech-act verbs)等等。
- (3) 快取暫存(cache)

公式化命名統計形成基線模式。它提出可能的候選詞彙，上下文提示使適當的候選詞彙得到額外的加分。快取暫存區紀錄下同一段落中所有可能候選詞彙的出現率。如果一個候選詞彙出現超過一次，它很有可能就是一個人名。

音譯的人名代表外國人的名字，與中文人名相較，音譯命名並不受限於 2 到 6 個字。下列各項方法被用來辨識音譯命名：

(1) 字元條件(character condition)

二個特別的字集(character sets)為一組(setup)。音譯命名的第一個字和剩餘的字分別屬於這二個字集。字元條件(character condition)的限制並不嚴格，滿足字元條件的字串就代表它極可能是一個地點、一棟大樓或是一個地址等等。字串還應該與上下文的提示配合（參閱 (2) – (4)）。

(2) 頭銜(titles)

中文人名的頭銜也適用於音譯人名。

(3) 命名介紹者(name introducers)

有些字會在音譯命名第一次被使用時擔任介紹的角色。

(4) 特殊動詞

音譯人名中同樣會用到在中文人名當中的語言行為動詞(speech-act verbs)集 (set)。

暫存系統也有助於音譯命名的確認。一個候選詞彙滿足字元條件時，其中一個上下文提示(cues)將被存放在快取暫存區之內。下一次，上下文提示也許不會再出現，但是我們仍能藉由檢查快取暫存區尋回特定的音譯人名。

組織命名的結構比人名的結構更為複雜。基本上，一個完整的組織命名可以被區分為二個部份—即名稱(name)和關鍵字(keywords)。組織名、國名、人名和地名皆可被分類為組織命名當中的名稱(name)。人名可用上一段落中所敘述的方法搜尋，本文會在稍後論及地點命名。音譯命名也可能出現在名稱(name)的部份。我們採用在前一個段落中所提到的字集(character sets)。如果某字串(a sequence of characters)符合字元條件(character condition)，那麼此字串就會和關鍵字形成一個組織命名。一般的內容文字會穿插在名稱部分與關鍵字部分之間。在目前的版本中，最多只允許兩個內容文字。此外，我們利用組織命名會在文件當中重複出現特徵，提出了 n-gram 模式來處理這一個問題。雖然暫存系統(cache mechanism)和 n-gram

使用相同的特徵，及前述的高重複出現率，兩者的概念卻是完全不同的。以組織命名來說，因為我們難以決定一個樣式(a pattern)的左側邊界範圍，所以並不確定何時將它存入快取暫存區。

地點命名的結構與組織命名類似。一個完整的地點命名是由一個人名(或一個地名)和一個地點關鍵字所組成。對於不含關鍵字的地點命名，我們採用一些表示位置的動詞來解決問題。暫存系統也有幫助，而且我們亦可用 N-gram 模式重新搜尋那些不符合字元條件的命名。

我們以三組 MET 資料(即 MET-1 正式執行、MET-2 訓練和 MET-演練)來測試我們的系統。測試結果為中文組織命名，人名和地點命名的召回率與準確率分別是(76.67%,79.33%)，(87.33%,82.33%)和(77.00%,82.00%)。

2 專有名詞翻譯

在查詢翻譯中，中文和英文分別是原始語言和目的語言，這二種語言的字母完全不同。目前羅馬化中文的兩套著名系統為威式拼音(Wade-Giles)與漢語拼音(Lu, 1995)。專有名詞翻譯的問題闡述如下：

- (1) 從網路上收集英文的專有名詞組。
- (2) 從查詢(queries)當中確認中文專有名詞。
- (3) 將中文專有名詞羅馬化。
- (4) 從適當的專有名詞組中挑選出候選名單。

如此一來，翻譯問題就會轉移到聲符的配對問題(phonic string matching problem)。如果一個英文的專有名詞能對應到一個中文實體，配對是很簡單的事，例如 Lee Teng-hui 代表'李登輝'(中華民國總統)。否則的話，配對就沒那麼容易了，舉例來說，我們用中文輸入'阿爾卑斯山'的查詢去檢索關於 Alps 山脈的資料。這個名詞的漢語拼音是 a.er.bei.si.shan，然而字串'aerbeisishan'與字串'Alps'並不相似。我們逐步發展一些語言模型以應付翻譯問題，而所要考慮的第一個議題是在羅馬

化中文專有名詞和英文專有名詞中有多少共同的字母？在這裡次序是很重要的。譬如，中文查詢是'埃斯其勒斯'，它的威式羅馬拼音是 'ai.ssu.chi.le.ssu'，其相對應的專有名詞是 Aeschylus。三個相對應的字母(如畫底線的字母所示)依序為：

aeschylus
ais suchilessu

我們以候選詞彙的長度(即 9)為準去評斷，得分是 0.33。在實驗當中，我們準備了 1,534 對中英文人名並引導配對，使用中文專有名詞作為查詢，試著從 1,534 個候選詞彙中檢索出對應的英文專有名詞。我們以需要建議多少候選詞以包含正確答案來估算正確率，換句話說，藉此可得知正確翻譯的平均排名。威式拼音和漢語拼音系統的基線模式表現分別是 40.06 和 31.05。基線模型的主要問題是：如果一個字母被配對錯誤，那些接下來的字母將無助於配對，在上述的例子中， chi ('其') 對翻譯並無幫助。

為減少錯誤，我們預先考慮到候選詞彙的音節。配對工作在音節內完成而非整個字。舉例來說，Aeschyluss 包含三個音節。以下為配對的情形：

aes chy lus
aissu chi lessu

如此，得分增加到 0.67 (6/9)。在類似的實驗當中，新的語言模型表現有所進步，威式拼音和漢語拼音系統的平均排名分別為 35.65 和 27.32。

觀察威式拼音和漢語拼音在成果表現上的差異，我們發現它們使用不同的音符來標示相同的發音。舉例說明如下：

(1) vowels

p vs. b, t vs. d, k vs. g, ch vs. j, ch vs. q,
hs vs. x, ch vs. zh, j vs. r, ts vs. z, ts vs. c

(2) consonants

-ien vs. -ian, -ieh vs. -ie, -ou vs. -o,

-o vs. -uo, -ung vs. -ong, -ueh vs. -ue,
-uei vs. -ui, -iung vs. -iong, -i vs. -yi

一個新的語言模型整合了替代選擇，其配對(mate match)的平均排名是 25.39。實驗的結果比單一的羅馬化拼音系統來的理想。

在上述的排名中，每個字母在配對時俱有同等的影響力。我們要求每個羅馬化中文的第一個字母比其它的字母重要性更高，舉例來說，chi 的 c 比 h 和 i 更重要。因此它應該有較高的分數。下列為新的計分函數：

$$\text{score} = \sum_i (f_i * (el_i / (2 * cl_i) + 0.5) + o_i * 0.5) / el$$

其中

el : length of English proper name,

el_i : length of syllable i in English proper name,

cl_i : number of Chinese characters corresponding to syllable i ,

f_i : number of matched first-letters in syllable i ,

o_i : number of matched other letters in syllable i .

我們再重複上述的例子，第一個字母以大寫表示。

aes	chy	lus
<u>A</u> i <u>S</u> s <u>u</u>	<u>C</u> hi	<u>L</u> e <u>S</u> s <u>u</u>

對應的參數如下：

$el_1=3, cl_1=2, f_1=2, o_1=0, el=9,$

$el_2=3, cl_2=1, f_2=1, o_2=1,$

$el_3=3, cl_3=2, f_3=2, o_3=0.$

這個候選詞彙的新得分是 0.83。在新的實驗中，平均的排名是 20.64。如果羅馬化中文的第一字母沒有配對成功，我們予以扣分。改良加強後的模型排名是 16.78。

表一・人名翻譯的表現

1	2-5	6-10	11-15	16-20	21-25	25+
524	497	107	143	44	22	197

我們進一步的考慮英語的發音規則，例如 ph 通常會發成 f 。如果把所有類似的規則加入語言模型，平均的排名會提高到 12.11 。表 1 概述了正確候選詞彙的排名分佈情形。第一排代表排名的範圍，第二排顯現某個範圍內候選詞彙的數目，大約有三分之一的候選詞彙有第一的排名。平均看來，只有 0.79% 的候選詞彙必須被提出以覆蓋正確的解決方案，這表示此一方法相當有效。

我們還另外作了二個的實驗。在最好的的模型中輸入查詢(query) ，找尋英文的位置。 這個測試包含了 1,574 個候選詞彙，結果其平均排名是 17.40 。換句話說，1.11% 的候選詞彙已經被議。如果我們合併人名字集和地名字集，再重複實驗一次，其結果降低到 27.70 ，由此可見專有名詞分類的重要性。

結論

本文提供了從詞彙、句子和文章段落，到不同類型專有名詞辨識的資訊。人名翻譯的問題可視為發音字串的配對問題(a phonic string matching problem)。我們考慮到了配對詞彙的長度、音節、不同的羅馬化拼音系統、發音規則、評級時的正數與負數，以及在漢英資料檢索系統之內整合命名搜尋的機制。如此一來，使用者在網際網路上將不再受到語言阻隔。目前的實驗結果，只有 0.79% 英文人名 和 1.11% 的英文地名在命名翻譯時需要被議。

這個模型可用以建立雙語的專有名詞字典。我們能定期從網路上收集英文與中文的專有名詞，然後引導配對，並輔以人工作正確的翻譯選擇，如此將能減少費用在建立為搜索專有名詞而發展出的大規模雙語專有名詞字典上。

参考文献

- Ballesteros, L. and Croft, W.B. (1996) "Dictionary-based Methods for Cross-Lingual Information Retrieval." *Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications*, pp. 791-801.
- Bian, G.W. and Chen, H.H. (1997) "An MT-Server for Information Retrieval on WWW." *Working Notes of the AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 1997, pp. 10-16.
- Brown, R.D. (1996) "Example-Based Machine Translation in the Pangloss System." *Proceedings of 16th International Conference on Computational Linguistics*.
- Chen, H.H. and Lee, J.C. (1996) "Identification and Classification of Proper Nouns in Chinese Texts." *Proceedings of 16th International Conference on Computational Linguistics*, 1996, pp. 222-229.
- Hull, D.A. and Grefenstette, G. (1996) "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval." *Proceedings of the 19th International Conference on Research and Development in Information Retrieval*, pp. 49-57.
- Krupka, G.R. (1995) "SRA: Description of the SRA System as Used for MUC-6." *Proceedings of Sixth Message Understanding Conference*, 1995, pp. 221-235.
- Mani, I., et al. (1993) "Identifying Unknown Proper Names in Newswire Text." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 44-54.
- McDonald, D. (1993) "Internal and External Evidence in the Identification and Semantic Categorization of Proper Names." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 32-43.
- Oard, D.W. (1997) "Alternative Approaches for Cross-Language Text Retrieval." *Working Notes of AAAI-97 Spring Symposiums on Cross-Language Text and Speech Retrieval*, pp. 131-139.

- Oard, D.W. (1996) *Adaptive Vector Space Text Filtering for Monolingual and Cross-language Applications*, Ph.D. Dissertation, University of Maryland.
- Oard, D.W. and Dorr, B.J. (1996) *A Survey of Multilingual Text Retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies. <http://www.ee.umd.edu/medlab/filter/papers/mlir.ps>.
- Paik, W., et al. (1993) "Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval." *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text*, 1993, pp. 154-160.
- Sheridan, P. and Ballerini, J.P. (1996) "Experiments in Multilingual Information Retrieval Using the SPIDER System." *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 58-65.
- Thompson, P. and Dozier, C. (1997) "Name Searching and Information Retrieval." *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.
- Weischedel, R. (1995) "BBN: Description of the PLUM System as Used for MUC-6." *Proceedings of Sixth Message Understanding Conference*, 1995, 55-69.

第二部份

反向異文字音譯相似度評量方法與跨語言資訊檢索

摘要

本文首先根據音譯的方向是否跨不同文字系統，將機器音譯分成「正向同文字」、「正向異文字」、「反向同文字」與「反向異文字」等四種來討論。接著以相似度的比較作為音譯系統的基礎，將語音相似度分為物理聲音、音素、和形素三個層級，並討論計算語音相似度的方式。最後，提出一個以音素相似度為基礎的方法，以中文和英文的音譯為例，進行反向異文字的音譯。實驗結果顯示在音素上的比較，比在形素上的比較來得有效。在一個 1,261 個人名的候選名單中，執行配偶配對實驗，平均排名是 7.80，其中 57.65% 的排名為第一名。

1. 介紹

網際網路隨著電子商務的快速發展，更深入我們的日常生活中，也擴大了世界各國在網際網路上參與的熱度，不同語言所呈現的「內容」(content)在網際網路上傳播。舉凡許多廠商覬覦的中國大陸市場所使用的中文，或是整個歐盟成員國間的各種語言，已經讓網際網路從大部分以英文為主的內容，擴大成為多語言的內容。對網際網路的使用者、或是電腦應用系統(例如搜尋引擎、網路資源蒐集軟體、或是新聞自動摘要系統(Chen and Lin, 2000))來說，因為語言不同所形成的閱讀與處理障礙，也日漸增加。在這種多語的大環境下，機器翻譯(machine translation)與跨語言資訊檢索(cross language information retrieval)等相關自然語言處理系統研究，就極為受到重視。

所謂跨語言資訊檢索(Chen, 1997)就是以一種語言所表達的查詢(query)，去檢索另一種語言所呈現的內容。因為語言上的差異，通常需要將查詢轉換成跟內容一樣的語言。歧義分析(disambiguation)，是查詢翻譯(query translation)一項重要的研究(Bian and Chen, 2000)。根據 1995 年網路使用者，對 Wall Street Journal、Los Angeles Times 和 Washington Post 等新聞語料檢索的統計(Thompson and Dozier, 1997)，分別有 67.8%、83.4%、和 38.8% 的檢索詞含專有名詞。我們知道辭典的覆蓋度，一直是查詢翻譯的重要問題，在專有名詞的翻譯更是挑戰。Chen 等人(1998)，Knight 和 Graehl(1998)，Wan 和 Verspoor(1998)都相繼提出機器音譯(machine transliteration)的方法，來處理這個問題。

音譯可以根據處理的方向，區分成正向音譯(forward transliteration)與反向音譯(backward transliteration)。當一個語言的專有名詞，因為沒有適當或是不容易以意譯來表示時，會採用正向音譯，將其音呈現出來。例如義大利的觀光勝地 Firenze，中文就音譯成「翡冷翠」，此為正向音譯。反過來說，當我們看到一個中文的音譯人名「阿諾史瓦辛格」，如果想要找出原文是 Arnold Schwarzenegger，就是反向音譯。一般來說，使用羅馬字母的拼音文字語言，會保持原詞語字母的拼法，然後以原語言的發音規則，或是自己語言的發音規則來發音。但如果在象形文字與拼音文字語言之間作音譯時，則需要將聲音由原語言盡量用另外一種語言相近的音素來表示，而且要符合目的語言(target language)的語音組合規則。很顯然地，拼音文字與象形文字之間的音譯處理相對來說較為困難，反向音譯比正向音譯更難。正向音譯允許某種程度的失真，所能夠接受的錯誤範圍較大；但反向音譯則不是。反向音譯較不允許錯誤，也就是在找出原文的過程中，必須要相當準確，否則反向音譯的結果應用性就較低。

本文第二節由音譯的正向和反向，以及是否跨文字來分析音譯問題，並介紹過去相關的研究。第三節由相似度的觀念，來執行機器反向音譯的程序。第四節提出一種以音素進行相似度比較的方法。第五節介紹實驗規畫，並對實驗結果進行討論，最後是結論。

2. 音譯分類與相關研究

根據音譯的方向，我們將音譯問題區分為「正向音譯」與「反向音譯」兩種。另外，根據音譯的原始語言與目標語言所採用的字母系統，還可以將音譯區分為「同文字系統間音譯」與「異文字系統間音譯」兩種。以下各小節，就針對這四種組合來介紹相關問題。

2-1 正向同文字間音譯

相同文字系統之間由於共用同一種文字，尤其以羅馬字母為基礎的拼音文字，不同語言在形素(grapheme)與音素(phonomene)間的組合規則雖不一樣，但是一個語言的詞語，要表達成另外一個同文字系統的語言，通常沒有問題。這類型的音譯通常保持原始語言的文字拼法，而目的語言的使用者則以目的語言的發音規則，或是以

原始語言的發音規則來發音。例如 Beethoven 雖然是德國名字，但是在英文的文本中，還是直接使用相同的文字拼法。即使在使用相同拼音字母的語言中，還是可能存在音譯。例如義大利觀光勝地 Firenze(義大利文)，英文則音譯為 Florence。

不同語言使用者在發音時，會採用自己語言的發音規則。例如英語使用者可能會依英語的發音規則來發音，這樣就跟原來德文的發音不同。但大體來說在音素上的發音較為接近，而且越來越多的人會選擇以原始語言來發音，以尊重原始語言。另外，日文中的漢字雖然與中文相通，但由於在發音上差距甚大，所以通常日文漢字翻譯成中文時，表面上與羅馬拼音文字一樣，保持原來日文漢字的寫法，但中文使用者通常會以中文的念法來對日文漢字發音。除非這位使用者學習過日文，才有辦法以正確的日文漢字來發音。

2-2 正向異文字間音譯

在正向異文字間音譯時，主要的工作在於將原始語言的音素，以目的語言的音素來呈現，並配合目的語言的組合規則表示。如果應用在書寫系統上，還要進一步將之前音譯後的結果，選擇目的語言適當的書寫文字，來呈現最後音譯的結果。Wan 與 Verspoor(1998)發展出一套自動將英文專有名詞，正向音譯成中文的系統。在將英文形素轉成音素的過程中，這個系統先將英文字母音節化(syllabification)。拆音節的方法主要有以規則為本(rule-based)，以及範例學習(instance learning)兩種。此系統採用規則為本的方式，但並不是利用上千條的規則來拆解音節，而是利用子音群(consonant cluster)與母音來當成音節的分界來拆解。由於中文為單音節的文字，且多為「子音+母音」的結構，所以系統還要進一步將之前拆解的音節，做進一步的次音節化(sub-syllabification)。將沒有辦法以中文字發音的英文子音群拆開，並加上跟情境相關的母音，以兜成「子音+母音」的音節。在將音素轉成目標語言(在這裡是中文)的文字過程時，Wan 與 Verspoor 的系統，先將拆解完成的音節查表轉換成漢語拼音，接著再查表將漢語拼音最後的中文音譯結果輸出。

2-3 反向同文字間音譯

如前所述，同文字系統間音譯，通常都是保持原來的詞彙組合與型態，所以並不需要做反向的音譯，來找出原始語言的詞彙到底為何。因此，這方面的處理比

較簡單。

2-4 反向異文字間音譯

中文和英文間的轉換，是屬於反向且跨文字系統的音譯，這是本文所要討論的重點。在反向音譯(以後如果沒有特別說明，指的都是異文字間的反向音譯)的研究，有兩種不同的處理方式：一種是直接將音譯後目標語言的詞彙，利用某個模型反推出原始語言的詞彙；另一種是將音譯後目標語言的音譯字，與一串原始語言的候選字相比對，判斷何者可能是原來原始語言所使用的詞彙。

Knight 與 Graehl(1998)利用衍生模型(generative model)，設計一個反向音譯的系統，將音譯後的日文字反向音譯出原來的英文詞彙。當嘗試將英文(原始語言)專有名詞，音譯成日文(目的語言)片假名(katakana)時，衍生模型分成幾個階段處理，包括寫下要音譯的英文詞彙，用英文將該詞彙發音，將英文發音修改成日文可以發的音，將這個日文發音轉成片假名，並寫出片假名。假設我們有一個根據 $P(w)$ 機率分佈來產生英文字(word)的產生器，又假設我們有一個英文發音器。給定一個英文字時，發音器會依據 $P(p|w)$ 的機率來設定該字的發音(pronunciation)。對一個英文發音 p，如果我們想要找出這個發音可能的英文字時，我們就可以尋找看看哪一個英文字 w 可以讓 $P(w|p)$ 這個機率有最大值。根據貝式定理(Bayes' Theorem)，這相當於尋找 $P(w) \cdot P(p|w)$ 。這個系統用到如下五個機率分佈，其中 w 為英文字、e 為英文發音、j 為日文發音、k 為片假名、o 為光學辨識出來的字元：

- (1) $P(w)$ ：產生英文詞彙。
- (2) $P(e|w)$ ：英文詞彙發音。
- (3) $P(j|e)$ ：將英文發音轉成日文發音。
- (4) $P(k|j)$ ：將日文發音轉成片假名。
- (5) $P(o|k)$ ：加入因為光學字元辨識所產生的錯誤。

當 OCR 取得一個片假名字串 o 時，反向音譯使用下面的公式，找出英文字串 w。

$$\arg \max_w P(w) \times P(e|w) \times P(j|e) \times P(k|j) \times P(o|k)$$

Chen 等人(1998)提出一個將英文音譯成中文(目的語言)的音譯字，反向音譯回

英文(原始語言)的模組，並應用於中英跨語言資訊檢索系統。這個系統是將可能的音譯字辨識出來，再進行反向音譯。首先利用漢字羅馬拼音系統(例如 Wade Giles (威翟)，或是漢語拼音(Pinyin))，把可能的音譯字(中文)轉成羅馬字母。接著將這個詞彙與一串可能的專有名詞進行比對，藉此找出可能的原文(英文)。

3. 語音相似度

本篇論文把音譯問題視為相似度的衡量。正向音譯即是在不同語言之間，讓音譯後的結果能夠保持最大的相似度。在反向音譯，如果預先給予一份候選名單，則系統比較音譯字與候選名單上的詞彙，計算兩兩相似度。相似度的比對，可以分成三個層次：形素、音素、和物理聲音。

音譯後的詞彙與原詞彙之間，最直接的比較方式，就是請母語使用者發音，然後以物理上可以測量到的音波來比較。如果從人類可以發出的語音來看，音素集合是固定且有限的，我們可以嘗試在音素的層次來比較。兩個音素的發音位置，或是發音方式越相近，兩個聲音也會越相似。當我們以書寫文字來比較時，就是直接比較形素的相似度。如果書寫文字系統不同，例如中文的方塊字，與英文的羅馬拼音文字，就必須先轉換到相同的字母集合，才能進行比對。

在形素上的比較，Odell 與 Russell 的 Soundex 系統(Knuth, 1973)，是屬於同語言的羅馬拼音字母，利用子音來捕捉詞彙發音的特性。當兩個詞彙的子音位置與發音相似時，表示這兩個詞彙的發音就可能越相似。而 Chen 等人(1998)的研究，可以視為在形素上比較相似度的反向音譯系統。由於所討論的中文音譯字，與原始語言英文的書寫系統不同，他們先將音譯字轉換成羅馬字母，這個動作稱為「羅馬拼音化」(romanization)。他們所採用的標準拼音系統，有威翟與漢語拼音，並加上一些經驗法則修正，來提高系統效能。

由於羅馬拼音系統，主要並不是考慮語音上的相近來設計，例如漢語拼音就用到了 Zh、Q 與 X 等羅馬字母，來表示與字母發音完全無關的漢語語音，所以英文音譯成中文的音譯字，在利用羅馬拼音系統轉換成羅馬拼音字母後，這些羅馬拼音字母，跟原來詞彙的拼音字母，在發音上並不十分相近。

有鑑於在形素層次上做羅馬拼音化時，非常需要一個以形素相近為出發點而設計的羅馬拼音系統。例如在中文和英文這兩種書寫系統完全不同的語言，我們可以設計一個「自動建立羅馬拼音對照表」的系統。這個系統分為兩個階段：第一個階段是訓練，我們從已知的英-中音譯字與原文詞彙的配對中，學習英中音譯字所應該轉換的羅馬拼音字母。例如 Elton 與「愛爾頓」這個配對，先將中文代換成注音符號後，然後分別對兩個字做音節拆解的動作，得到「El·ton」與「ㄞ·ㄦ·ㄩㄨㄣ」，這裡忽略英文重音與中文聲調符號，而 · 為音節間隔符號。接著進一步將英文音節做次音節化後，我們就可以得到英文音節與中文字的音節對應共三組，包括「ㄞ→e」、「ㄦ→l」與「ㄩㄨㄣ→don」。第二個階段實際從事形素相似度衡量，系統根據前一個階段訓練所得到的對照表，將英-中音譯字轉換成英文詞之後，再與候選名單相比較。如前例，「愛爾頓」先轉換成注音符號「ㄞ·ㄦ·ㄩㄨㄣ」，然後查表後得到「e·l·don」。拿掉音節符號後就得到「eldon」，然後再做配偶配對(mate matching)。

表一 列出上述例子的訓練結果。跟其他羅馬拼音系統來比較，我們可以發現：由這個系統所產生的對應，在形素上比其他拼音系統來得更接近實際情形。像是英-中音譯字中的儿，例如貝爾 (Bell) 中的「爾」字，如果採用其他拼音系統來做形素上的比較時，可以發現其他系統完全配對失敗 ($e \neq l$)，只有經過訓練階段所產生對應才能正確配對 ($e=l$)。換句話說，這個系統的對照表是比較有效的，所以能夠在形素層次上的相似度比較，有更好的效能。

表一・訓練結果與羅馬拼音系統

注音符號	威翟	耶魯	漢語拼音	注音符號第二式	「自動產生羅馬拼音對照表」系統的結果
ㄞ	ai	ai	Ai	Ai	e
ㄦ	erh	er	Er	Er	l
ㄩㄨㄣ	tun	dwei	Duan	duan	don

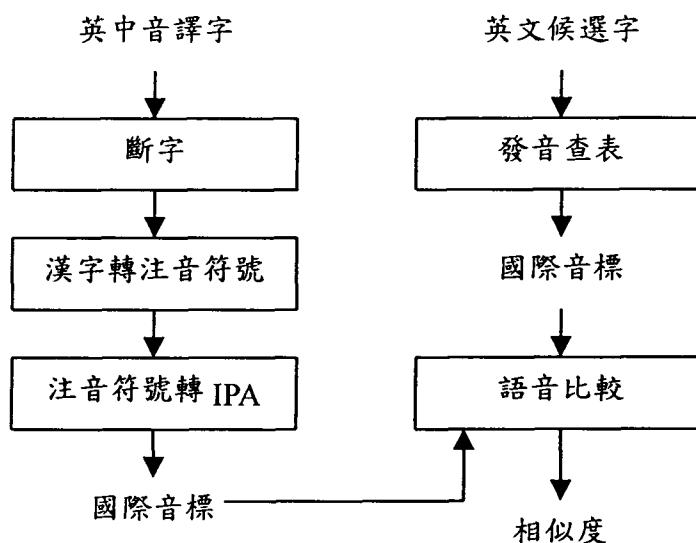
4. 音素相似度評量

考量物理發音在跨語言資訊檢索的實用性，以及形素層次上比對的事前訓練，

4. 音素相似度評量

考量物理發音在跨語言資訊檢索的實用性，以及形素層次上比對的事前訓練，因此音素層次上的相似度比較易顯重要。而衡量兩個詞彙的音素相似度，我們提出一個以國際音標(International Phonetic Alphabet, IPA)為基準的比較，先將兩個詞彙的國際音標列出來，然後比較國際音標的相似度，進而達到反向音譯的目的。圖一顯示音素相似度比較的流程。我們先說明流程圖左邊的部分，也就是英中音譯字處理的部分：

- (1) 斷字：在收到英中音譯字時，第一個步驟即是取出其中的中文字，也就是斷字。例如「亞瑟」這個音譯字，經過斷字後取出「亞」與「瑟」。
- (2) 漢字轉注音符號：將前一個步驟斷出來的漢字，經查表後得到相對應漢字的注音符號。例如「亞」查表後，得到「一ㄚ」，在此我們忽略聲調符號。
- (3) 注音符號轉 IPA：將前一個步驟中的注音符號，經查表二後，得到相對應注音符號的 IPA。表二列出母音和子音與 IPA 對照，這部份參考謝國平(1998)，並略作修正。例如「一」查表後，得到「˥」。一般 IPA 的表示必須配合特殊字體，才能顯現，CMU pronunciation dictionary 0.6 版(簡稱 CMU dict)(<ftp://ftp.cs.cmu.edu/project/fpdata/dict/>)採用 ASCII 來表示，附錄列出 CMU dict 符號和 IPA 符號對照。例如，「˥」對應「IY」。



圖一・音素比對

表二・注音符號與 IPA 對照表

(a) 子音部份

注音符號	ㄅ	ㄉ	ㄇ	ㄔ	ㄅ	ㄔ	ㄉ	ㄉ	ㄌ	ㄉ	ㄉ	ㄍ	ㄉ	ㄏ	ㄉ	ㄉ	ㄉ	ㄉ
IPA	ㄤ	ㄤ	ㄤ	ㄤ	-	ㄤ	ㄤ	ㄤ	ㄭ	ㄤ	ㄤ	ㄤ	ㄤ	ㄤ	ㄤ	ㄤ	ㄤ	ㄤ
注音符號	ㄝ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ
IPA	ㄢ	ㄢ	ㄢ	ㄢ	-	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ

(b) 母音部份

注音符號	一	ㄨ	ㄩ	ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ	ㄞ	ㄞ	ㄞ	ㄞ	ㄞ	ㄞ	ㄞ
IPA	ㄢ	-	ㄢ	ㄢ	。	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ	ㄢ

對候選名單中的音譯字，處理的方式也是類似。每一個英文詞彙，我們查表(CMU dict 0.6)，以得到該詞彙的發音。例如「Arthur」的發音，經查表後得到「AA R TH ER」(忽略重音標示)。

「語音比較」是整個流程最重要的部分，當我們拿到兩串 IPA 時，如何比較這兩個 IPA 字串的相似度呢？首先來看以下三個定義，由字母對齊相似度、到字串對齊相似度，最後到字串相似度。

定義 1：字母對齊相似度

假設 S_1 與 S_2 這兩個字串的字母集合為 Σ ，而 Σ' 表示 Σ 加上「_」(_表示空白字元)，給予 Σ' 中的兩個字元 x 與 y ， $s(x, y)$ 表示 x 與 y 對齊後，所得到的分數，稱為字母對齊相似度。

定義 2：字串對齊相似度

假設 A 為字串 S_1 與 S_2 的某一種對齊方式(alignment)， S_1' 與 S_2' 為插入空白後的字串。如果 S_1' 與 S_2' 的長度為 l ，則對齊方式 A 的分數如下：

$$\sum_{i=1}^l s(S_1'(i), S_2'(i)) .$$

我們以一個例子說明上述定義。如前例，「亞瑟」經查表後，得到的發音是「IY AA S r」，而 Arthur 的發音為「AA R TH ER」，所以此時的 $\Sigma = \{AA, ER, IY, R, r, S, TH\}$ ，而音素間彼此的分數，以下面的對稱矩陣表示：

S	AA	ER	IY	R	r	S	TH	=
AA	5	0	0	-10	0	-10	-10	-5
ER	0	5	0	-10	8	-10	-10	-5
IY	0	0	5	-10	0	-10	-10	-5
R	-10	-10	-10	10	-10	-10	-10	-5
r	0	8	0	-10	5	-10	-10	-5
S	-10	-10	-10	-10	-10	10	8	-5
TH	-10	-10	-10	-10	-10	8	10	-5
-	-5	-5	-5	-5	-5	-5	-5	-5

下面這個對齊方式：

亞瑟	IY	AA	_	S	r
Arthur	_	AA	R	TH	ER

依定義 2 所給定的字串對齊相似度分數為： $-5 + 5 + -5 + 8 + 8 = 11$ 。

然後我們來定義字串相似度。

定義 3：字串相似度

給定一個字母集合 Σ' ，和成對的分數矩陣。字串 S_1 與 S_2 的相似度，定義成 S_1

與 S_2 的最佳對齊方式 A 的值，也就是最大的字串對齊相似度值。

相似度跟相關的最佳對齊方式，可以用 dynamic programming 的方式來找出。Gusfield (1997) 曾定義基底條件(base condition)為

$$(i, 0) = \sum_{1 \leq k \leq i} s(S_1(k), \underline{\quad})$$

$$(0, j) = \sum_{1 \leq k \leq j} s(\underline{\quad}, S_2(k))$$

一般的 recurrence 式可以寫成：

$$\begin{aligned} V(i, j) &= \max[V(i-1, j-1) + s(S_1(i), S_2(j)), \\ &\quad V(i-1, j) + s(S_1(i), \underline{\quad}), \\ &\quad V(i, j-1) + s(\underline{\quad}, S_2(j))] \end{aligned}$$

$0 \leq i \leq \text{length}(S_1)$ ， $0 \leq j \leq \text{length}(S_2)$ ， $V(0, 0) = 0$ 。其中 $V(i, j)$ 為 $S_1[1..i]$ 與 $S_2[1..j]$ 這兩個前字串(prefix)，最佳對齊方式的值。假設 S_1 與 S_2 的長度各為 n 與 m，則最佳對齊方式的值就是 $V(n, m)$ 。如果利用 dynamic programming 的方式來求，這個值可以在 $O(nm)$ 的時間內算出來。

在我們的反向異文字音譯語音相似度評量中， Σ' 為 63 個 IPA 音標符號(含空白)，其中英文有 39 個，中文除了共用的之外，另外還有 24 個中文所獨用的符號，所以整個分數矩陣的大小為 63×63 。我們對分數矩陣中的分數指定方式如下：

- (1) 原則上，IPA 匹配(match)給 10 分，不匹配(mismatch)扣 10 分。但若匹配的為母音，則只給 5 分，而母音不匹配不扣分。這裡我們希望利用母音來捕捉音節的對齊，所以母音不對齊不扣分。但由於母音在不同語言間的匹配，意義較不顯著，因此相同的母音只給 5 分。
- (2) 與空白字元()對齊，可以看做 insertion 或是 deletion。由於不匹配可以看成是一個 insertion 加上一個 deletion。例如 abcdfgh 和 abcdigh，其中 f 與 i 未匹配，當要對齊時，可以採用如下的方式：

abcd <u>f</u> <u>g</u> h
abcd <u>i</u> gh

所以未匹配要扣的分數，跟兩個字元對上空白，亦即做一次 insertion 和一次 deletion 要相同，這樣才沒有偏好。因此，為了公平起見，我們讓 insertion 或是 deletion 的扣分，等於不相同配對的一半，也就是 $10/2=5$ 分。另外，關於空白對空白分數還是設-5(參考分數矩陣範例)，原因是兩個字串 ab 與 ac 在對齊時，如果 a 對 a 匹配給 10 分，不匹配扣 10 分，則 ab 和 ac 字串對齊相似度為： $10 + (-10) = 0$ 分。如果加上空白，再進行對齊，如 a_b 和 a_c，這樣的分數為 $10 + (-5) + (-10) = -5$ 分。也就是在對列時，同時加上空白是沒有用的，只是會把分數拉低，所以空白對空白是-5 分。

- (3) 其他根據發音位置與發音方式的相近，中英文在音譯上的習慣、中英文各自的發音特性、將某些音標之間的配對分數設為 8 分，如表三所列。

表三・其他音標之配對

理由	例子
中文不分清濁	P 與 B、D 與 T、F 與 V、G 與 K、S 與 Z
發音方式與位置相近	B 與 Ph、K 與 Kh、D 與 Th、P 與 Ph
發音位置相近	L 與 R、DH 與 Th
發音方式相近	CH 與 Tch、CH 與 TSch、H 與 Th、G 與 Tc、JH 與 Tc，L 與 R、M 與 ANG、N 與 AN、N 與 AHN、N 與 ANG、NG 與 ANG、NG 與 AN、NG 與 AHNG、S 與 Sc、S 與 c、S 與 TH、S 與 TS、Z 與 Sc、Z 與 TS、Z 與 TSc
音譯習慣以及跨語言所造成的音標空缺	K 與 Tc、L 與 e、R 與 e、TH 與 Th、ZH 與 Tch、ER 與 r、ER 與 L、ER 與 e、UW 與 V、JH 與 TSc、G 與 Tch
中文不分長短母音	IH 與 IY、UW 與 W
半母音與母音	IY 與 Y

5. 實驗結果

我們採用配偶配對(mate matching)的方法，來評估語音的相似度。方法如下所述：給予已知的原始語言詞彙 o_i ，與音譯後的目的語言詞彙 t_i 的配對清單集合， $\{(o_1, t_1), (o_2, t_2), \dots, (o_n, t_n)\}$ 。當讀入音譯後的目的語言詞彙 t_k 時，測量語音相似度的系統，對整個清單中的每個原始語言詞彙作相似性比對，並計算每一對相似度的分數。之後再看看正確的原始語言詞彙 i_k ，落在依分數高低排序的配對結果中的名次。名次越高，表示語音相似度比較越準確。

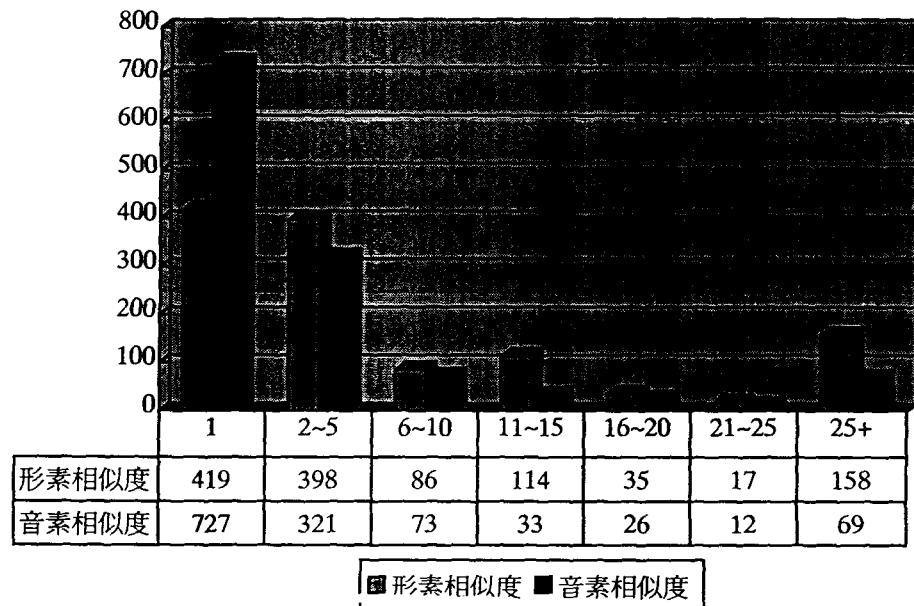
集合中的每一個目的語言詞彙，都設定名次後，我們就可以取這些名次平均值，作為一個語音相似度比較方法的評量標準。另外一個標準化的指標，是將這個平均的名次除以整個配對集合的個數 n 。這表示當一個音譯後的詞彙輸入後，系統需要提出多少個候選詞彙，才會包括到正確答案。

根據這項評量語音相似度的方法，我們採用與 Chen 等人(1998)實驗相同的候選名單，共 1574 個人名。扣除無法找到發音的人名 313 個，合格的候選名單共 1,261 個人名。表三列出音素相似度和形素相似度的結果。本文所採用的音素相似度平均排名為 7.80，比 Chen 等人所採用的形素相似度平均排名 9.69，表現還要好。

表三・評估結果

	音素相似度	形素相似度
平均排名	7.80	9.69

圖二進一步列出排名分佈情況。



圖二・排名分佈

從以上結果我們可以清楚發現，在語音相似度的比較上，音素層次比形素層次表現的好。不僅平均排名上，音素相似度比形素相似度的平均排名好。如果進一步觀察名次的分佈，音素相似度有 57.65% 的結果都是最相似的，也就是正確答案。反觀形素相似度，只有 33.28%。

進一步觀察實驗結果中匹配失敗的配對，我們可以將失敗的原因歸類如下：

- (1) 約定俗成但聲音並不相近的音譯：由於中文與英文是兩種不論書寫與發音都是相差甚遠的語言系統，因此不論對專業譯者或是一般作家，音譯並不是一件容易的事。但是一些已經約定俗成的翻法，例如 Bach (巴哈)、Caesar (凱薩)、John (約翰) 等音譯，在音素上卻不十分相近，所以在音素層次上比對的效果不好並不令人意外。
- (2) 英文非重音節的子音被忽略：英文中不在重音節的子音（通常是靠近結尾的部分），由於在中文使用者的語音知覺上並不明顯，所以音譯時

經常就直接省略不翻，例如：Briand（白里安）中結尾的 d 在音譯成中文時就沒有被翻出來。

- (3) 插入的母音造成混淆：由於中文字為單音節且多為子音加母音 (CV) 的結構，當英文字要轉換成中文時，勢必要在適當的地方插入母音才能構成 CV 結構。例如 Paul (保羅) 與 Young (楊格) 中結尾的亡，原來是一個音節的字，到中文變成了兩個音節，也造成在音素層次上比對的困擾。
- (4) 不是追求聲音接近的翻法：在一些特別的場合，特別是書寫的文本，翻譯者並不純粹追求聲音上相近的音譯方式，而可能為了與中文命名法相近（像 Gertrude，葛麗露）、或是為了簡潔（像 Gillian，姬兒），或是一味因襲傳統音譯方式卻忽略聲音上是否相近（像 Patricia，珮格麗特），這些在在都造成音素上的比對並不成功。

雖然如此，這些配對失敗的中文音譯字，比較音素相似度方法所找出來的最相似字，仍然反應音素上的相近。例如「保羅」雖然跟正確答案 Paul 並不十分相近，但系統比對得到的 Polo 在音素上其實是比 Paul 來得比「保羅」更接近；又或「姬兒」雖然無法對到 Gillian，但是這個方法找到的 Jill 也是更接近「姬兒」，讓我們對這個方式在音素上比較相似度的能力深具信心。

6. 結論與未來的研究

機器音譯研究中，最具挑戰性，也最具實用價值的問題，就是在跨文字系統的反向翻譯。這種反向音譯在跨語言資訊檢索，或是機器翻譯時，都是一個不能忽略的問題。利用語音相似度的原理，從事反向音譯時，如果相似度的比較層次分為物理聲音、音素、與形素，而物理聲音無法進行時，我們發現音素層次上的比較，比之前在形素層次上的比較來得準確。

根據 Knight 與 Grahel(1998)對音譯系統的評量標準，這個以音素相似度來進行反向音譯作業的方式，相當接近人類在判斷音譯字是否相近，因為音素比較接近實際的聲音，而形素通常差距較大。而這個方法在應用到其他語言配對時，只要給定不同的配分矩陣就可以。最後，這個方法可以根據分數的高低，來提供一串可能的

清單。所以，這個方法不管在理論與實際應用上都是深具價值。

機器音譯並不完全只是在語音上追求相等，有的專有名詞翻譯，因為歷史因素或是語言使用者的習慣，採取意譯而不是音譯。例如國家名稱 the United States，在大部分的中文文件中都是意譯成「美國」，而不採取音譯。同時，並不是所有專有名詞都採取意譯，例如 British Virgin Island 中的 Virgin，在中文音譯成「維京」，而不是採取意譯，Island 則直接翻譯成島。因此，在反向異文字音譯處理之前，先將地名送進雙語字典。如果已有現存的翻譯，就直接採用此翻譯。如果沒有，再檢查有沒有關鍵詞。關鍵詞查雙語辭典，其餘部份才經反向異文字音譯處理。

參考文獻

- Bian, Guo-Wei and Chen, Hsin-Hsi (2000) "Cross Language Information Access to Multilingual Collections on the Internet," *Journal of American Society for Information Science*, 51(3), 2000, pp. 281-296.
- Chen, Hsin-Hsi (1997) "Cross Language Information Retrieval," *Proceedings of ROCLING Workshop on ED/MT/IR*, Academic Sinica, Taipei, June 2, 1997, pp. 4-1~4-27.
- Chen, Hsin-Hsi; Huang, Sheng-Jie; Ding, Yung-Wei and Tsai, Shih-Chung Tsai (1998) "Proper Name Translation in Cross-Language Information Retrieval," *Proceedings of 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, Montreal, Quebec, Canada, August 10-14 1998, pp. 232-236.
- Chen, Hsin-Hsi and Lin, Chuan-Jie (2000) "A Multilingual News Summarizer," *Proceedings of 18th International Conference on Computational Linguistics*, July 31-August 4 2000, University of Saarlandes.
- Gusfield, Dan (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, 1997, Cambridge University Press.
- Knight, Kevin and Graehl, Jonathan (1998) "Machine Transliteration," *Computational Linguistics*, Vol. 24, No. 4, 1998, pp. 599-612.
- Knuth, Donald E. (1973) *The Art of Computer Programming, Volume 3, Sorting and*

- Searching*, Addison-Wesley, Reading, Mass, 1973, pp. 391-392.
- Thompson, P. and Dozier, C. (1997) "Name Searching and Information Retrieval," *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island, 1997.
- Wan, Stephen and Verspoor, Cornelia Maria (1998) "Automatic English-Chinese Name Transliteration for Development of Multilingual Resources," *Proceedings of 17th COLING and 36th ACL*, 1998, pp. 1352-1356.
- 謝國平(1998), 語言學概論, 三民書局, 台北, 1998 年 10 月。

第三部份：
以學習語音相似度為基礎之反向機器音譯

摘要

在許多跨語言的應用程式中，我們需要把目的語言轉換回原始語言。在本文中，我們運用以相似度為基礎的架構以實現反向音譯的任務，而且提供一個學習演算法則從音譯字語料庫中自動獲取語音相似度。實驗顯示學習演算法則會很快地聚集同樣的結果，而且用以獲取語音相似度的方法顯然地勝過先前在 1574 對英文名詞與音譯中文名詞的語料庫中，預先設定語音相似度或字型相似度的作法。學習演算法則並不假設任何潛在的語音體系結構或規則(phonological structures)，只要有適當的訓練語料庫(training corpus)和發音字典(pronouncing dictionary)，它還可以擴展至其他語言。

1 介紹

隨著網路上多元語言文件的快速增加，克服語言障礙的需求也愈形迫切。跨語言資訊檢索(CLIR)致力於以一種語言的查詢(query)去檢索另一種語言的文件，而專有名詞處理在查詢翻譯當中扮演著很重要的角色(Bian and Chen, 2000; Oard, 1999)。研究顯示，使用者在新聞搜尋引擎上所使用的查詢(query)有絕大部分包含專有名詞。然而，大多數的專有名詞來自音譯而非翻譯，以字典為基礎來處理音譯的方式存在著覆蓋度有限和更新頻率不高的問題，因此任何健全的 CLIR 系統必須要適當地處理此問題以達到更好的成果。

音譯可按其翻譯的方向來分類。在一個成對的組合(s, t)當中， s 是來源語言的原始專有名詞， t 是目標語言中的音譯字，正向音譯的程序是按照音節把 s 轉換成 t ，反向音譯是依照所給的 t 正確地搜尋或產生 s 。表一顯示出兩種類型的例子。

Direction	From	To
Forward	Harry Potter (English)	Ha1-li4-bo1-te4 (Chinese) Harri Pottaa (Japanese)
Backward	Huo4-ge2-hua2-zi1 (Chinese) Hoguwaatsu (Japanese)	Hogwarts (English)

表一・正向與反向機器翻譯

在先前的研究工作中，不斷有學者在雙向翻譯與不同語言字組(language pairs)的領域探索：從英文到中文的正向翻譯(Wan and Verspoor, 1998)，以及從日文到英文(Knight and Graehl, 1998)，中文到英文(Chen et al., 1998; Lin and Chen, 2000)，從阿拉伯文到英文(Stalls and Knight, 1998)的反向翻譯。反向音譯比正向音譯更具挑戰性。正向音譯能對照表格查詢以完成配對的關係(mapping relationship)，反向音譯則要針對正向翻譯時所產生之多種可能的歧義加以分析，並且儘可能的判斷出其最初的字組。本文將集中焦點在反向音譯上。

我們應用以相似度為基礎的架構來建立達成反向音譯的模型。當我們在作正向音譯時，主要目標在根據音節的相似度，儘可能使音譯字接近原始字組。因此，我們基於相同的想法來建立反向音譯的模型。以相似度為本的方法在形素(Chen et al., 1998)及音素(Lin and Chen, 2000)層級上有不錯的效果。然而，在先前的研究中相似度是被特別指派的(ad hoc assigned)。在本文中，我們將發展一種學習運算法，從訓練語料庫(training corpus)中自動獲取語音資訊(phonicetic knowledge)。藉由此學習運

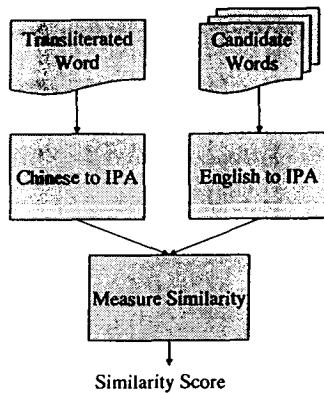
算法則，我們就不用再以人工去指派二種語言間的語音相似度，並且希望這些相似度能改進音譯的性能。

本文大綱如下：在第 2 節中，我們將描述以相似度為本的架構而且定義出二個字之間的相似度；學習運算法則和訓練語料庫的準備(*training corpus preparation*)包含在第 3 節中；實驗設計與結果是在第 4 節中；我們會在第 5 節中討論實驗結果並在第 6 節中提出結論。

2 以相似度為本的架構

在以相似度為基礎的架構中，我們將一個音譯字 t 和候選名單上的字作比對，名單上的字與 t 有最高相似度的會被選為答案。候選名單可由人工或由命名實體擷取系統(Fung and Yee, 1998)來收集。在 CLIR 應用程式裡，查詢中的外語文字一旦被確認之後，我們就會對這些字進行配偶配對(mate-matching)的程序—這也是查詢翻譯(query translation)的步驟之一。程序流程詳見圖一。換句話說，反向音譯可被歸納為相似度的測量。

我們將語音相似度分為物理聲音、音素、和形素三個層級，例如 Soundex (Knuth, 1973) 是在測量形素層級的相似度。在這裡我們選擇音素，因為要產生並且比較物理聲音較為困難，而且在形素層級比對比在音素層次比對更佳(Lin and Chen, 2000)。我們提出一個以國際音標(International Phonetic Alphabet, IPA)為基準的比較，國際 IPA 是一套能標註任何語言語音的拼音系統。在下面的章節中，我們首先描述該如何從中文與英文當中列出國際音標，然後再比較兩者間國際音標的相似度。



圖一・以相似度為本的反向機器音譯流程

2.1 形素轉音素的變換程序

在中文的音譯字中，每個字元會藉由查表轉換成漢語拼音，至於音調則會被省略。漢語拼音字串會被拆解為兩個部分：開端的子音和剩餘的母音，接著再查表將漢語拼音轉換成國際音標(Hieronyms, 1997)。

將英文字轉換成國際音標的過程需要考慮更多的因素。首先，如果一個字存在於一部發音字典(Cmudict, 1995)，字典中提供的發音會被轉換到國際音標 IPA。但如果這不字典並未收錄這個字，我們應用一種語音合成系統(speech synthesis system) MBRDICO (Pagel et al., 1998) 來產生發音。雖然專有名詞的語音合成仍然是一個充滿挑戰與爭議性的研究議題 (Llitjos and Black, 2001),但是與其予以省略我們寧可保有這些字典未收錄的字，並且研究在反向音譯工作中，這些未盡完善的語音合成字所形成的結果與影響。字母—音素的轉換系統(letter-to-phoneme)會輸出 SAMPA (Wells, 1997) 的發音(the pronunciation in SAMPA), 再依次轉換成國際音標 IPA，而在轉換期間的資訊將不被使用。

2.2 相似度評量方法

“距離編輯” the edit distance (Levenshtein, 1966) 被廣泛地使用在二個字串間的相關距離測量。在此的距離被定義為轉換一個字串到另一個字串時，所需要插入、

刪除和替換字元的最小數目。下列的相似度定義與“距離編輯”一樣，在插入，刪除和替換字上有很大的變化，然而在某些應用上，相似度定義比“距離編輯”更適當，舉例來說，搜尋高相似度的子字串。（Gusfield，1997）首先，我們定義出二個字串的組合來測量兩者的相似度。

定義 1

假設 S_1 與 S_2 這兩個字串的字母集合為 Σ ， Σ' ($\Sigma' = \{\Sigma, '_'\}$) 表示 Σ 加上「 $_$ 」（ $_$ 表示空白字元）， S_1' 與 S_2' 為插入空白後使其等長度的兩個字串。兩個等長字串間的字元相互對齊，而 A 就代表了這兩個對齊字串。

二種對齊方式的相似度得分定義如下：

定義 2

$s(a, b)$ 函數能在字母集合 Σ 當中，測量出字元 a 與字元 b 之間的相似度。假設 A 為字串 S_1 與 S_2 的某一種對齊方式(alignment)， S_1' 與 S_2' 為插入空白後的字串。如果 S_1' 與 S_2' 的長度為 l ，則對齊方式 A 的分數如下：

$$Score = \sum_{i=1}^l s(S_1'(i), S_2'(i)) \quad (1)$$

其中 $S'(i)$ 表示字串 S' 中的第 i 個字元。

舉一個英文名字 Hugo，及其中文音譯名 Yu3-guo3 為例來說明上述的定義。經過形素轉音素的變換程序，得到的發音是(v k uo, h j u g oU)¹，所以此時的 $\Sigma' = \{h, j, u, v, g, k, oU, uo, _\}$ ，二個字串間有許多對齊方法，表二顯示出其中的兩種。

圖二的得分矩陣可表示相似度函數 $s(a, b)$ ，矩陣的內容可由人工指定或自動學

習取得。得分範圍在 10 與 -10 之間，得分愈高，二個音素愈相近。

	Grapheme	Phoneme
Λ_1	Hugo	h j u g oU
	Yu-guo	_ _ v k uo
Λ_2	Hugo	h j _ u g oU
	Yu-guo	_ v k _ uo _

表二・音素字串(h j u g ou, v k uo) 的兩種可能對齊方式

s(a,b)	h	j	u	v	g	k	^o U	uo	-
h	10	0	-8	0	0	-9	0	-4	-10
j	0	10	-1	0	0	-1	0	-1	3
u	0	0	10	3	0	-4	0	-2	-10
v	0	0	-6	9	0	-6	0	-5	-10
g	0	0	-	0	10	10	0	-7	-10
				10					
k	0	0	-	-	0	10	0	-	-10
					10	1		10	
oU	0	0	2	4	0	-4	10	10	-10
uo	0	0	0	0	0	0	0	10	-10
-	-	-	-	-	-	-	-	-	-
-	10	10	10	5	10	10	10	10	10

圖二・相似度得分矩陣

藉由方程式 1 和圖二的得分矩陣，我們就能計算出表二中二種對齊方式的相似度得分：

$$Score_{\Lambda_1} = s(h, _) + s(j, _) + s(u, v) + s(g, k) + s(oU, uo) = 16$$

$$Score_{\Lambda_2} = s(h, _) + s(j, v) + s(_, k) + s(u, _) + s(g, uo) + s(oU, _) = -47$$

最後，二個字串的相似度得分定義如下：

定義 3

給定一個字母集合 Σ' ，和相似度分數矩陣 M 。兩字串間的相似度分數即為最佳對齊方式 A 的值，也就是最大的字串對齊相似度值。

相似度跟相關的最佳對齊方式，可以用 dynamic programming 計算得出 (Masek, 1980)。Set T 是一個 $n+1$ 乘以 $m+1$ 的表格，其中 n 是字串 S_1 的長度， m 是字串 S_2 的長度。藉由循列填充表格 T ，我們就能獲得最佳的對準方式和 S_1 、 S_2 的相似度得分。下列為其基本條件(base condition)定義：

$$\begin{aligned} T(i, 0) &= \sum_{1 \leq k \leq i} s(S_1(k), _) \\ T(0, j) &= \sum_{1 \leq k \leq j} s(_, S_2(k)) \end{aligned} \quad (2)$$

一般的 recurrence 式可以寫成：

$$T(i, j) = \max \left(\begin{array}{l} T(i-1, j-1) + s(S_1(i), S_2(j)) \\ T(i-1, j) + s(S_1(i), _) \\ T(i, j-1) + s(_, S_2(j)) \end{array} \right) \quad (3)$$

where $1 \leq i \leq n$, $1 \leq j \leq m$.

如果我們用”編輯距離”(edit distance)的語言來說，recurrence 公式會試著比較替換、刪除和插入的差別，並且選擇差別最小的那一個，在這裡也就是最大的相似度值。表格會在 $O(nm)$ 時間內完成，而 $T(n, m)$ 將是字串 S_1 和 S_2 之間，最佳對齊方式的相似度得分。我們可以藉由紀錄 recurrence 式中所作的選擇獲得最理想的對齊方式。舉例來說，從圖二的得分矩陣來看，二個音素字串 S_1 (j h u g oU) 和 S_2 (v k uo) 間的最佳對齊方式是表二中的 Λ_1 。

3 語音相似度的學習

得分矩陣的設計在區分何種對齊方式較佳時扮演著重要的角色(Gusfield, 1997)。其得分反映出當我們在作反向音譯時是如何感知音素的。發展學習演算法則的動機是要免除在矩陣中用人工指定得分的過程，並且獲取人工不易察覺的差異。

在以或然率為本的架構下(probability-based framework)，距離編輯學習(Edit distance learning)一直是被研究的對象(Ristad and Yianilos, 1998)。當一個或然率模型(generative probabilistic model)能產生並且學習語音相似度的時候，一個較好的學習演算法則應該是能直接在上述相似度為本之架構上運作並且還要能區分音素字串。在這一個部分中，我們首先描述該如何準備一個訓練語料庫(a training corpus)，接著再論及學習演算法則。

3.1 訓練階段的準備

為了要訓練出有識別力的分類工具，我們必須準備具有正面例子(positive examples)和負面例子(negative examples)的訓練資料。然而，包含原始字組和音譯字組的語料庫只是正面例子(positive examples)。幸好，我們能藉由原始字組與音譯字組的錯誤配對來產生負面例子，而不用去收集更多的資料。

在一個包含 n 對音素字串的語料庫中， e_i 是原始語言英文， c_i 是音譯後的中文字， $1 \leq i \leq n$ 。每個 e_i 都會有一個與其發音最相近的音譯字(即 c_i)，還有其他 $n-1$ 個發音相異的音譯字(即 e_j)，其中 $1 \leq j \leq n$ ， $j \neq i$ 。配對的相似度得分設定如下：

$$Score_{(e_i, c_j)} = \begin{cases} 10 * p & i = j \\ -10 * p & i \neq j \end{cases} \quad (4)$$

where $p = \max(\text{length}(e_i), \text{length}(c_j))$.

因此，一個包含 n 組成對字組的語料庫能產生 n 個正面的例子，和 $n(n-1)$

個負面例子。為了要說明正面和負面例子數目上的差異，我們複製正面的例子，使其總數達到 $2n^2$ 個。

3.2 學習運算法則

我們把每個訓練例子(training sample)視為一個一次方程式。方程式一可重寫如下：

$$y = \sum_{i=1, j=1}^m w_{i,j} x_{i,j} \quad (5)$$

m 是音素集合(phoneme sets)的大小， $w_{i,j}$ 代表在得分矩陣中的第 i 列、第 j 欄， $x_{i,j}$ 是由二部分組成的，顯示在對齊方式中 $w_{i,j}$ 代表的數值，而 y 則是相似度得分。舉例來說，圖二是個九乘以九的得分矩陣，因此 $m = 9$ 。 $w_{i,j}$ 能對應到圖二的每個儲存格，在這裡 $1 \leq i, j \leq 9$ 。在表二的對齊方式 Λ_1 中 $x_{1,9}, x_{2,9}, x_{3,9}, x_{5,9}, x_{7,8}$ 是 1，其他的 $x_{i,j}$ 皆是零的。此外，在語料庫的方程式系統也亦可以矩陣形式呈現：

$$\begin{pmatrix} x_{1,1}^1 & \dots & x_{m,m}^1 \\ x_{1,1}^2 & \dots & x_{m,m}^2 \\ \vdots & \ddots & \vdots \\ x_{1,1}^R & \dots & x_{m,m}^R \end{pmatrix} \times \begin{pmatrix} w_{1,1} \\ w_{1,2} \\ \vdots \\ w_{m,m} \end{pmatrix} = \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^R \end{pmatrix} \quad (6)$$

or

$$\mathbf{X}\mathbf{w} = \mathbf{y}$$

其中上標的 i 代表了在語料庫中第 i 個成對例子，在這裡 $1 \leq i \leq R$ ， R 是訓練語料庫中成對例子的數目。

我們所選擇的最佳化準則是 sum-of-squared error，即 $\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ 。因此，學習任務的目標將是獲得方程式 6 的 \mathbf{w} ，使 sum-of-squared error 最小化。傳統的解決方案是利用 \mathbf{X} 的假相反數 pseudo inverse(也就是 $\mathbf{X}^\dagger \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$)，來獲得使 sum-of-squared error 最小化的 \mathbf{w} ，在這裡 $\mathbf{w} = \mathbf{X}^\dagger \mathbf{y}$ 。然而，當 \mathbf{X} 是一個龐大的矩陣時，假相反數的數值相當可觀，而且當 $\mathbf{X}'\mathbf{X}$ 是單數時，根本無法計算出假相反數。因此，我們採用 Widrow-Hoff procedure(Duda et al., 2001)來避免這些問題。Widrow-Hoff 或 Least-Mean-Squared (LMS)會在下降趨勢中(in gradient descent

fashion) 將錯誤減到最少。圖三列出了學習演算法則的偽程式碼，下標的 k 代表 X 矩陣的中 的第 k 列； i 為重複的數目； $w(i)$ 、 $\eta(i)$ 、 和 $\delta(i)$ 是 i 的函數； η 則是學習率(learning rate)。

```

Initialise  $w(0)$ ,  $y$ ,  $\eta(0)$ ,  $i$ 
Do
     $i \leftarrow i + 1$ 
     $k \leftarrow i \bmod R$ 
     $\eta(i) = \eta(0) / R$ 
    For the  $k^{\text{th}}$  sample  $(s_k, t_k)$ 
         $X_k \leftarrow$  the optimal alignments given  $w(i-1)$ 
         $\delta(i) \leftarrow y_k - w(i-1)'X_k$ 
         $w(i) \leftarrow w(i-1) + \eta(i)\delta(i)X_k' + \alpha\delta(i-1)X_k'$ 
    While  $w$  is not overfitting

```

圖三。以 Windrow-Hoff rule 為本加以修改之學習演算法則的偽程式碼

函數 $w(i)$ 會被反覆地更新直到習得的 w 超越確認組的容量。學習率會隨著重複的數目減少以確保 w 會趨近滿足 $X'(Xw - y) = 0$ 的向量。除了 Windrow— Hoff 規則的基本骨架外，我們還應用了線上學習技術 (Biehl and Riegler, 1994)來加速聚集。我們一遇到新的訓練例子就立刻更新函數 $w(i)$ ，而不是在累積了所有錯誤訓練例子之後才作。另一個加速的技術是抑制波動的動量(momentum)， α 代表動力係數， $\eta(0)$ 會依經驗設定在 5×10^{-6} ， α 則是 0.8，而 $w(0)$ 則依下列所示：

$$w_{i,j}(0) = \begin{cases} 10 & \text{if } i = j \\ -10 & \text{if } i \text{ or } j \text{ is ' ' } \\ 0 & \text{otherwise} \end{cases}$$

我們假定音素是自我相似的，並且防止音素和空白字元相配。其他的語音相似度被設定為零，這是沒有任何預設立場的合理起始值。

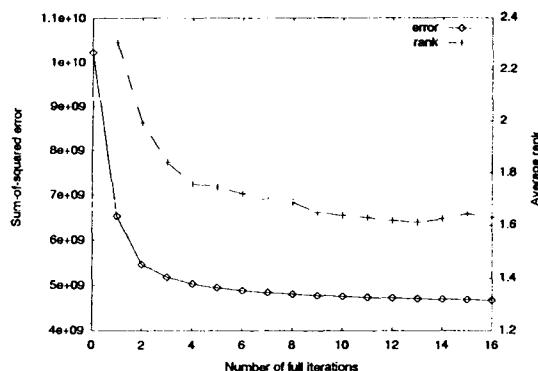
為了要避免 overfit 語料庫而失去其普遍性，我們會在訓練集合的完整重複之後對確認組中所習得的 w 做評估。倘若其成果並沒有在連續的三個重複中得到改善，我們將會停止下降程序(gradient descent procedure)，並且將 w 返回最佳性能的數值。

4 實驗結果

為了把學習語音相似度與先前的研究作比較(Chen et al., 1998) and (Lin and Chen, 2000)，我們採用相同的訓練語料庫和評估矩陣(evaluation metrics)。語料庫由 1547 對英文專有名詞及其相對應的中文音譯字所組成，其中有 313 字並未收錄於發音字典。總共需要 97 個音素用以代表這些專有名詞，這當中有 59 個 和 51 個音素分別用於中文和英文的專有名詞。

在評估反向音譯系統的成效時，我們對語料庫作了 10 層的交叉確認(ten-fold cross-validation)。在每個層級中，語料庫被區分成三組(set): 8/10 是訓練組(training set)、1/10 是確認組(validation set)、剩餘的 1/10 是測試組(test set)。訓練組產生獲取語音相似度的正面及反面例子；確認組用來避免學習者(learner) overfit 訓練組(如前一章節所述)；學習者尚未見到的測試組用以評估系統成效。這十個層級所平均出的 w 將是最後的結果。成績矩陣即為在先前論文中所提出的平均排名，這裡的排名代表在候選名單中正確原始字組的位置，候選名單上的字組依照計算出的相似度分數遞減排列。平均的排名愈小，成績表現也愈高。

圖四顯示出一個層級的學習曲線。為簡明起見，在此我們省略了其他九個亦有相似趨勢的層級。squared error 的總數 (左邊的 Y 軸) 在頭幾次重複(iterations)時，快速地減少，展現我們的學習演算法則會快速地趨向同一結果。當學習演算法則得到較佳的語音相似度時，確認組的平均排名 (右邊的 Y 軸) 也會跟著提昇。平均排名從第 14 次重複開始增加，因此學習演算法則在第 16 次重複時停止。系統將回傳第 13 次重複時所得到的得分矩陣。



圖四・某層級的學習曲線顯示出學習演算法則的快速聚集成效

以學習語音相似度為本的方法與先前 Chen98 (Chen et al., 1998)與 Lin00 (Lin and Chen, 2000)的研究成果相比較，表三列出比較情形。實驗結果顯示，以學習語音相似度為基礎的反向音譯明顯優於之前以預先定義相似度為本的方法。

	Chen98	Lin00	Learning
Average Rank	12.11	7.8	2.04
Similarity	Grapheme	Phoneme	Phoneme

表三・實驗結果顯示，以學習語音相似度為基礎的反向音譯顯然優於其他方法

我們近一步調查排名分布的情形，結果如表四所示。學習語音相似度的優越性相當引人注目，因為超過 70% 的中文音譯專有名詞能與原始英文字組正確配對。

Rank	1	2-5	6-10	11-15	16-20	21-25	25+
Chen98	33.3	31.6	6.8	9.1	2.8	1.4	12.5
Lin00	57.7	25.5	4.6	2.1	2.1	1.0	5.5
Learning	77.7	18.0	2.0	1.0	0.3	0.1	0.8

表四・排名分布分析

5 討論

既然基底的語音結構(underlying phonological structures)並非假定而來，對齊方式也不一定要用人工標明，如果再加上可供參考的訓練語料庫和發音字典，擴大學習演算法則到其他語言配對的難度應可減至最低。然而，並不是所有的發音字典都可供利用，而且語音合成(speech synthesis)在產生專有名詞的發音上也有困難。雖然約莫 1/5 訓練語料庫內的項目並未收錄在發音字典當中，但是相較於 Lin00 完全忽略沒有發音的成對字組(pairs)，學習演算法則仍有較好的表現。未來發展的方向之一將朝向在沒有字典的情況下獲取發音，語音合成的研究工作亦朝此方向努力(Llitjos and Black, 2001)。

學習演算法則能獲取不易由人工依據語音常識指派的細微相似度。以圖二的矩陣為例²，成對母音(oU, uo)及(oU, u)的得分為 2 和 10，意謂著他們有不同的程度相似，然而在先前的研究當中他們都得了 5 分。成對子音(g, k)得到 10 分，但在以語音常識為基礎、又沒有中文有聲子音與無聲子音區別的系統中只得到 8 分。在沒有任何語音分析的情況下，學習演算法則仍然能不假人工而得到語音類似度。

反向音譯在機器翻譯的領域裡有著特殊的重要性。本文所探討的反向音譯是基於以相似度為本的架構，發展出學習演算法則，自動從訓練語料庫中獲取語音相似度。實驗結果顯示學習演算法則能有效地擷取相似度，而且所得到的相似度比人工指派分數的矩陣具有更佳的辨別性。

參考文獻

- Guo-Wei Bian and Hsin-Hsi Chen. 2000. Cross language information access to multilingual collections on the internet. *Journal of American Society for Information Science*, 51(3):281-296.

- M. Biehl and P. Riegler. 1994. On-line learning with a perceptron. *Europhysics Letters*, 28:525-530.
- Hsin-Hsi Chen and Jen-Chang Lee. 1996. Identification and classification of proper nouns in chinese texts. In *Proceedings of 16th International Conference on Computational Linguistics*, pp. 222-229.
- Hsin-Hsi Chen. 1997. Cross-language information retrieval. In *Proceedings of ROCLING Workshop on ED/MT/IR*, pages 4-1~27, Taipei, June 2. Academia Sinica.
- Hsin-Hsi Chen, Sheng-Jie Huang, Yung-Wei Ding, and Shih-Chung Tsai. 1998 Proper name translation in cross-language information retrieval. In *Proceedings of 17th COLING and 36th ACL*, pages. 232-236, Montreal, Quebec, Canada, August 10-14.
- Cmudict. 1995. The cmu pronouncing dictionary 0.6.
<ftp://ftp.cs.cmu.edu/project/speech/dict>.
- Richard O. Duda, Peter E. Hart, and David G. Stork. 2001. *Pattern Classification*. Wiley-Interscience Publication, 2nd edition.
- Pascale Fung and Lo Yuen Yee. 1998. An IR Approach for translating new words from nonparallel, comparable texts. In *Proceedings of 17th COLING and 36th ACL*, pages 414-420, Montreal, Quebec, Canada, August 10-14.
- Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.
- James L. Hieronyms. 1997. Worldbet phonetic symbols for multilanguage speech recognition and synthesis. Technical report, AT&T Bell Labs.
- Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. *Computational Linguistics*, 24(4):599-612.
- Donald E. Knuth. 1973. *The Art of Computer Programming*, pages 391-392. Addison-Wesley, Reading, Mass.
- V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics—Doklady* 10, 10:707-710.
- Wei-Hao Lin and Hsin-Hsi Chen. 2000. Similarity measure in backward transliteration between different character sets and its application to clir. In *Proceedings of Research*

on Computational Linguistics Conference XIII, pages 97-113, Taipei, Taiwan, August 24-25.

- Ariadna Font Llitjos and Alan W. Black. 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In *Eurospeech*, Aalborg, Denmark.
- W. Masek and M. Paterson. 1980. A faster algorithm computing string edit distances. *Journal of Computer System Science*. 20:18-31.
- Douglas W. Oard. 1999. Issues in cross-language retrieval from document image collection. In *Symposium on Document Image Understanding Technology*, Annapolis, MD, April.
- V. Pagel, K. Lenzo, and A. Black. 1998. Letter to sound rules for accented lexicon compression. In *Proceedings of the 1998 International Conference on Spoken Language Processing*, Sydney, Australia.
- E. S. Ristad and P. N. Yianilos. 1998. Learning string-edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5)
- Bonnie Glover Stalls and Kevin Knight. 1998. Translating names and technical terms in Arabic text. In *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages*.
- P. Thompson and C. Dozier. 1997. Name searching and information retrieval. In *Proceedings of Second Conference on Empirical Methods in Natural Language Processing*, Providence, Rhode Island.
- R. Sproat et al. 1994. A stochastic finite-state word segmentation algorithm for chinese. In *Proceedings of 32nd Annual Meeting of ACL*, New Mexico, pp. 66-73.
- Stephen Wan and Cornelia Maria Verspoor. 1998. Automatic english-chinese name transliteration for development of multilingual resources. In *Proceedings of 17th COLING and 36th ACL*, pages 1352-1356, Montreal, Quebec, Canada, August 10-14.
- John Wells, 1997. *Handbook of Standards and Resources for Spoken Language Systems*, chapter IV. Mouton de Gruyter, Berlin.

國科會國際數位圖書館合作研究計畫
NSC's International Digital Library Collaborative Research

數位圖書館聯合會議訪問報告

(Joint Conference on Digital Libraries)

2002 年 7 月 14 日~2002 年 7 月 26 日

國科會國際數位圖書館合作研究計畫(IDLP)II-子計畫一
英中雙語資訊系統相關語言處理技術和資源整合之研究

NSC90-2750-H-002-731

主持人

陳信希 教授

台灣大學資訊工程學系

民國 91 年 7 月

一・參訪單位簡介

數位圖書館聯合會議(Joint Conference on Digital Libraries)是數位圖書館界主要的國際論壇，結合了過去 ACM 數位圖書館會議和 IEEE-CS 數位圖書館會議，在數位圖書館聯合會議之後，就是美國國科會(NSF)所主導的美國數位圖書館計畫 PI 會議。

今年的數位圖書館聯合會議在奧勒岡州(Oregon)的波特蘭(Portland)召開，時間是 7 月 14 日到 7 月 18 日。第一天為 tutorials，中間三天是主要的會議，最後一天是 workshops。

這次的 tutorials 涵蓋如下的課程：Open Archive Initiative Protocol for Metadata Harvesting、Thesauri & Ontologies、Build Digital Library Using Open Source Software、Bioinformatics and Digital Libraries 等。

主議程又分為 13 個平行的 sessions，包括 Building and Using Cultural Digital Libraries、Summarization and Question Answering、Studying Users、Classification and Browsing、Digital Libraries for Education、Novel Search Environments、Video and Multimedia Digital Libraries、Searching across Language, time and Space、Models and Tools for Generating Digital Libraries、Music Digital Libraries、Preserving, Securing and Assessing Digital Libraries、Image and Cultural Digital Libraries、Digital Libraries for Spatial Data 等。

最後一天的 workshops 共有 7 場，含 Interface Design、Usability、Visual Interfaces、Music Libraries、Digital Gazetters、TREC Genomics、DLI2。

二・參訪行程

於七月十四日搭長榮班機由台北出發，經西雅圖轉搭 AmTran 火車到波特蘭，開始本次參訪的行程。在七月十八日數位圖書館國際會議結束後，隔天七月十九日搭 AmTran 火車回西雅圖，在西雅圖順道訪問華盛頓大學 Jeff A. Bilmes

教授實驗室，於七月二十五日凌晨返國。

教授實驗室，於七月二十五日凌晨返國。

三・心得

本次參訪，除了吸收數位圖書館相關技術外，也觀察到數位圖書館評估的重要性，以及生物資訊和數位圖書館的接軌。在大會中就有相關 workshops 提到生物資訊數位博物館的建立，並擬在資訊檢索最大的評比-TREC 中加入 Genomics Track 項目。其中涉及的主題包括：

- (1) How to define implement environment?

User-oriented vs. system-oriented?

- (2) 內容部份包括：What to use? How much? From where?

Open vs. closed collection?

- (3) 評估部份包括：評估人員及評估標準等。

經與會人員的討論，歸納出幾點結論，可能涵蓋的工作項目有：

- (1) abstract, full-text

- (2) citation better with named entities

- (3) detection of disagreements

- (4) named entities problem would help promote standards

- (5) going from microarray output to other information

- (6) identify interactions between molecular in text

- (7) connect equivalent genes or sequences across organisms

- (8) abbreviation definition

- (9) discern function, process, and localization of gene in papers or abstracts

- (10) discussion of gene-what organism?

四・帶回資料

- (1) Proceedings of the second ACM/IEEE-CS Joint Conference on Digital Libraries
- (2) Proceedings of TREC Genomics Pre-Track Workshop



NTU Proposal for TREC Genomics Pre-Track Workshop

Hsin-Hsi Chen and Wen-Juan Hou

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
E-mail: hh_chen@csie.ntu.edu.tw

Natural Language Processing Lab at CSIE, NTU, Taiwan

- **Multilingual Information Management**
 - This research studies classification, retrieval, filtering, extraction, and summarization of multilingual data.
- **Language Engineering**
 - The research studies how to represent, acquire, employ, and integrate linguistic knowledge.

Hsin-Hsi Chen (NTU)

2

Multilingual Information Management

- Chinese-English/English-Chinese query translation
- Proper name transliteration
- CLIR and Web translation
- Language identification
- English/Chinese information extraction
- Multilingual multi-document summarization
- Question and answering
- Multilingual topic detection and tracking
- National palace digital museum

Hsin-Hsi Chen (NTU)

3

Language Engineering

- segmentation
- part of speech tagging
- full/partial parsing
- term extraction
- prepositional phrase attachment
- word/sentence/document/structure alignment
- word/passage/document clustering/classification
- tree bank construction

Hsin-Hsi Chen (NTU)

4

Running Projects

- Knowledge Extraction Technologies from Heterogeneous Web Resources and Their Applications
- Biological Data Mining Using Natural Language Processing Technologies

Hsin-Hsi Chen (NTU)

5

Evaluation Conferences

- Evaluation tasks we participated
 - MET2 and MUC7 (1998)
 - SUMMAC (1998)
 - Topic Detection in TDT (1999)
 - Topic Tracking in TDT (2000)
 - Link Detection in TDT (2001)
 - QA Task in TREC (1999, 2000, 2001, 2002)
 - CLIR Task in TREC (2000)
- Evaluation tasks we organized
 - NTCIR 2 (2000-2001)
 - NTCIR 3 (2001-2002)
- Evaluation tasks we supported
 - CLEF 2001
 - CLEF 2002

MET2 and MUC7 (1998)

- Major information extraction evaluation conferences
 - MUC7: English
 - MET2: Chinese
- Named entity
 - Names: people, organizations, locations
 - Number: monetary/percentage expressions
 - Time: date/time expressions
- Named entity extraction
 - Given English/Chinese documents, insert SGML tags into the text to mark each string that represents a person, organization, or location name, or a date or time stamp, or a currency or percentage figure

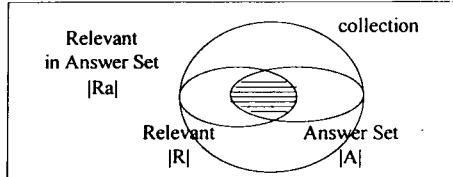
Hsin-Hsi Chen (NTU)

7

MET2 and MUC7 (1998)

Evaluation Criteria

- Precision $\frac{|R_a|}{|R_d|}$
- Recall $\frac{|R|}{|A|}$
- F-measures $F = \frac{2 \times P \times R}{P + R}$



SUMMAC (1988)

- first automatic summarization evaluation conference
- Categorization Task
 - Summarization systems produce summary for each document
 - The assessor will read the summary and then assign the summary into one of five topics or the sixth topic, 'non-relevant' topic.

Hsin-Hsi Chen (NTU)

9

SUMMAC (1988)

- Ad hoc Task
 - Adaptive to user's information need
 - Critical for Internet environment
- Evaluation Metric
 - Recall
 - Precision
 - F-measures

Hsin-Hsi Chen (NTU)

10

Topic Detection and Tracking

- Topic Detection (1999)
 - grouping all stories as they arrive, based on the topics they discuss
- Topic Tracking (2000)
 - monitoring the stream of news stories to find additional stories on a topic that was identified using several sample stories
- Link Detection (2001)
 - deciding whether two randomly selected stories discuss the same news topic
- Evaluation Criteria
 - cost function i.t.o. missed detection and false alarm

$$(C_{Det})_{norm} = C_{Det} / \text{MIN}(C_{Miss} \cdot P_{target}, C_{FA} \cdot P_{non-target})$$

QA Task in TREC

- Find the exact answer, which can meet the users' need more precisely, from a huge unstructured database
 - Main task: 250-bytes
 - List task: A question does not only ask for its information need but also a specified number of answers
 - Context task: A series of questions are submitted, which are somewhat relative to the previous questions.
- Evaluation metric
 - MRR (Mean Reciprocal Rank) $MRR = \sum_{i=1}^N r_i / N$

Hsin-Hsi Chen (NTU)

12

CLIR Task in TREC 9

- English-Chinese information retrieval
 - English query to Chinese document collection
- Evaluation metric
 - average precision
 - R-precision
 - ...

Hsin-Hsi Chen (NTU)

13

NTCIR 3 CLIR Task

- MLIR

– C → C, J, E	– C → J, E
– E → C, J, E	– E → J, E
– J → C, J, E	– J → J, E
– K → C, J, E	– K → J, E
– C → C, J	– C → C, E
– E → C, J	– E → C, E
– J → C, J	– J → C, E
– K → C, J	– K → C, E

Hsin-Hsi Chen (NTU)

14

NTCIR CLIR Task

- | | |
|--|--|
| • Bilingual CLIR <ul style="list-style-type: none"> – C → J – E → J – K → J – C → K – E → K – J → K – J → C – K → C – E → C | • Single Language IR <ul style="list-style-type: none"> – C → C – J → J – K → K |
|--|--|

Hsin-Hsi Chen (NTU)

15

Material Used

Japan	Mainichi Newspaper (1998-1999): Japanese	236,664
	Mainichi Daily News (1998-1999): English	12,723
Korea	Korea Economic Daily (1994): Korea	66,146
Taiwan	CIRB011 (1998-1999): Chinese	132,173
	United Daily News (1998-1999): Chinese	249,508
	Taiwan News (1998-1999): English	7,489
	Chinatimes English News (1998-1999): English	1,715

Hsin-Hsi Chen (NTU)

16

Assessment Environment

The screenshot shows a web-based search interface for the CLEF 2001, 2002 tasks. At the top, there's a form for specifying search parameters like topic ID (034), topic set (CLIRFormalResPo-APENIC), and document IDs (151, 21, 1644). Below this is a list of documents with their titles and URLs. The main area displays an XML snippet of a document, likely a news article, with fields such as HEADLINE, DATE, and TEXT. A note at the bottom states: "A majority of Taiwan businesses with operations in China are optimistic about cross-strait relations in the short term, results of a recent opinion poll showed."

CLEF 2001, 2002

- We supported the Chinese version of topic set.

Hsin-Hsi Chen (NTU)

18

Current Works on Bioinformatics

- Use the contextual cues in biological texts to detect gene symbols or names
- Identify the meanings of co-occurrences of gene names in scientific texts
- Automatically extract facts from scientific abstracts and full papers in the biology domain, and use these to update databases
- Reconstruct the metabolic pathways

Hsin-Hsi Chen (NTU)

19

Evaluation Tasks

- Gene names identification
- The homology search
- The interaction between genes
- The reconstruction of metabolic pathways
- Extracting instances of relations among objects

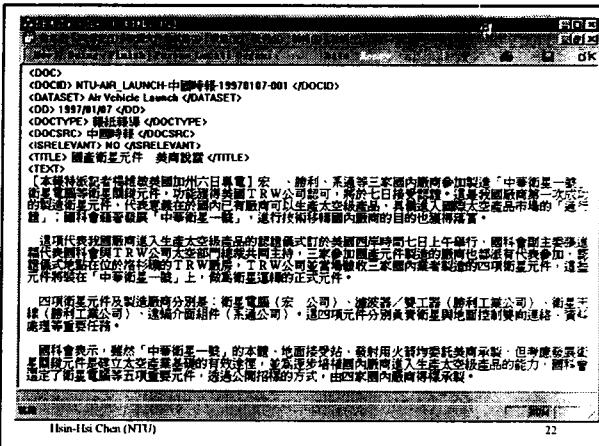
Hsin-Hsi Chen (NTU)

20

Named Entity Tagging Environment

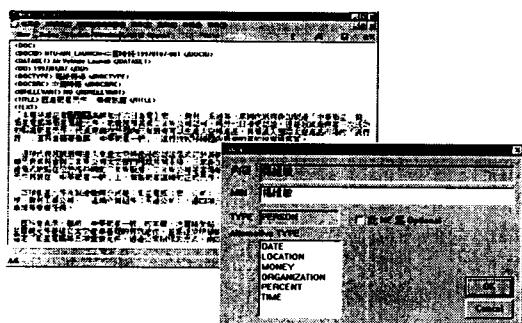
Hsin-Hsi Chen (NTU)

21



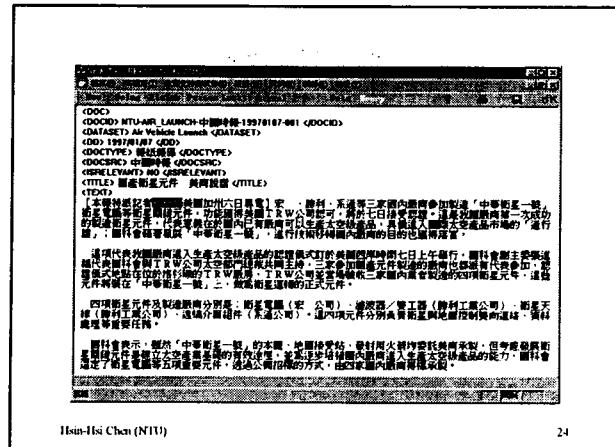
Hsin-Hsi Chen (NTU)

22



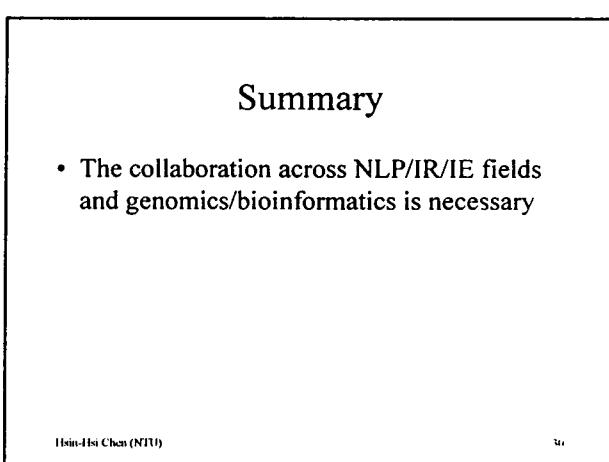
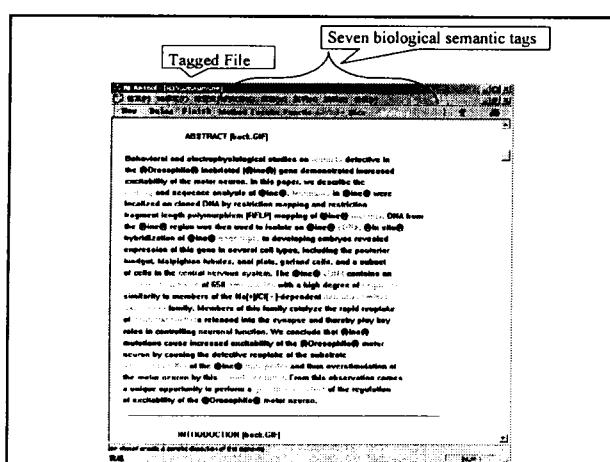
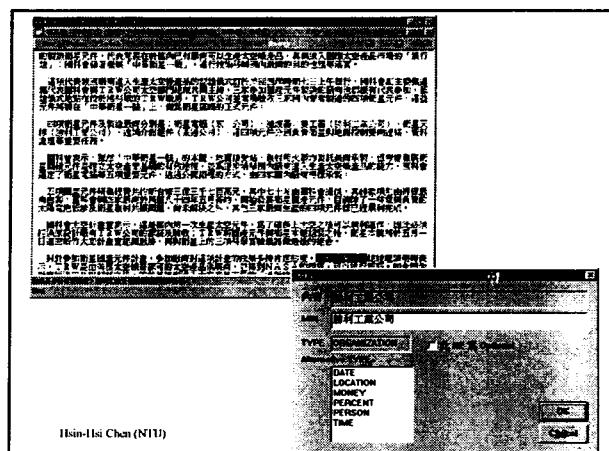
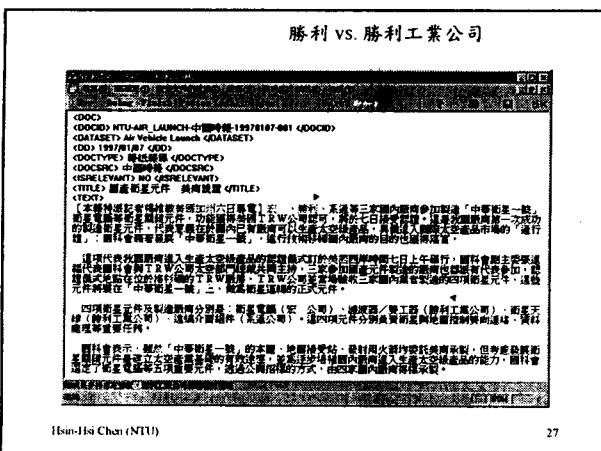
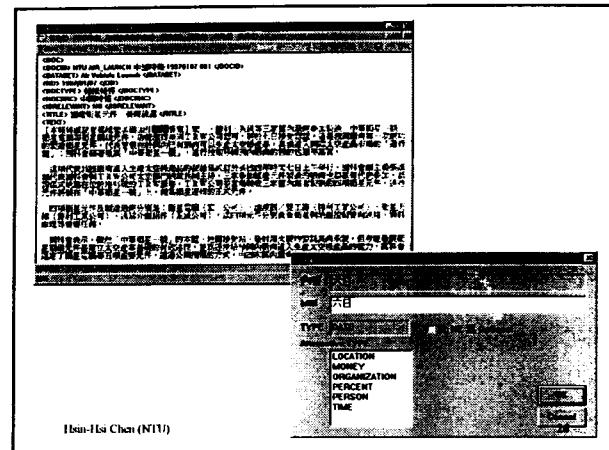
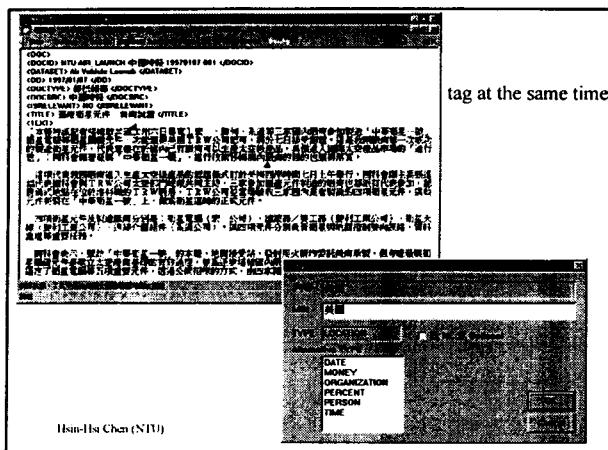
Hsin-Hsi Chen (NTU)

23



Hsin-Hsi Chen (NTU)

24



國科會國際數位圖書館合作研究計畫
NSC's International Digital Library Collaborative Research

中國大陸參觀訪問報告
2002年3月30日~2002年4月7日

國際數位圖書館合作研究計劃(IDLP)II-子計劃一
英中雙語資訊系統相關語言處理技術和資源整合之研究
NSC 90-2750-H-002-731

主持人
陳信希 教授
台灣大學資訊工程學系

共同主持人
陳光華 教授
台灣大學圖書資訊學系

民國 91 年 4 月

目 次

壹、參訪緣起.....	2
貳、參訪成員名單.....	3
參、參訪行程簡介.....	3
肆、參訪記錄.....	4
一、北京大學、清華大學.....	4
二、中國國家圖書館、中國數字圖書館有限責任公司.....	5
三、上海交通大學.....	7
四、浙江省圖書館.....	8
五、浙江大學.....	9
伍、附件.....	11
一、台灣地區 IDLP 總計畫	
二、台灣地區 IDLP 子計畫一	
三、台灣地區 IDLP 子計畫二	
四、台灣地區 IDLP 子計畫三	
五、台灣地區 IDLP 子計畫四	
六、台灣地區 IDLP 子計畫五	
七、北京大學數字圖書館相關計畫現況	

壹、參訪緣起

本研究計劃之背景為美國白宮科技顧問陳劉欽智博士有鑑於美國國家科學基金會(NSF)推動 DLI-1 成果斐然，同時我國國科會也於這幾年推動數位博物館專案計劃，也有相當的成果，認為有必要透過國際性的合作計劃，展開相關的技術標準的討論，於是本計劃成員於 2000 年與陳劉欽智博士共同向國科會與美國 NSF 提出合作計劃，美國部份獲 NSF 支持，總計劃名稱為 CMNet (Chinese Memory Net)。本計劃即是在這樣的背景之下，結合國內中央研究院、台灣大學、新竹清華大學等學術單位，以現有的研究成果為基礎，提出五項研究計劃，與大陸的北京大學、北京清華大學、上海交通大學，以及美國西蒙斯大學進行國際性的學術研究計劃。

隨著網際網路的發展，各種以其為使用環境的研究亦引起各國學術研究機構與個人的重視，其中尤以數位圖書館/博物館的研究為最。數位圖書館/博物館跨越學科領域的特性，更促使學者相互交流，形成所謂的跨學科整合性的研究，同時數位圖書館/博物館提供的服務更是未來網際網路最重要的價值之一，全球各主要先進國家莫不視為是未來國力的重要指標之一。我國國科會在 1998 年起推動數位博物館專案計劃，結合資訊科技，資訊組織，資訊內涵三種主要的學者專家，透過主題性計畫與技術支援計劃的執行，取得相當重大的成果。這幾年的成果使得我國在亞洲地區數位圖書館/博物館的研究中，具有舉足輕重的領導地位，因此國科會更於 2001 年繼而推動國家典藏數位化計劃，2002 年更進而將國內多個數位博物館相關計畫合併為數位典藏國家型科技計畫。

2002 年 3 月大陸 IDLP 計畫正式開始執行，預計至 2004 年底結束；而台灣地區 IDLP 計畫則將於 2002 年 7 月結束，併入數位典藏國家型科技計畫，故藉春假之課餘機會，台灣地區 IDLP 計畫成員組成一支訪問團赴大陸相關數位化研究及實施單位進行正式的交流與訪問，以為未來數位圖書館相關研究交流奠定基礎與共識。

貳、參訪成員名單

本次出訪一行共十三人，計畫人員如下所列：

總計畫	主持人	台大圖資系	陳雪華教授
	共同主持人	台大資工系	項潔教授
子計畫一	主持人	台大資工系	陳信希教授
子計畫二	共同主持人	東吳資科系	柯淑津副教授
子計畫三	共同主持人	師大圖資研究所	陳昭珍副教授
子計畫四	主持人	清大資工系	蘇豐文教授
子計畫五	主持人	中研院資科所	黃世昆助研究員

其他成員有：師大圖資研究所所長吳美美教授，計畫專任助理洪筱盈、高世芯、張慶瑞，台大圖資系博士生邱子恆、碩士生張懷文。



圖：訪問團一行共十三人於北京清華大學留影

參、參訪行程簡介

時間：2002年3月30日～4月7日，共計九天。

參訪單位：

- 4月1日(週一) 北京大學、北京清華大學
- 4月2日(週二) 中國國家圖書館、中國數字圖書館有限責任公司
- 4月3日(週三) 上海交通大學
- 4月5日(週五) 浙江省圖書館
- 4月6日(週六) 浙江省圖書館特藏室、浙江大學

肆、參訪記錄

自從 1996 年後，大陸地區一些圖書館陸續開始數字圖書館的實踐工作，研究人員開始研究實際操作中的一些問題，如數位化加工、數位資源的組織、分類和服務等。而 1998 年底中國大陸教育部「211 工程」公共服務體系中的「中國高等教育文獻保障系統（簡稱 CALIS）」專案啟動以後，大量數位化資源湧入大學圖書館，以北京大學、清華大學、南京大學、上海交通大學為代表，這些大學圖書館，日益關注傳統圖書館資源和數位資源及其服務的整合。

雖然有多個單位及圖書館開始進行數位化推行的工作，但因為規模較小，大多不具系統化。故在大陸區進行數位化工作並力行數位圖書館相關應用研究最具代表性、並較有實力的單位，主要有國家圖書館、北京大學數字圖書館研究所、北京清華大學圖書館等單位。本次 IDLP 計畫訪問大陸地區，主要即是與這些積極參與數位化工作及研究的單位進行雙邊的交流與尋求雙方合作的可能性。

一、北京大學、清華大學

時間：2002 年 4 月 1 日（週一）

地點：北京清華大學圖書館

大陸地區參與人員，共約 20 餘人：

清華大學：計算機系系主任周立柱教授、邢春曉副教授，圖書館劉桂林館長，多位圖書館員及研究生等。

北京大學：計算機系楊東青教授、張銘副教授，多位圖書館員及研究生等。

交流事項摘要：

交流會議先由台灣地區 IDLP 計畫研究成員進行約 90 分鐘有關台灣 IDLP 計畫研究現況及研究重點成果介紹，每一子計畫出訪人員分別以投影片或系統展示；隨後由北京大學楊東青教授、清華大學邢春曉副教授分別介紹各自關於數位圖書館之研究主題及計畫成果。藉由雙方的交流介紹其研究重點與成果，不但讓北京大學、清華大學進行數位圖書館相關研究的教授及研究學生更加瞭解台灣地區的研究重點外，也使我們對於大陸地區的數位圖書館相關計畫及其研究重點有初步的認識。



圖：於清華大學圖書館會議剪影

北京大學目前主要研究內容包括：基於 WebGIS 的拓片檢索及論文知識導航系統、科技文獻導航系統。其中基於 WebGIS 的拓片檢索及論文知識導航系統以北京大學古籍數位圖書館為中心，將以 OAI-PMH 為基礎提供 Metadata，並研究其數位圖書館總體模式、系統層次結構、數位化標準、數位資源建設等。其中，北京大學古籍數位圖書館收藏善本、拓片、輿圖、敦煌書畫為主。目前已完成 Metadata 標準、實驗系統的設計，正進行資料加工、數位化加工的資源建設階段，即將在網路上開展服務。Metadata 標準體系研究也已完成「中文 Metadata 標準框架」以及古籍、拓片、輿圖等具體的 Metadata 標準設計工作。

北京大學十分重視數位化工作及其相關實務的研究，更於 1999 年 9 月正式成立「數位圖書館研究所」，由 CALIS 管理中心、北京大學圖書館、北大資訊科學中心組成。主要在數位圖書館模式、Metadata 和數位圖書館的體系結構等方面開展系統化的研究。北京大學圖書館也很早就開始數位圖書館研究，但真正開始系統化且組織化的研究和實作則在 2000 年初才開始。目前正以建設一個學術型研究型的數位圖書館為目標。

清華大學數位圖書館相關研究工作以其建築數位圖書館為主要核心。清華大學建築數字圖書館（THADL）於 1999 年開始，目標的是建立一個原型示範系統，作為研究和建立數位圖書館的突破口。這個建築數字圖書館以中國營造學社與梁思成生平為主線，構建 THADL，收集了「營造學社」，花費 15 年實地測繪 2783 處古建築的圖紙資料，同時提供古建動畫。技術上採用了分佈物件技術和智慧代理技術，構建分層的系統服務體系，並建立了面向物件的分散式多媒體資料庫。Metadata 則以 Dublin Core 為基礎進行適當的擴充。

二、中國國家圖書館、中國數字圖書館有限責任公司

時間：2002 年 4 月 2 日（週二）

大陸地區接待人員：

中國國家圖書館：陳力副館長、孫公關處長等。

中國數字圖書館有限責任公司：技術總監、經理等。

交流事項摘要：

1996 年初，中國國家圖書館在文化部申請「數字式圖書館試驗專案」，可說是大陸地區最早開始進行數位圖書館實踐的單位。隨著這幾年數位圖書館的不斷升溫，許多圖書館，尤其是一些條件好的大學圖書館，都不同程度地開展數位圖書館的研究與實行，逐步把一些具本館或地方特色的資

源數位化，或將各類數位資源整理上網並提供服務。中國國家圖書館從 1997 年開始，主要對於中國「數位式圖書館試驗專案」中的一些問題開展研究，包括數字圖書館的概念、發展 SGML 應用分散式查詢與調度等，近年來加強了 Metadata 和系統結構的研究。中國國家圖書館的珍善本館藏亦十分豐富，目前數位化典藏以碑帖、拓片為主，目前約有 5000 筆數位化典藏，預計將增加到 9000 筆。

在交流訪問的過程中，我們發現中國國家圖書館、北京大學、清華大學、上海交通大學等大陸境內參與數位圖書館研究工作之重要單位均為「中國數位圖書館工程」的一部份。「中國數位圖書館工程」是國家級的數位資源系統工程，其建設目標是在 Internet 上形成大規模的、高質量的中文數位資料庫群，並通過國家骨幹通信網，向全國及全球提供高效服務。1998 年 7 月，中國國家圖書館正式向文化部提出申請，實施「中國數位圖書館工程」。工程建設內容包括：數位資源建設、數位圖書館軟/硬體基礎設施建設、應用系統開發、標準規範與法規的制定和推行、知識產權問題的處理、服務體系的建立及人才建設等。



圖：於中國國家圖書館訪問剪影



圖：與中國國家圖書館館長合影

2000 年 3 月起，由中宣部出版局、國家計委社會發展司等 21 個相關部門組成了「中國數位圖書館工程建設聯席會議」，作為工程建設的決策機構，負責整體規劃工程的建設方向，協調資源建設等。而「中國數位圖書館工程建設專家顧問委員會」則協助聯席會議對工程所涉及的規劃及實施方案、資源建設、技術路線、標準規範和知識產權等關係到全局性的重大問題給予諮詢和指導。

中國數字圖書館有限責任公司成立於 2000 年 4 月，是依附在中國國家圖書館館內之營利性公司，由於大陸官方深感於數位化資源的重要性，但是不可能每年都投入類似中國數位圖書館工程般的龐大經費，故由中國國

務院批准成立中國數字圖書館有限責任公司，使用中國國家圖書館的各類資源、人才與網路，提供信息資源內容組織、數位化加工、數位圖書館技術總體解決方案、數位資源庫建設的應用開發、多媒體信息資源內容等付費服務。

中國數字圖書館有限責任公司目前所開發的產品包括：網上讀書在線瀏覽系統、中國數圖選書軟件、數字圖書資源儲備等。其中數字圖書資源儲備目前有近 6000 萬頁（約 20 萬冊，其中近幾年出版的新書 5 萬冊），數位化圖書內容涵蓋經濟、文學、歷史、醫藥生物、工業、農業等各類，保持每天 20 萬頁的數位化速度增長。由於其顧客群以圖書館與相關組織為主，故為避免數位圖書館資源的重複建設，中國數字圖書館有限責任公司以合作建立分館的模式和用戶單位合作，目前已在全國建立了四、五十家「中國數字圖書館分館」。

中國數字圖書館有限責任公司未來首要目標在於建立數位化共建模式、標準與格式，其次是促進資源的共享與保存。

三、上海交通大學

時間：2002 年 4 月 3 日（週三）

地點：上海交通大學圖書館

大陸地區參與及接待人員，共約 20 餘人：

國際合作與交流處黃新昌科長、嚴良瑜副教授，圖書館陳兆能館長、楊宗英副館長、林皓明副館長，計算機系馬範援教授、張冬茱副主任、吳亞棟副教授，多位圖書館員及研究生等。

交流事項摘要：

上海交通大學亦是大陸地區進行數位圖書館研究工作的重點單位，交流會議亦循北大與清大的交流模式，先進行約 90 分鐘有關台灣地區 IDLP 計畫研究現況及研究重點成果介紹，每一子計畫出訪人員分別以投影片或系統展示；隨後由上海交通大學計算機系馬範援教授介紹該校關於數位圖書館之研究主題及計畫成果。藉由雙方的交流介紹其研究重點與



圖：上海交通大學訪問剪影

成果，不但讓上海交通大學進行數位圖書館相關研究的教授及研究學生更加瞭解台灣地區的研究重點外，也使我們對於上海交通大學的數位圖書館相關計畫及其研究重點有初步的認識。

上海交通大學目前主要研究重點在於：分散式 WebRobots、智慧型網路搜尋工具、網路資源擷取與搜尋引擎、自然語言處理研究、語音識別、音樂資訊檢索等。其數位圖書館的發展可分為兩階段：第一階段從 1996 年到 1998 年，為中國 211 工程的一部份，數位化重點在於民族音樂簡譜與五線譜的轉換、博士論文數據庫、上海交大重點學科導航系統等；第二階段為 985 計畫之一，自 1998 年至今，發展重點在於學術性電子期刊整合搜尋系統、教師指定參考書資料庫、隨選視訊系統等，目前技術團隊有 7 名教授、13 位技術人員與若干學生。



圖：於上海交通大學圖書館舉行交流會議

四、浙江省圖書館

時間：2002 年 4 月 5 日（週五）、2002 年 4 月 6 日（週六）

大陸地區接待人員：

圖書館程小瀾館長、王效良副館長、劉曉清副館長、古籍部童正倫副主任、地方文獻部袁逸主任等。

交流事項摘要：

浙江圖書館擁有十分豐富的館藏，包括文瀾閣四庫全書、敦煌經卷、宋元明刻本、稿抄本等，是推行數位化工作的重點單位。浙江圖書館目前的數位化重點工作在於：全國公共圖書館網路連結、文瀾閣四庫全書部分文字數位化、與浙江大學產生數位資源共享網等，其中地方文獻數位化的工作為首要目標。



圖：與浙江圖書館館長、副館長合影

五、浙江大學

訪問時間：2002年4月6日（週六）

大陸地區接待人員：

圖書館葉鷹館長、夏勇前館長等。

交流事項摘要：

本次訪問所談及的重點在於中國與美國所合作的「中美百萬冊書數位圖書館合作專案」。中美百萬冊書數位圖書館合作專案是由美國 Raj Reddy 博士（卡內基-梅隆大學電腦與機器人學教授、電腦學院前院長、美國總統資訊技術顧問委員會 PITAC 聯合主席）、陳劉欽智博士（Ching-chih Chen, 西蒙學院研究生院圖書館與資訊科學教授，美國總統資訊技術顧問委員會 PITAC 成員，美國 NSF 國際數位圖書館專案 - Chinese Memory Net 專案負責人）和中國高文博士(電腦科學教授、中國科技大學副校長、中國科學院研究生院常務副院長、中國國家高技術研究發展計畫智慧電腦主題前首席科學家)共同發起的旨在建設包含 100 萬冊圖書的數位圖書館研究與開發專案。計畫由中國及美國各出資約 1000 萬美金，預計進行 4 年，整合 100 萬冊數位圖書（中英文各 50 萬冊）上網，而且能夠在網際網路上流通。

藉由此專案，中國將與美方合作，開發一套高水準的數位圖書館技術平台，有效支援 100 萬冊書的加工、管理和服务。中國方面由 211 工程、985 計畫主導，包括浙江大學、北京大學、清華大學等在內的 11 所大學和中科院研究生院參加此項目，承擔項目的研究開發、選擇提供有特色的資訊資源（圖書以及圖畫、雕塑等）並承擔圖書和資訊資源的數位化工作。美國方面將整合卡內基-梅隆大學、西蒙學院和美國數位圖書館聯盟成員參加此項目，並將和 NSF 一道為此專案提供經濟和技術支援，具體包括電腦、掃描設備、軟體、技術培訓等。雙方在項目期間將共同在製作工具、技術、內容方面進行創造性的研究開發。該計畫之技術支援包括三大部分：

1. 數位加工子系統，包括資源的數位化、中英文識別、自動較對、自動排版等。



圖：於浙江浙江大學圖書館舉行交流會議

2. 資源管理子系統，支援海量多媒體資源的存儲、管理和檢索。
3. 知識服務子系統：包括資訊資源的智慧檢索、分析、處理、操縱、視覺化和互動等，開發數位圖書館應用的新模式，促進知識、知識生產者、知識傳播者、知識整理者、知識消費者之間關係的變革，使之更加有利於知識傳播。

為了確保計畫的執行，中國方面已於 2001 年 8 月派出以浙江大學校長潘雲鶴院士為團長的代表團訪問美國，與美方具體協商了項目的規劃和執行中的重大問題。代表團由浙江大學、北京大學、清華大學、復旦大學、南京大學、中國科學院研究生院的有關校長和教育部及國家計委的官員組成，雙方在卡內基-梅隆大學舉行了第一次工作會議。2002 年 3 月中國與美國雙方在浙江大學舉行了第二次工作會議，並正式將此計畫命名為「中美數字圖書館（China-America Digital Academic Library，簡稱 CADAL）」。

目前中國大陸方面有幾個大型的中文電子新書服務系統，重要的包括：

1. 超星數字圖書館 (<http://www.ssreader.com/>)：1998 年 7 月創立，約有 30 萬冊藏書，自行開發超星圖書瀏覽器 3.6 版。
2. 書生之家中華圖書網 (<http://www.21dmedia.com>)：2000 年 4 月創立，約 5 萬冊圖書，使用書生數字信息閱讀器 3.0B 版。
3. 中國數圖有限公司網上圖書館 (<http://www.d-library.com.cn/>)：2000 年 9 月成立，約 20 萬圖書，使用中國數圖瀏覽器 1.01 版。
4. 方正 Apabi 數字圖書館 (<http://www.apabi.com>)：2000 年 12 月成立，約 1 萬圖書，使用 Apabi Reader 1.5 版。

這些大型的民間中文電子新書服務系統，希望屆時能夠整合入中美百萬冊圖書數位圖書館中，但是圖書版權將是首待解決的重要問題。

國科會國際數位圖書館合作研究計畫 NSC's International Digital Libraries Collaborative Research

陳雪華
台灣大學圖書資訊系教授
項潔
台灣大學資訊工程學系教授

1

數位典藏國家型科技計畫 發展背景

結合國科會三個計畫而成：

- 一、數位博物館專案計畫 (1998~2002)
數位典藏的加值、應用與推廣
- 二、國家典藏數位化專案計畫 (2001)
國家級典藏的大量數位化
- 三、國際數位圖書館合作計畫 (2000~2002)
International Digital Library Project with NSF

3

台灣地區數位典藏重要計畫

- 國科會「數位博物館專案計畫」(1998-)
- 國科會「國家典藏數位化專案計畫」(2001-)
- 文建會「國家文化資料庫」(2001-)
- 國科會「國際數位圖書館合作計畫」(2000-)
- 國科會「數位典藏國家型科技計畫」(2002-)



IDLP計畫研究人員

- 總計畫：國際數位圖書館合作研究計畫
主持人：陳雪華 台灣大學圖書資訊學系教授
共同主持人：項潔 台灣大學資訊工程學系教授
- 子計畫一：英中雙語資訊系統相關語言處理技術和資源整合之研究
主持人：陳信希 台灣大學資訊工程學系教授
共同主持人：陳光華 台灣大學圖書資訊學系副教授
- 子計畫二：詞彙為本的知識連結
-朝向多語數位圖書館中之詞彙網路基礎架構
主持人：黃居仁 中央研究院語言所籌備處研究員
共同主持人：陳克健 中央研究院資訊科學研究所研究員
共同主持人：柯淑津 東吳大學資訊科學系副教授

4

ANIC's International Digital Library Collaborative Research



IDLP計畫研究人員

- 總計畫：國際數位圖書館合作研究計畫
主持人：陳雪華 台灣大學圖書資訊學系教授
共同主持人：項潔 台灣大學資訊工程學系教授
- 子計畫一：英中雙語資訊系統相關語言處理技術和資源整合之研究
主持人：陳信希 台灣大學資訊工程學系教授
共同主持人：陳光華 台灣大學圖書資訊學系副教授
- 子計畫二：詞彙為本的知識連結
-朝向多語數位圖書館中之詞彙網路基礎架構
主持人：黃居仁 中央研究院語言所籌備處研究員
共同主持人：陳克健 中央研究院資訊科學研究所研究員
共同主持人：柯淑津 東吳大學資訊科學系副教授

3

IDLP計畫研究人員

<p>子計畫三：數位圖書館中文詮釋資料系統之研究</p> <p>主持人：陳雪華 共同主持人：侯昭珍</p>	<p>國立台灣大學圖書資訊系教授</p>
<p>子計畫四：以代理人為基礎的數位圖書館的資訊搜集與服務</p> <p>主持人：蘇豐文 共同主持人：劉吉軒</p>	<p>清華大學資訊工程學系教授 國立政治大學資訊科學系</p>
<p>子計畫五：智慧財產權保護機制研究</p> <p>主持人：董世昆 共同主持人：廖弘源</p>	<p>中央研究院資訊所助理研究員 中央研究院資訊所研究員</p>
<p>-數位圖書館內涵保護及電子版權管理</p> <p>主持人：董世昆 共同主持人：廖弘源</p>	<p>中央研究院資訊所助理研究員 中央研究院資訊所研究員</p>

5

IDLP子計畫簡介

<p>子計畫一：中文對譯資訊系統</p> <p>著重於中文對譯資訊系統，由產出輸入英文，並仔細切人；並發展多項有效而與整合多項資訊的技術</p>	<p>子計畫二：中文對譯資訊系統，結合多項軟體工具處理技術和資源整合。</p>
<p>子計畫三：中文對譯資訊系統 / 圖書館中文對譯資訊系統</p> <p>建構中文對譯資訊系統，著重於中文對譯資訊系統，以協助數位圖書館 / 物流倉庫資訊系統管理與資訊資源。</p>	<p>子計畫三：中文對譯資訊系統 / 圖書館中文對譯資訊系統</p>
<p>子計畫四：以代理人為基礎的數位圖書館的資訊搜集與服務</p> <p>目標在於建構圖書館代理人知識庫，以達到及資訊系統之搜尋、並專注於電子版權管理的問題。</p>	<p>子計畫四：以代理人為基礎的數位圖書館的資訊搜集與服務</p>
<p>子計畫五：智慧財產權保護機制研究</p> <p>子計畫五：智慧財產權保護機制研究，並專注於電子版權管理的安全問題。</p>	<p>子計畫五：智慧財產權保護機制研究，並專注於電子版權管理的安全問題。</p>

6

5

附件二

國科會國際數位圖書館合作研究計畫(IDLP) II
子計畫一
英中雙語資訊系統相關語言處理技術和資源整合之研究

報告人

陳信希

台灣大學資訊工程學系

Linguistic Technologies and Resources for English-Chinese Bilingual Information Systems

Hsin-Hsi Chen (陳信希)
Department of Computer Science and
Information Engineering
National Taiwan University

Hsin-Hsi Chen

IDLP Presentation

1

Multilingual World

- There are 6,703 languages listed in the Ethnologue
- Digital libraries
 - OCLC Online Computer Library Center serves more than 17,000 libraries in 52 countries and contains over 30 million bibliographic records with over 500 million records ownership attached in more than 370 languages
- World Wide Web
 - Around 40% of Internet users do not speak English, however, 80% of Web sites are still in English

Hsin-Hsi Chen

IDLP Presentation

3

Outlines

- Goals
 - Query Translation
 - Machine Transliteration
- Evaluation
- An Application:
National Palace Digital Museum

Hsin-Hsi Chen

IDLP Presentation

2

Goal

- to share its valuable resources with users of different languages
- to utilize knowledge presented in a foreign language

IDLP Presentation

4

Cross-Language Information Retrieval

- CLIR
 - Select information in one language based on queries in another
- Major Problems
 - Queries and documents are in different languages (translation)
 - Words in a query may be ambiguous (disambiguation)
 - Queries are usually short (expansion)
 - Queries may have to be segmented (segmentation)
 - A document may be in terms of various languages (language identification)
 - Documents are in terms of various languages (ranking and merging)

Hsin-Hsi Chen
IDLP Presentation

5

Ambiguities

- Query translation
 - Unify the language in queries and documents.
- Translation ambiguity
 - A word in a source query may have more than one sense.
- Target polysemy
 - A word in a target query may have more than one sense.
- MT
 - Readers may disambiguate the meaning of a target polysemous word
- CLIR
 - Target polysemy adds extraneous senses and affects retrieval performance

Hsin-Hsi Chen
IDLP Presentation

7

An Example of Target Polysemy

- Chinese-English information retrieval (CEIR)
 - Employ Chinese queries to retrieve English documents.
 - The Chinese word “银行” (yin2hang2) is unambiguous, but its English translation “bank” has 9 senses.
 - When the Chinese word “银行” (yin2hang2) is issued, it is translated into the English counterpart “bank” by dictionary lookup without difficulty, and then “bank” is sent to an IR system.
 - The IR system will retrieve documents that contain this word. Because “bank” is not disambiguated, irrelevant documents will be reported.

Hsin-Hsi Chen
IDLP Presentation

8

Query Translation

Hsin-Hsi Chen
IDLP Presentation

6

An Example of Translation Ambiguity

- English-Chinese information retrieval (ECIR)
 - Employ English queries to retrieve Chinese documents.
 - When “bank” is submitted to an ECIR system, we must disambiguate its meaning at first.
 - If we can find that its correct translation is “銀行” (yin2hang2), the subsequent operation is very simple.
 - “銀行” (yin2hang2) is sent into an IR system, and then documents containing “銀行” (yin2hang2) will be presented.

Hsin-Hsi Chen

IDLP Presentation

9

Multiplication Effects of Translation Ambiguity and Target Polysemy

- A Chinese word may have more than one sense.
 - “運動” (yun4dong4) has the following meanings: (1) sport, (2) exercise, (3) movement, (4) motion, (5) campaign, and (6) lobby.
- Each corresponding English word may have more than one sense.
 - “exercise” may mean *a question or set of questions to be answered by a pupil for practice; the use of a power or right*; and so on.
- The multiplication effects of translation ambiguity and target polysemy make query translation harder.

Hsin-Hsi Chen

IDLP Presentation

10

Machine Transliteration

- Query is usually translated into another language in CLIR.
- Disambiguation is important in query translation.
 - Proper Nouns appear extensively in user queries.
 - How do we translate proper nouns when they are not covered in dictionaries?

Hsin-Hsi Chen

IDLP Presentation

11

Machine Transliteration in CLIR

- Query is usually translated into another language in CLIR.
- Disambiguation is important in query translation.
 - Proper Nouns appear extensively in user queries.
 - How do we translate proper nouns when they are not covered in dictionaries?

Hsin-Hsi Chen

IDLP Presentation

12

Classification

- Direction of Transliteration
 - Forward (Firenze → 義冷翠)
 - Backward (阿尔诺史瓦辛格 → Arnold Schwarzenegger)
- Character Sets b/w Source and Target Languages
 - Same
 - Different

Hsin-Hsi Chen IDLP Presentation 13

Forward Transliteration b/w Same Character Sets

- Especially b/w Roman Characters
- Usually no transliteration is performed.
- Example
 - Beethoven (貝多芬)
 - Firenze → Florence, Muenchen → Munich, Praha → Prague, Moskva → Moscow, Roma → Rome
 - 小淵惠三

Hsin-Hsi Chen IDLP Presentation 14

Forward Transliteration b/w Different Character Sets

- Procedure
 - Sounds in Source language → Sounds in Target language → Characters in Target language
- Example
 - 吳宗憲 → Wu × {Tsung, Dzung, Zong, Tzung} × {Hsien, Syan, Xian, Shian}
 - Lewinsky → 露文斯基, 柳思基, 陸愛絲姍, 陸文斯基, 呂茵斯基, 李文斯基, 露溫斯基, 露恩斯基, 李變斯基, 李文絲基, etc.

Hsin-Hsi Chen IDLP Presentation 15

Backward Transliteration b/w Same Character Sets

- Few or nothing to do because original transliteration is simple or straightforward

Hsin-Hsi Chen IDLP Presentation 16

Backward Transliteration b/w Different Character Sets

- The Most Difficult and Critical
- Two Approaches
 - Reverse Engineering
 - Mate Matching

Hsin-Hsi Chen

IDLP Presentation

17

IR Evaluation and Testing

- Importance of IR evaluation
 - Verify performance of IR systems
 - Compare different IR systems or techniques
- Which modules are evaluated?
- Which linguistic issues are concerned for Chinese IR or English-Chinese CLIR?

Hsin-Hsi Chen

IDLP Presentation

19

Evaluation

- The Most Difficult and Critical
- Two Approaches
 - Reverse Engineering
 - Mate Matching

IDLP Presentation

18

Hsin-Hsi Chen

C-C and E-C Text Retrieval at NTCIR2

- Coverage
 - Full news articles from Web sites
 - News articles of different subject categories
- Sources
 - May 1998 ~ May 1999
 - China Times (中國時報)
 - Commercial Times (工商時報)
 - China Times Express (中時晚報)
 - Central Daily News (中央日報)
 - China Daily Newspaper (中華日報)

IDLP Presentation

20

Hsin-Hsi Chen

CLIR Task at NTCIR 3

- Chinese/English/Japanese/Korean topics to multilingual document collection containing Chinese, English, Japanese, and Korean documents

K
J
E
C

A vertical column of black dots arranged in a grid pattern. The dots are circular and have varying diameters. They are arranged in approximately 10 rows, with some rows having more dots than others. The overall effect is a decorative or abstract graphic element.

Hsuan-Hsi Chen DLP Presentation 21

Challenging Issues

- Multilingual IR
 - Chinese → Chinese • English • Japanese • Korean
 - English → Chinese • English • Japanese • Korean
 - Japanese → Chinese • English • Japanese • Korean
 - Korean → Chinese • English • Japanese • Korean
 - Segmentation issue in Chinese, Japanese and Korean
 - Vocabulary issue among language pairs

112

22

Search Modes

- Free search
 - users describe their information need using natural languages (Chinese or English)
 - Specific topic search
 - users fill in specific fields denoting authors, titles, dates, and so on

Hsin-Hsi Chen

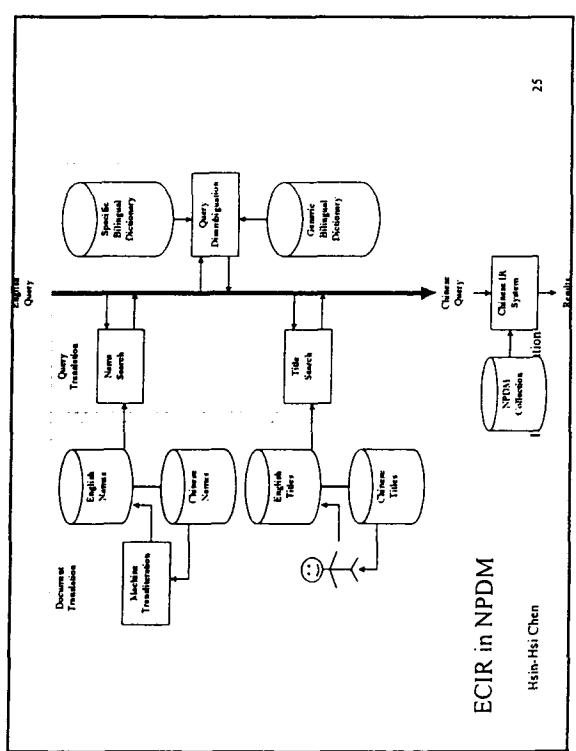
24

An Application: National Palace Digital Museum

- National Palace Digital Museum
All Application:

JIDP Presentation
Hsin-Hsi Chen

23



附件三

國科會國際數位圖書館合作研究計畫(IDLP) II 子計畫二 詞彙為本的知識連結： 朝向多語數位圖書館中之詞彙網路基礎架構

報告人

柯淑津

東吳大學資訊科學系

子計畫二

詞叢為本的知識連結 --
朝向多語數位圖書館中之詞叢網絡
基礎架構

黃居仁（中央研究院語言所籌備處 研究員）

陳克健（中央研究院資訊科學所 研究員）

柯淑津（東吳大學資訊科學系 副教授）

1

Background

- 國際數位圖書館需要多語言資訊處理能力
多語語文座標 + 多語知識管理技術
理想使用環境
- 建立多語詞叢語言網絡作為資訊處理
的基礎
建立中共雙語詞網
利用國際學術資源
 - 歐語詞網

2

Objective

- 建構中英雙語詞網的第一步
中文詞叢與英語詞網節點的對應
 - 中文詞網
 - 中文詞叢語言概念關係
 - 中英詞網連結
 - 形成多語網路
 - 轉換、修正、與擴充為中英雙語詞網

3

Resources

- Monolingual dictionaries
- Bilingual dictionaries
 - Chinese-English Dictionaries
 - English-Chinese Dictionaries
- Corpus
 - 中研院平衡語料庫
 - BNC

4

Project Contents

- 中文詞彙與英語詞網節點的對應
- 語意關係定義
- 語意關係標記

6

Current Status

- 中文詞彙與英語詞網節點的對應
 - 完成 87095 / 99642 (87.4%) Synsets
 - 95235 中文翻譯
 - 45160 中文詞彙

6

Current Status

- 語意關係定義
 - 要求：明確的關係判定標準
 - 效果：連結的一致性、適當性與可檢驗性
- 方法：採歐語詞網語言測試法
- 進度：41 種中文語意關係測試
 -

7

Semantic Relations Test

- 同義詞、近同義詞
- 反義詞、近反義詞
- 上位詞、下位詞
 - 同詞性、跨詞性
- 整體詞、部份詞
 - 部件、成員、單位、成份、區域
- 導致、肇因於
- 涉入者、角色
- 施事者、經歷者、工具、場所、動向

8

Current Status

- 語意關係標記

中英概念詞彙化的差異

非中文單詞

檢測中譯詞與詞集詞義間的語意關係

- 等同、上位、下位

9

Experiment

- 小型詞網實驗

步驟

- 選取231個中文常用語詞
- 透過對應的英語詞網節點與其語意關係
- 找出語意相關詞
- 目的
 - 檢測以『英語詞網架構』建立中文詞網的適當性
 - 預見大型詞網建立時的問題

10

Experimental Result

- 詞網規模

231 個詞彙

• 108 名詞、52 動詞、20 形容詞、51 副詞

42606 組詞彙對譯

493 個詞義 + 944 個相關詞義連結

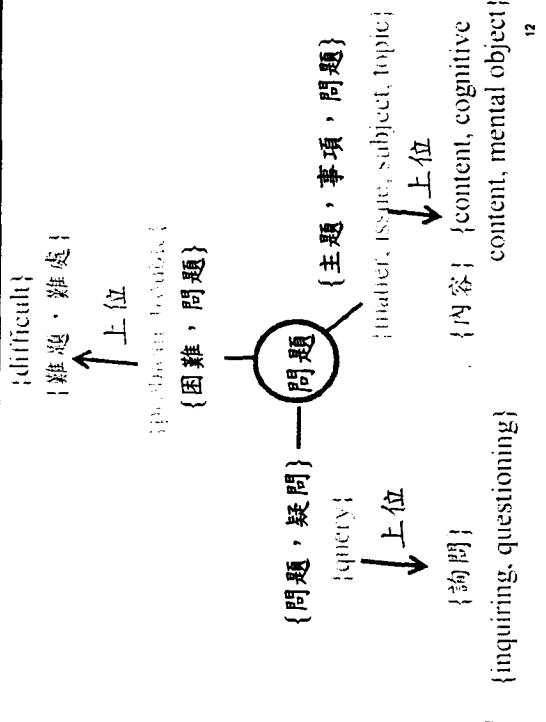
- 語意關係

同義關係 -- 67% 適當率

反義關係 -- 75% 適當率

上位關係 -- 55% 適當率

11



12

附件四

國科會國際數位圖書館合作研究計畫(IDLP) II

子計畫三

數位圖書館中文詮釋資料系統之研究

報告人

陳昭珍

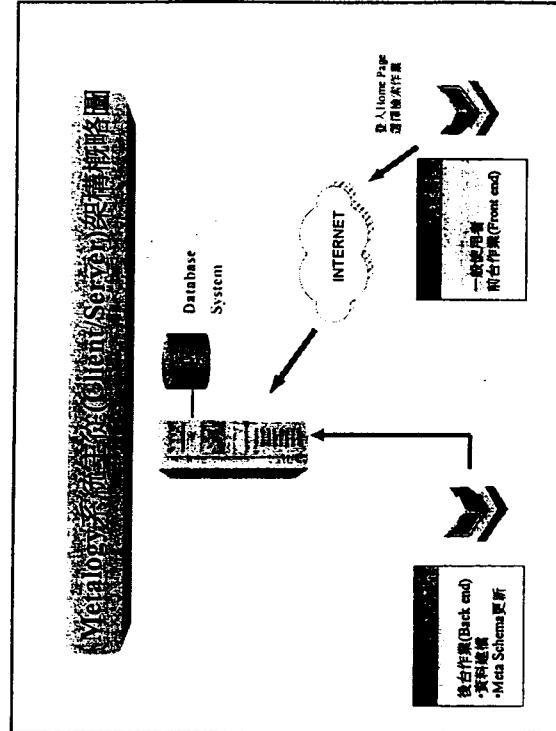
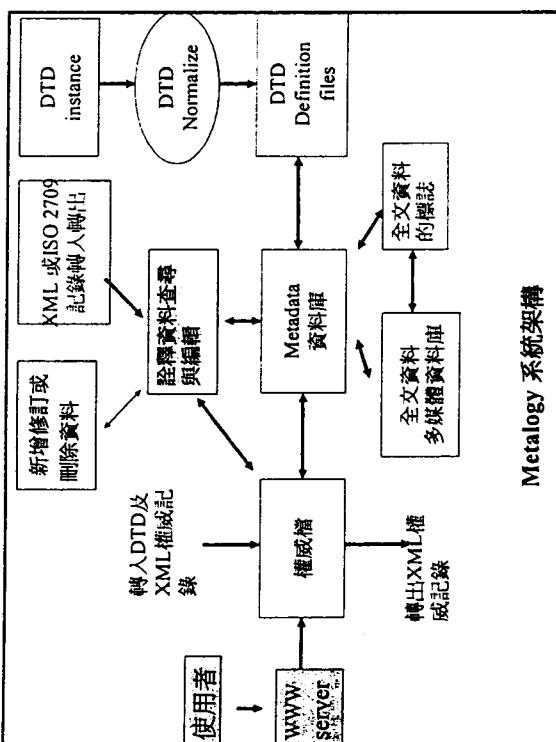
台灣師範大學圖書資訊學研究所

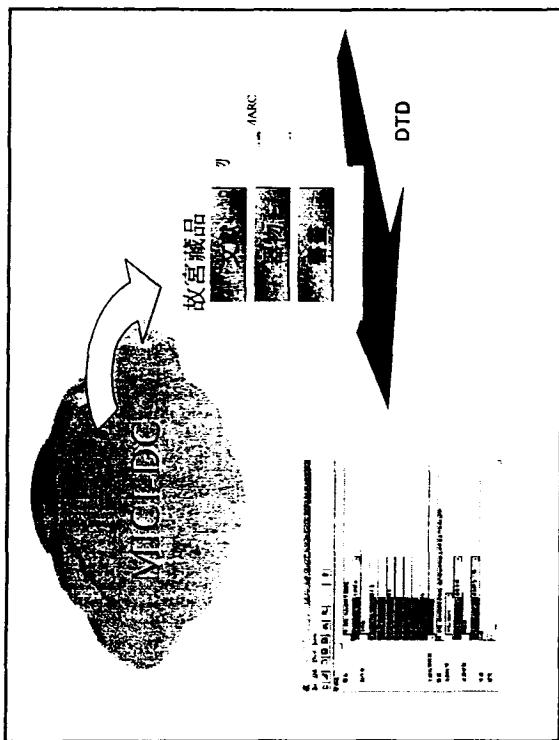
XML/Metadata管理系統

報告人：陳昭珍
國立臺灣師範大學圖書資訊學研究所副教授

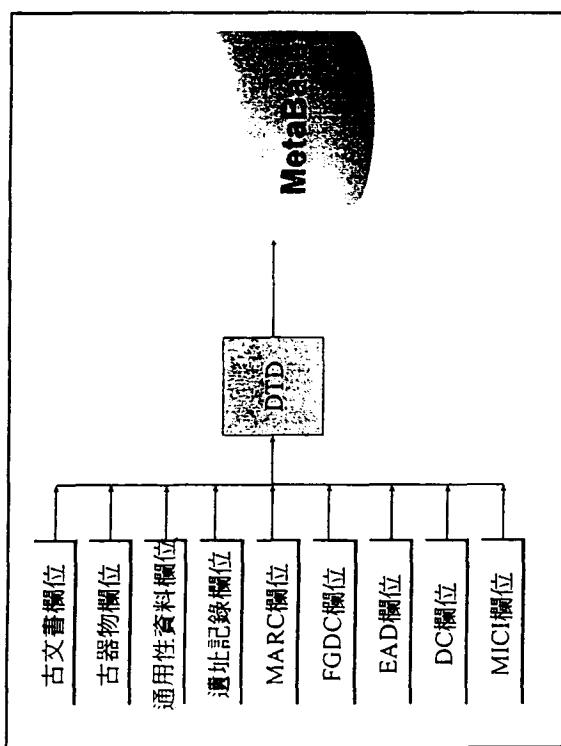
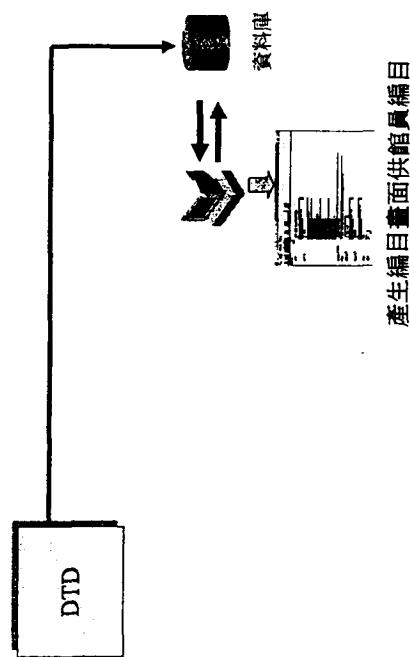
通用系統 -- Metalogy 系統特色

- 可支援各種DTD，亦即可支援各種Metadata
- 可轉入及轉出XML記錄語法之Metadata
- 當系統新增一個XML DTD 後，它會自動轉成內部的資料庫結構，並產生輸入介面
- 從DTD、資料庫綱要表、要素索引的欄位、檢索點的訂定、到顯示欄位都有親和性的介面讓使用者自行訂定
- 可連結全文、多媒體資料
- 可標誌全文做為電子出版的平台
- 可查詢系統內一個資料檔，也可同時查詢多個資料檔
- 具有權威控制的功能
- 具有整合檢索功能

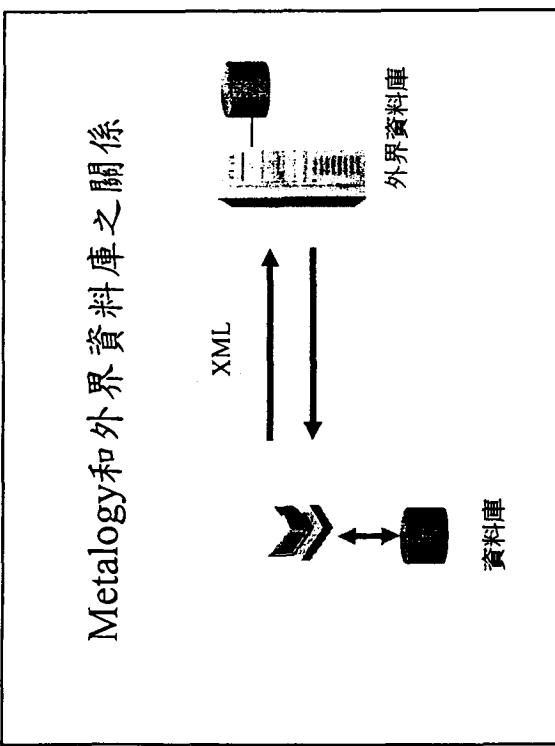


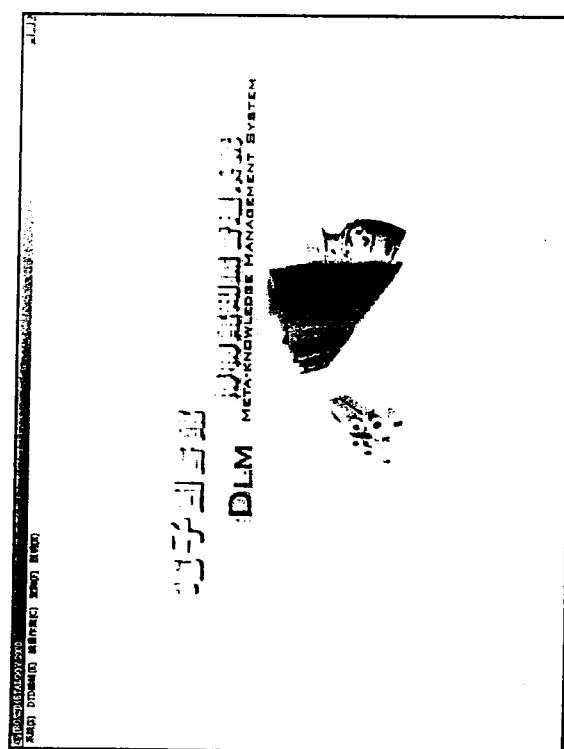
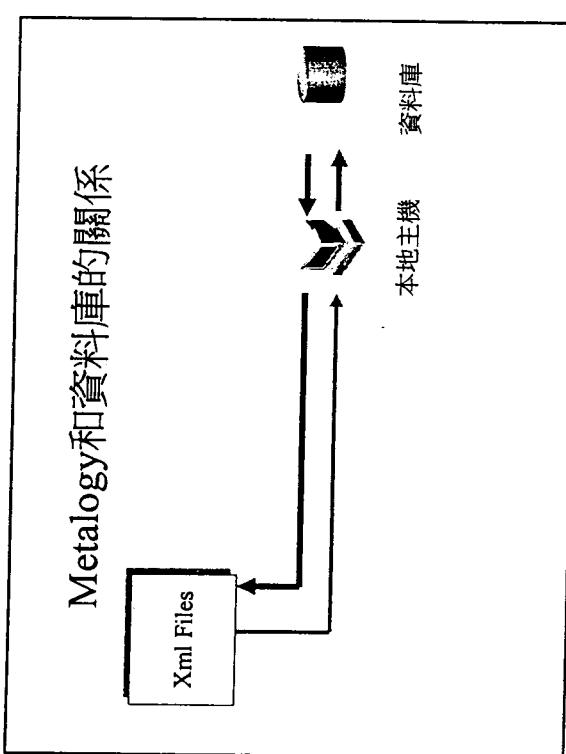
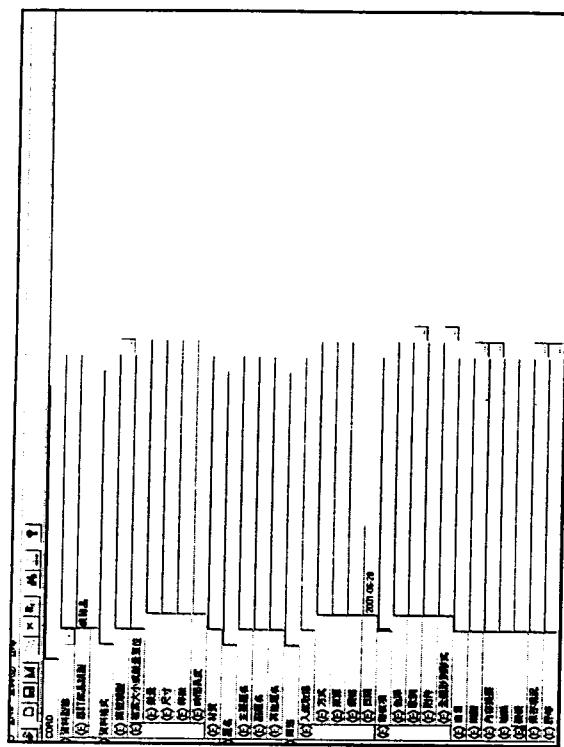
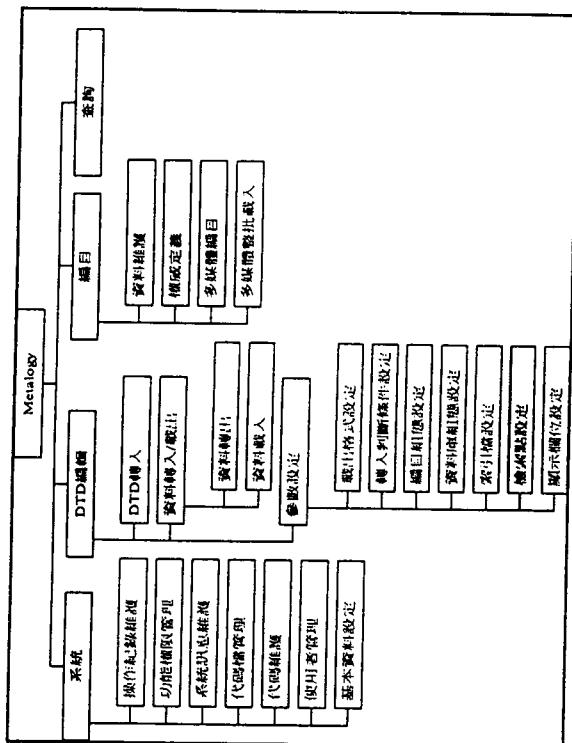


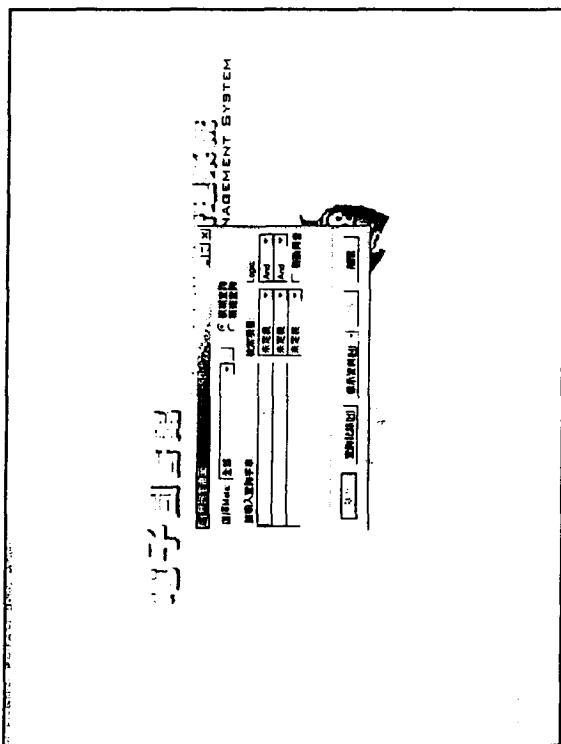
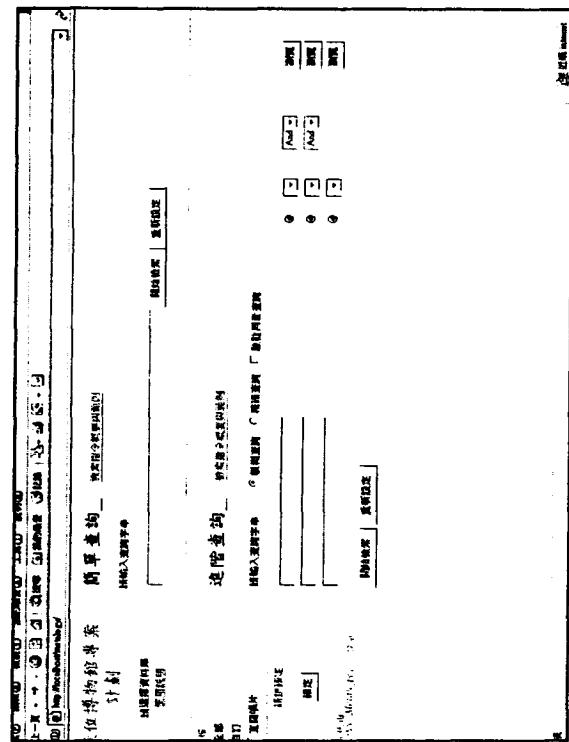
Metalogy和資料庫的關係



Metalogy和外界資料庫之關係







系統需求

- MS-SQL 版
 - SQL server
 - Metalogy應用系統
- Oracle版
 - Oracle server
 - Metalogy應用系統
- Source Code open

本年發展重點

- Multi-tier system
- Distributed Processing System- SOAP
- XML Schema
- XSL presentation
- Hierarchical metadata management system
- Support OAI Protocol -- metadata harvesting system

附件五

國科會國際數位圖書館合作研究計畫(IDLP) II 子計畫四 以代理人為基礎的數位圖書館資訊搜集與服務

報告人

蘇豐文

新竹清華大學資訊工程學系

以代理人為基礎的 數位圖書館的 資訊搜集與服務

清華大學資工系（新竹）
蘇 豐文 教授

Introduction

- The subproject II "intelligent agents on internet" in a large scale program for Promotional Activities of universities funded by MOE(NSC) in Taiwan.
- The IDLP (in collaboration with CMNED) on "intelligent agents for digital libraries" funded by NSC in Taiwan.

Research Objectives

- Develop intelligent agent techniques to support applications in domains of e-business , digital libraries, e-learning, e-government, etc.
 - Multitagent coordination, human/agent dialogue and interactions, game theoretic negotiation, resource allocations, auctions, teamwork, etc.
 - Information discovery, retrieval and gathering, user profiling, recommendation for personalized information service.

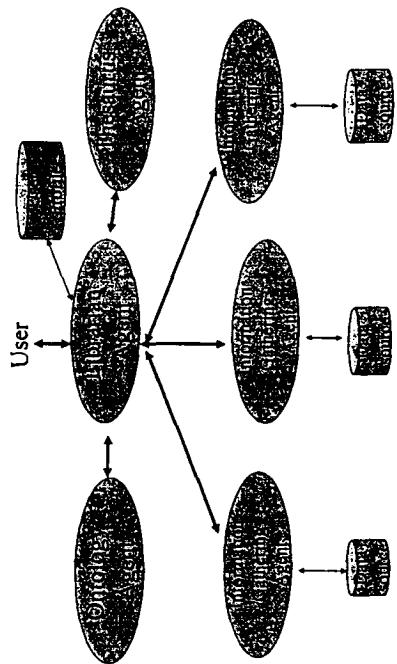
Our View on Digital Libraries

- Not just a passive archive of various kinds of data and documents,
- But active and dynamic information service centers.
- However, the roles and functions of libraries are very complex,
- So a multi-agent approach of modeling digital library services could prove to be very useful

MADL_{IR} – Multiagent Digital Libraries for Information Retrieval

- ◆ Librarian agents
- ◆ Thesaurus agents
- ◆ Ontology agents
- ◆ Information gathering agents

The MADL_{IR} Architecture



Librarian Agent

- ◆ decompose the problem solving process into several subtasks
- ◆ delegate the sub-tasks to other agents depending on the information retrieval status
- ◆ coordinate the communication interactions among agents
- ◆ interface with human users and maintain a user profile

Thesaurus Agent

- ◆ has a common and several domain specific dictionaries
 - Chinese keywords are semantically categorized in a hierarchical manner
- ◆ conduct generalization, shift, or specialization search over keyword concepts

Ontology Agent

- ◆ has the ontology (domain models) in the application domain.
 - ontology is a collection of keyword concepts and their relationships
- ◆ perform generalization, shift and specialization search over inter-relations among keyword concepts
- ◆ make heuristic inferences over domain schemas

Information Gathering Agents

- ◆ convert user's query to the right extraction schema
- ◆ extract information from a specific information source (can utilize existing search engines)

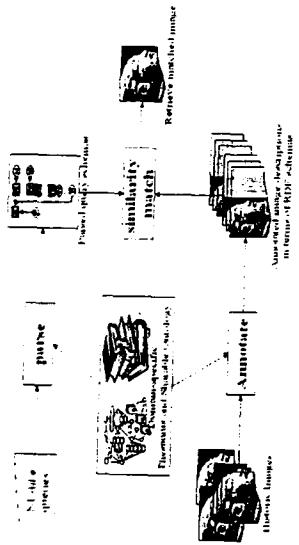
IDLP CMNET Challenges

- ◆ Extend the MADL_{IR} systems to the information retrieval of Chinese historical and cultural multi-media domain.
- ◆ The CD-ROM of First Emperor of China by Dr. Chen Ching-Chih provides a suitable and challenging testbed.

Challenging Issues

- ◆ Could we help naive users to retrieve pictures in the CD-ROM using the state of the art of intelligent agent technologies?
 - people who have little knowledge on Chinese history
- ◆ How to collect the domain specific terminologies and insert properly into entries in the Thesaurus?
- ◆ What are the reliable and essential subset of historical ontology we need to build?
 - ◆ How to convert the historical knowledge into machine readable form?
- ◆ Semantic Web ...

Processes of Ontology-based Image Retrieval



Processes of Ontology-based Image Retrieval

- 1) design tools to annotate historical images using domain specific thesaurus and sharable ontology,
- 2) design a query parser to parse natural language queries with domain specific thesaurus and ontology into query schemas,
- 3) heuristically match the query schemas with the annotated schemas of images and find out the best fit image as the result.

Augmentation of Thesaurus

- Attach domain specific terms to the right concept hierarchy and semantic category of Chinese Thesaurus:
 - Names of historic figures: 秦始皇, 李斯, 吕不韦
 - Names of geological locations: 黄河, 渭水, 长江, 泰山
 - Names of historic time periods: 秦二世元年, 西元 ...
 - Names of historical social and political systems: 秦, 韩, 赵, 魏, 燕, 齐, 邶鄖
 - Names of historic objects: 青銅劍 (bronze sword), 銅車馬 (bronze chariot), 兵馬俑 (Terracotta warriors and horses)

The Chinese Thesaurus

- (See Demo)

Augmentation of Ontology

- ◆ Temporal relations: [event1 before, within or after event2]
秦昭王四十八年 = 259BC
 - ◆ Geological relations:[orientation, territories, locations of countries, rivers, mountains, etc.]
 - ◆ Functional relations among objects:
[compositions, functions, e.g., fruit plate, fruit pie, fruit wine, fruit fly, etc.]
 - ◆ Relations in social and political systems:[capitals or cities of a country, role a person play in the system, etc]
 - ◆ Events: [Who, When, Where, What, How, Why]

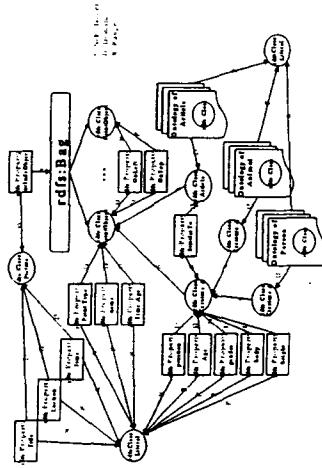
Sharable Domain Ontology

- ◆ Represent domain ontology in terms of RDF and RDF Schema of W3C Semantic Web.
 - ◆ RDF is an XML extension that represents a semantic concept in terms of a set of Subject-Property-Value tuples

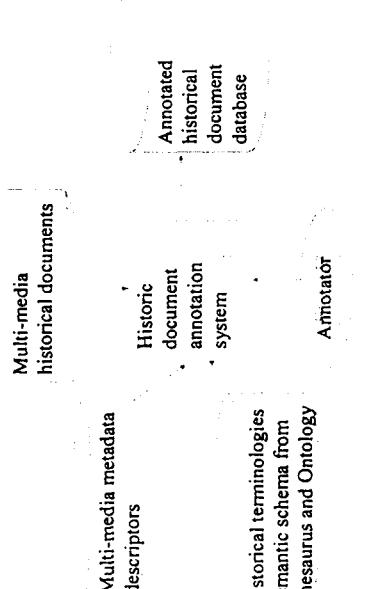
Sharable Ontology for Terracotta Soldier Domain

- ◆ ..\論文\ICDL2002\RDF files\Pic RDF files\rdfs.txt

A Snapshot of Sharable Ontology



An Annotation System for Historical Multi-media Resources



Annotated Images



Query and RDF

Query: 穿戰袍的射手

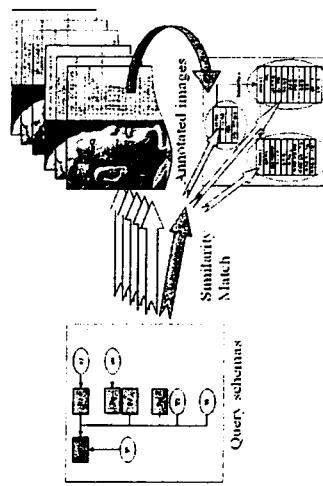
穿	FA1801	FA2901	HF0602
戰袍	BS01		
射手	AE1033		

((Property 穿) (Value 戰袍) (Subject 射手))

Query Parsing

- “秦代穿著盔甲的將軍” → RDF query Schema
- Parsing Result:
 - [將軍] Subject
 - [穿著] Property
 - [盔甲] Value
 - [時代] Property
 - [秦代] Value

Heuristic structural match of a query schema with annotated images



Performance evaluation

- ◆ (Under development)
 - Experimenting on 65 queries with 29 annotated pictures under testing
 - Recall and Precision measure

Conclusions

- ◆ Intelligent agents will play important roles in future digital library services.
- ◆ For intelligent agents to provide services, utilization of domain knowledge is a basis.
- ◆ We have shown a case study of how to use the sharable knowledge of domain ontology and thesaurus to retrieve historic pictures.
- ◆ To build a sharable ontology and thesaurus, the international, multi-cultural, multi-disciplinary collaboration is a must.

附件六

國科會國際數位圖書館合作研究計畫(IDLP) II
子計畫五
智慧財產權保護機制研究：
數位圖書館內涵保護及電子版權管理

報告人

黃世昆
中研院資訊科學研究所

Digital Content Protection Platform

- a rights management system
數位產權管理系統設置

黃世仁、何建明、廖弘源、呂俊賢
3/31-4/8 2002

簡報大綱

- 動機
- 數位產權管理背景與相關問題
- 內涵保護平臺設計
- 中文語言相關問題與國際合作
- 結論

動機

- 數位博物館產權管理（Rights Management for Digital Library）
- 針對數位內涵(Digital Content)之智慧財產權保護，發展相關技術、工具與管理程序。

Digital Rights Management 的優勢

- Content Protection 的迫切需求
- 數位典藏與 E-learning 已成為各界支持的重點計畫
- Digital Rights Management 為其安全機制
- 2000 年十大新興科技之一

相關背景

- 數位內涵資料的使用關係涵蓋三種不同角色與定位，
包括
 - 內涵提供者(Content Provider)、
 - 內涵使用者(Content User)、
 - 數位產權(Digital Rights)。

- 數位內涵種類
 - 資料庫資料、各種媒體資料、數位典藏、電子公文、電子書等。

- 內涵提供者可能是擁有者，也有可能是代理販售者。
- 數位產權管理(Rights Management)將規範這三者的關係，包括產權的規範、轉移、交易、與侵權處理。

數位資料的智慧財產權保護

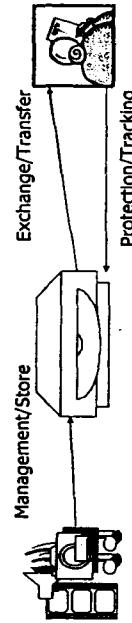
- 對於不同媒體有不同作法
 - 以數位化形式呈現傳統印刷文件 - 電子版權維護 (Electronic Copyright Protection)，
 - 這是未來電子書交易模式的重要保障，也是B-to-C能否成功推展的成功因素，各大主要公司紛紛投入制订相關協定、開發相關平台。
 - 一般媒體資料，如影像、視訊、音樂等 - 強調維護擁有者資訊，
 - 以隱藏數位浮水印為主要研究重點。也有甚多媒體公司、媒體公司、研究單位等，針對上述媒體資料之智慧財產權保護而建立管理制度。

Content Protection Issues

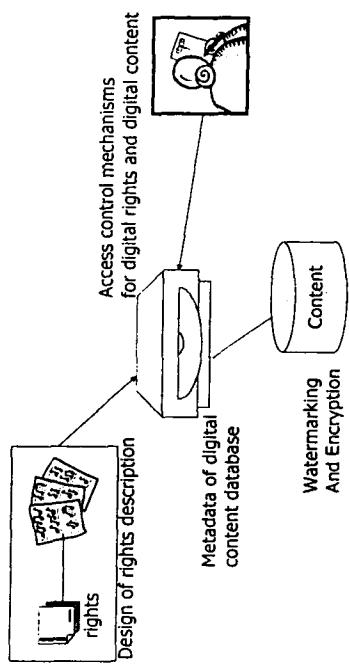
- Rights Protection and Tracking Protocols
- Rights Description
- Reader
- Watermark

Digital Rights Management

- Rights Description and Backend Management
- Digital Content Exchange and Transfer
- Digital Content Protection and Tracking



Rights Description and Backend Store Management



Protection and Tracking

- Encryption/decryption for multimedia data protection
- watermarking and fingerprinting for media tracking
- software tamper resistance

Rights Description

```

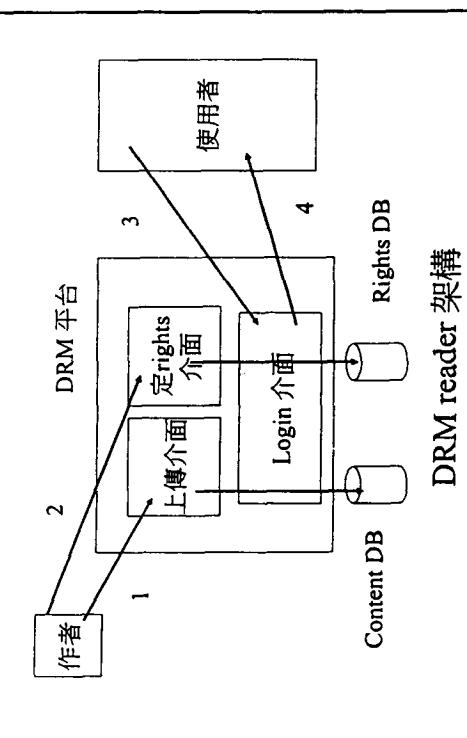
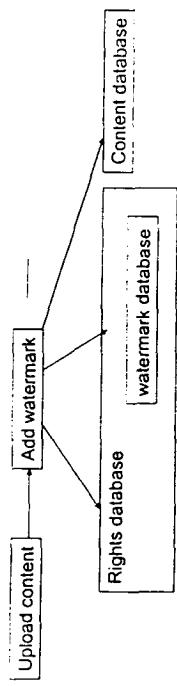
<RIGHTGROUP>
  <BUNDLE>
    <TIME> <UNTIL>2001/11/30</UNTIL>
    </TIME>
  </BUNDLE>
  <RIGHTGROUP>
    <BODY>
      <SIGNATURE>
        <DIGEST>
          <ALGORITHM>MD5</ALGORITHM>
          <VALUE encoding="base64"
size="160">+UZI0SS+U75saKKragDnlg==</VALUE>
        </DIGEST>
      </SIGNATURE>
    </BODY>
  </RIGHTGROUP>
</RIGHTGROUP>
  
```

Technology support - XrML

- Rights description : XrML^[2]
- XrML : eXtensible rights Markup Language
- Language in XML for describing specification of rights, fee, and usage

Management Flow

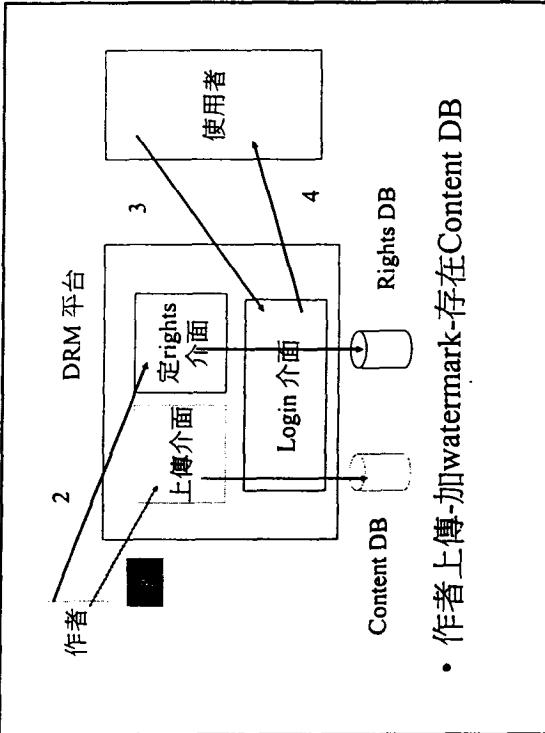
- Batch watermark adding interface^[3]
- Watermark database management

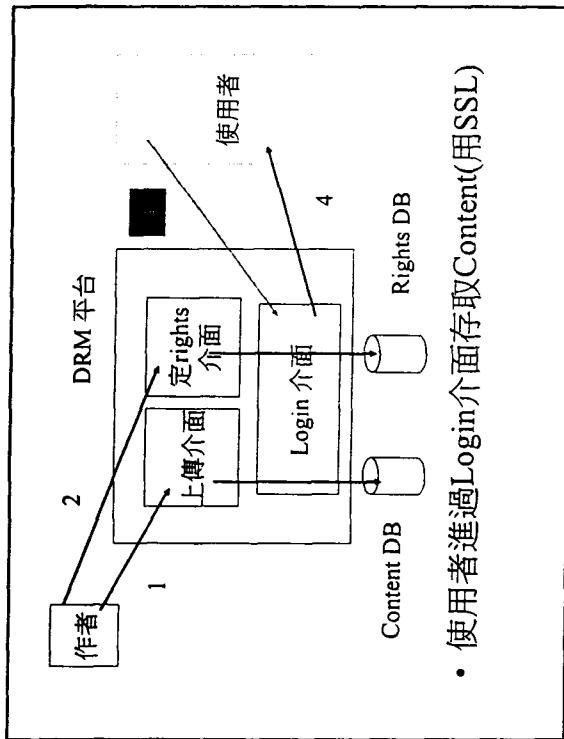


DRM reader 架構

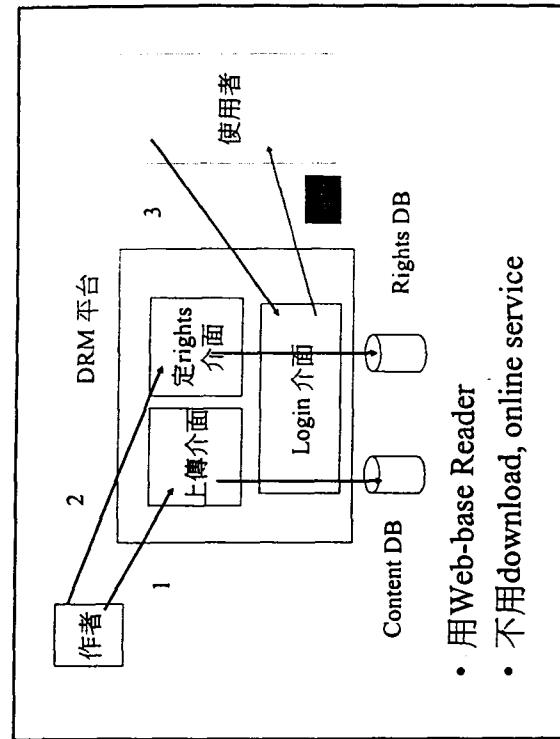
Reader 設計

- 作者上傳-加watermark-存在Content DB

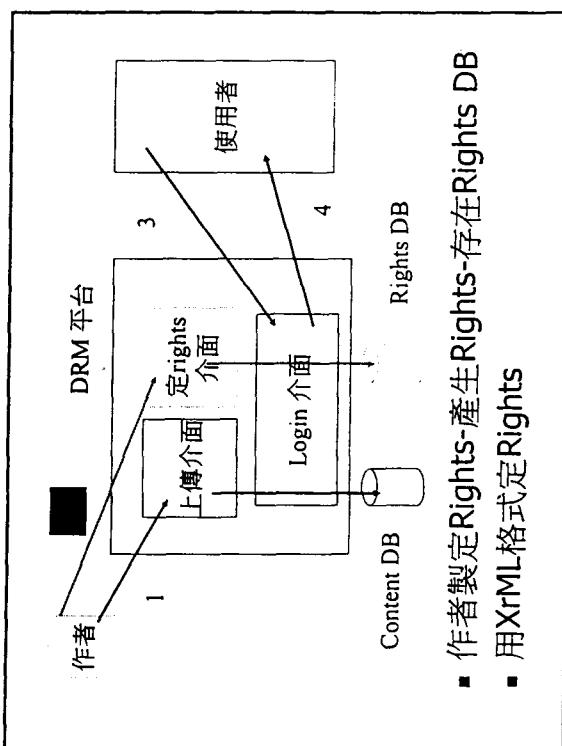




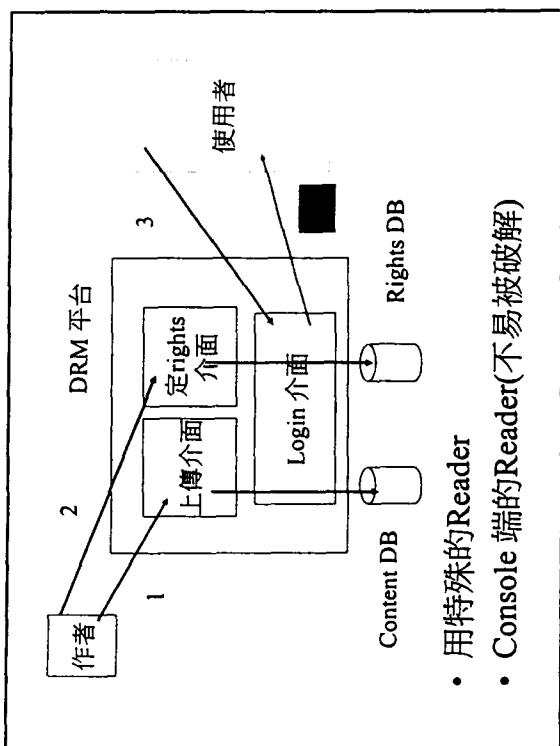
- 使用者逕過 Login 介面存取 Content(用 SSL)



- 用 Web-base Reader
- 不用 download, online service

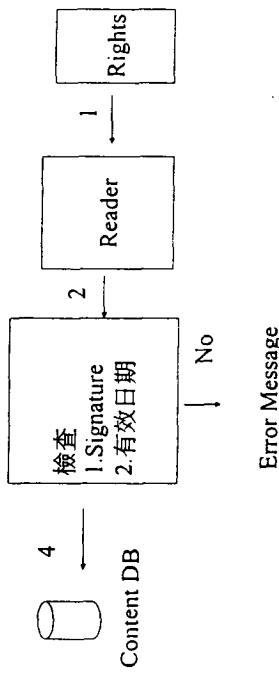


- 作者製定 Rights-產生 Rights-存在 Rights DB
- 用 XML 格式定 Rights



- 用特殊的 Reader
- Console 端的 Reader(不易被破解)

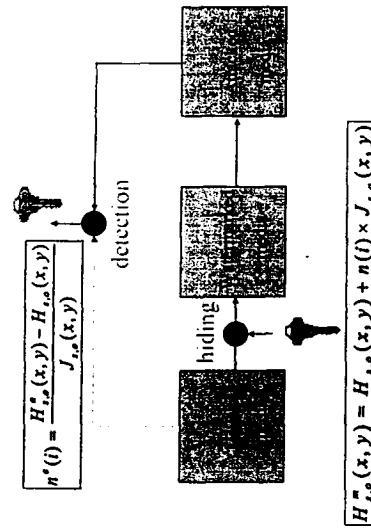
Reader 讀取Content的流程(圖示)



Copyright Marking

- Robust Watermarking
 - multimedia watermarking
 - ownership protection
 - fingerprinting
 - tracing illegal copies
 - Fragile watermarking
 - tamper detection, content authentication

A General Watermarking Scheme



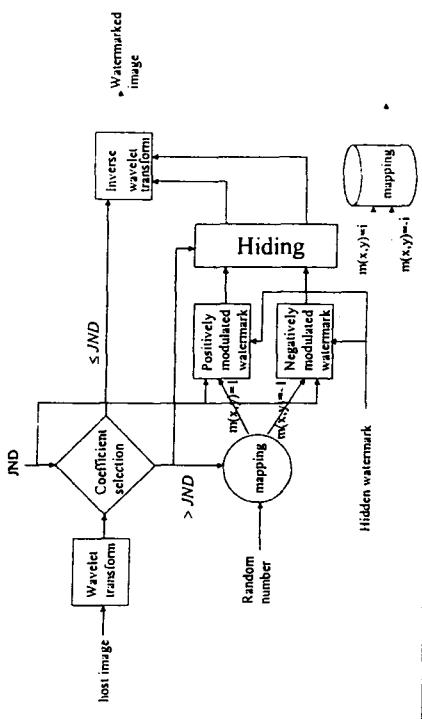
Requirements of Watermarking

- Transparency
- Undetectability
- Robustness**
- Public detection**
- Universality
- Capacity
- Registration
- Non-invertibility
- ...

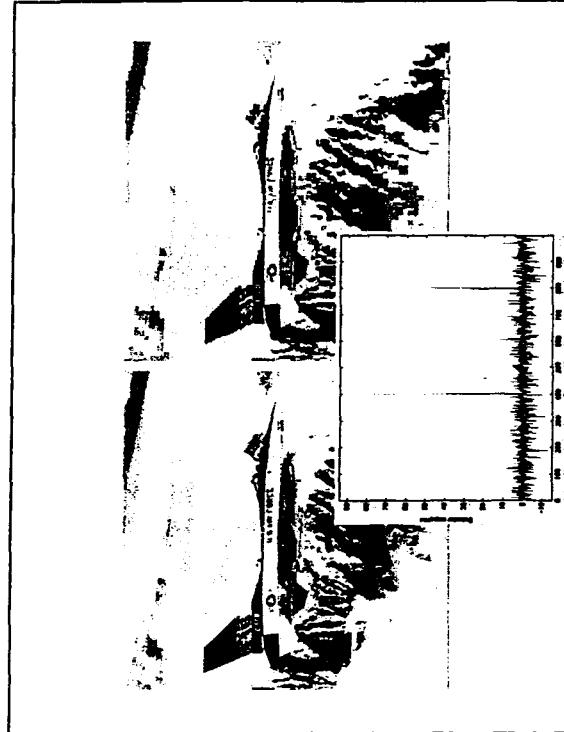
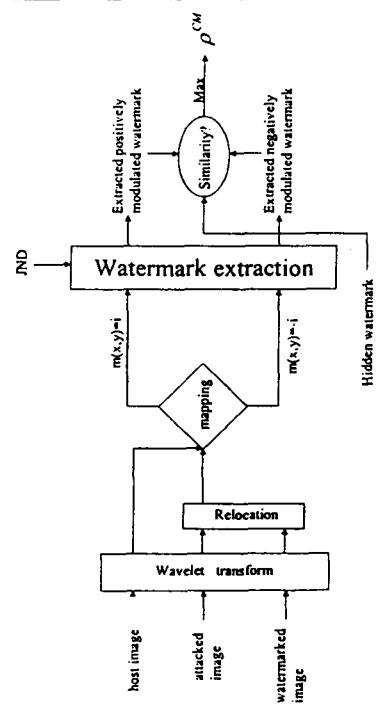
Cocktail Watermarking

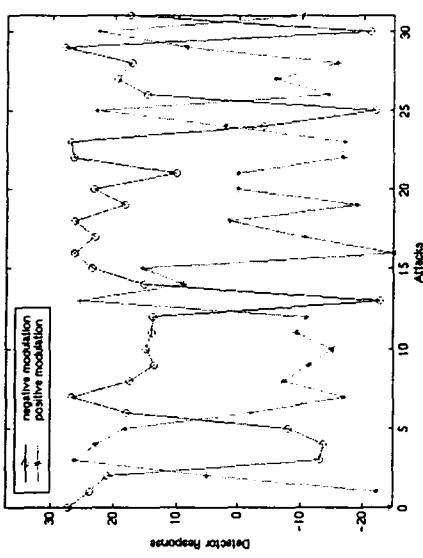
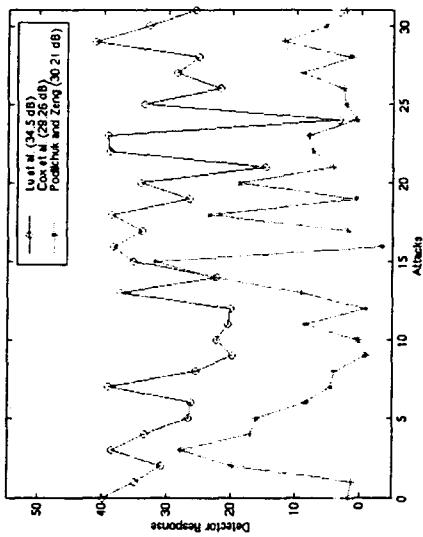
- We have observed that an attack tends to
 - increase or decrease the transformed coefficients of a media
 - Positive modulation
 - increase the magnitude of transformed coefficients
 - Negative modulation
 - decrease the magnitude of transformed coefficients

Watermark Encoder



Watermark Decoder





各種浮水印與暗藏保護資訊

- 資料庫
 - 加入雜散之 dummy record
 - 多媒體資料
- Media protection for Image, video, speech, music
- 電子公文
 - 追查洩漏源
 - 保護與管制閱讀
- 電子書
 - 追查非法散佈者
 - 保護與管制閱讀、版權控制、收費

侵權管理與用戶端 fingerprint

- 侵權管理方面整合中研院資訊所研發之互補式訊息為主之數位浮水印技術[6,7,8,9]，可容納64bits文字訊息資料，有足夠空間隱藏使用者或提供者的數位指紋。

數位內涵保護平台 (Content Protection Platform)

- 整合多種產權管理機制的系統，
 - 包括利用 XML 規範內涵提供者與使用者之間的產權關係
 - 提供有效的使用執照(license)；
 - 利用電子書交換與版權交易協定處理產權轉移與交易
 - 利用浮水印與數位指紋概念保障數位產權，以進行事後侵權認定處理，不同的內涵有不同的侵權處理方式：
 - 數位典藏著重於保障原創作者，
 - 電子公文則在追查不當流出管道，
 - 電子書將以內涵使用者為追訴重點。

Rights Protection Services

- <http://ocelot.iis.sinica.edu.tw/dirm/> (離形系統)
 - 簡易授權：上傳檔案、新增浮水印(單一或批次)
 - 認證授權：註冊資料、與浮水印連結
 - 網頁檢驗授權：以網頁 URL 為授權及授權單位進行數位產權保護。
- 利用 XML 描述與交換 media 與被授權者版權憑證資料 (Rights Center)，可應用於多授權中心與多數位內涵主機(Multiple Content Providers)

結語

- 認證授權與內涵伺服分離架構，提供各類型應用，如：
- 數位博物館系統：數位內涵之「版權」資訊必須具擴充性，如整合收費機制。
- 電子公文：電子公文可視為一種數位內涵，但收受者之間的「權限」關係複雜。
- 電子書：包括各種藝術創作都能以電子書的形式呈現，必須維護合法讀取等交易過程。

中文語言相關問題所引起之際際合作

- Text-based watermark
 - 格式編排相依之數位浮水印
 - 語意相依之浮水印，如加入不影響語意之虛字
- 數位智慧財產權問題不分國界
 - 產權管理中心有文化差異問題，權利後設資料 (rights meta-data) 的訂定 (如中文契約規範)
 - 產權管理協定。

Reference

- [1] Digital Rights Management for Ebooks : Publisher Requirements Association of American Publishers, Inc.
- [2] <http://www.xml.org>
- [3] Chun-Shien Lu, Hong-Yuan Mark Liao, and Martin Kutter, "Denoising and Copy Attacks Resilient Watermarking by Exploiting Prior Knowledge at Detector"
- [4] <http://www.ebxwg.org>
- [5] Masayuki Terada, Hiroshi Kuno, Masayuki Hanadate, Ko Fujimura, "Copy Prevention Scheme For Rights Trading Infrastructure"

References

- [6] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. Liao, "Cocktail Watermarking for Digital Image Protection," IEEE Transactions on Multimedia, December 2000, Vol 2., No. 4, pp. 209-224.
- [7] C. S. Lu, H. Y. Mark Liao, S. K. Huang, and C. J. Sze, "Cocktail Watermarking on Images," Proc. 3rd International workshop on Information Hiding, Dresden, Germany, Nov. 1999, Lecture Notes in Computer Science.
- [8] C. S. Lu, H. Y. Mark Liao, S. K. Huang, and C. J. Sze, "Highly Robust Image Watermarking Using Complementary Modulations," in Proc. 2nd Information Security Workshop, Malaysia, March 2000. Lecture Notes in Computer Science.
- [9] Gwo-Jong Yu , Chun-Shien Lu , Hong-Yuan Mark Liao , "A Message-Based Cocktail Watermarking System", Proc. 8th IEEE International Conference on Image Processing, Thessaloniki, Greece, Oct.7-10, 2001, pp. 971-974.

附件七

北京大學數字圖書館關鍵技術研究

報告人

楊冬青

北京大學計算機系

**北京大学数字图书馆关键技术研究
Peking University
Digital Library**

数字图书馆中心 (Shuwei TANG)
张海清 (Hongqing YANG) 张海清 (Hong ZHANG)
王伟华 (Weihua WANG) 王伟华 (Weihua WANG)
北京大学计算机系
Department of Computer Sci & Tech

北京大学数字图书馆研究所
Institute of Digital Library
Peking University

- 1999年9月成立
- 由CALIS管理中心、北京大学图书馆和北京大学信息科学中心联合发起并组织
- The sponsor and organizer are CALIS(China Academic Library and Information System), Peking University Library and Center for Information Science of Peking University

汇报内容 Outline

- 一、基于WebGIS的拓片检索与导航系统
 - I. WebGIS Based Retrieval and Navigation for Chinese Rubbings
- 二、科技文献导航系统
 - II. Knowledge Navigation for Scientific Documents

北京大学数字图书馆研究所(续)
Institute of Digital Library
Peking University(cont.)

- 主要工作 Major Research Work
 - 前沿研究 Initiative Research
为“数字图书馆及其相关领域指明发展方向，提出专业性的或综合性的思路和指导性的成果。
 - Putting forward professional or general thoughts and instructional achievements to direct the development of Digital Library
 - 技术研究 Technology Research
 - 实验系统:建立“数字化图书馆工程演示系统” Experimental System: To build up a Digital Library Prototype
 - 成果转换: 成熟的技术和实验成果产业化、市场化 Industrialization: To transfer the mature technology and experimental achievement into marketable products

一、基于WebGIS的拓片检索与导航系统

- 1. WebGIS Based Retrieval and Navigation for Chinese Rubbings
 - 1. 基于WebGIS的拓片检索
WebGIS based Rubbing Retrieval
 - 应用图层生成
Generation of Application Map Layers
 - 基于Web-GIS的检索
Retrieval Based on WebGIS
 - 历史地图的选择
Choose of Historical Maps
 - 概念分类层次导航式浏览
Navigation of Hierarchical Concepts

一、基于WebGIS的拓片检索与导航系统(续)

- I. WebGIS Based Retrieval and Navigation for Chinese Rubbings (Cont.)
- 2. 基于OAI-PMH的数据提供
Metadata Provider based on OAI Protocol for Metadata Harvesting

北京大学古籍数字图书馆

Peking University Rare Book Digital Library

- 中文古籍 Chinese rare books
- 拓片 Rubbings
- 古地图 Ancient atlases
- 敦煌卷宗 Dunhuang Scrolls
- 1949年以前发行的杂志
Old journals published before 1949

Metal and Stone Rubbing

金石拓片

- 一种比较特殊的中国传统文献
One kind of more special China traditional documents
- 内容丰富广泛
Contains extensively abundant contents
- 忠实地在纸质载体上再现器物的铭文图像
Inscription picture can be reproduced on the carrier paper trustfully
 - 反映不同历史时期器物的状况
Reflecting state of different historical periods

北大拓片元数据的总体部署
General Deployment of Rubbing
Metadata in Peking University

- ◆ 北京大学图书馆馆藏金石拓片近3万种，6万多份，
- Library of Peking University keeps a rubbing collection of nearly 30,000 kinds of metal and stone rubbings, more than 60,000 items.

设计元数据时的关键点 Key Points in Designing Rubbing Metadata	
<ul style="list-style-type: none"> ◆ 著录对象：三位一体，元数据需全面反映 (tombstone or others), rubbing and digital rubbing (digital image) ◆ 确定著录单位和拓片之间的关系 ◆ Making sure the description unit and relations among the rubbings ◆ 确定拓片的唯一元素 Determining unique elements for rubbings 	

拓片元数据标准 Rubbing Metadata Standard	
◆ 元数据的组成	The composition of Rubbing Metadata Standard
- 描述性元数据	Descriptive metadata
- 管理性元数据	Management metadata
- GIS元数据	GIS metadata
◆ 描述性元数据	19个 Descriptive metadata
- 核心元素，从DC借鉴的12个元素	
Core elements, which are generated based on Dublin Core with 12 elements	
- 本地核心元素 Local core elements	
- 拓片专用元数据 Unique elements for rubbing	
◆ GIS元数据	GIS metadata : 时空 space and time

拓片描述元数据 Elements of the Rubbing Description Metadata	
核心元素	本馆核心元素
Core element (12个)	Local core element (2个)
Title, 题名	Edition, 版刻/版本
Creator, 责任者	Collection history 收藏历史
Subject and Keywords, 主题关键词	Handwriting, 书法特征
Description, 内容及注释	Date, 金石刻制时间
Resource Type, 金石类型	Location, 金石刻制/出土地点
Format, 资源形式	Materials and Techniques 金石材质
Resource Identifier, 拓片标识	Original Object 原物标识
Language, 语种	Identifier 原物标识
Relation, 相关资源	Coverage: 时空范围
Coverage:	Rights Management, 传播信息

GIS 元数据 GIS Metadata

元素名 element name	子元素 sub-element	要素内容 record content
空间项 space	经度 longitude	纬度 latitude
时间项 time	中历纪年 Chinese lunar calendar year	公历纪年 Gregorian calendar year

拓片内容所涉及的地点、金石刻立、出土等地点的经维度
拓片内容所涉及的时间、金石刻立、出土等时间的公元纪年，尽可能反映到月、日

为什么使用基于Web-GIS的检索 Why We Develop Web-GIS-based Retrieval

- 拓片包括丰富的时空信息，如拓片的出土地点/时间，拓片的刻立地点/时间，拓片的时空范围
Rubbings include a lot of spatial and temporal information, such as excavated location and time, engraving location and time, and coverage
- GIS能为用户提供方便实用的检索方式
GIS can provide the user with effective and efficient retrieval method
- GIS是时空信息可视化与分析的理想工具
GIS is the perfect tool to visually display the information of such items

1. 基于Web-GIS的检索 Web-GIS-based Retrieval

- 应用图层生成
Generation of Application Map Layers
- 基于Web-GIS的检索
Retrieval Based on WebGIS
- 历史地图的选择
Choose of Historical Maps
- 概念分类层次导航式浏览
Navigation of Hierarchical Concepts

应用图层生成工具 Application Map layers Creating Tool

- 通过元数据中的时空信息（如出土时间、地点），把拓片、古籍、舆图等与图层（如城市图层）相联系。即在相应的经维度添加一个地理要素（如一个点）
Connecting rubbings, rare books and atlases with map layers (such as city map layers). That is to add a geographical feature(such as a dot) on corresponding latitude and longitude according to the spatial-temporal information in metadata, such as excavated time and location.

应用图层生成工具演示

Example of Application Map Layer Creating Tool

It can generate various map layers to meet the needs of metadata administrating requirement.

For example, creating a map layer for gravestone rubbings of Tang dynasty.

- 应用图层的生成可以保证充分利用WebGIS提供的各种查询及可视化工具

The application map layers assure the system to make full use of various query and visualization tools provided by WebGIS.

应用图层生成工具（续）

Application Map Layers Creating Tool (cont.)

- 可以生成满足元数据管理人员需求的多种图层。例如出土的唐朝墓碑的拓片图层

It can generate various map layers to meet the needs of metadata administrating requirement.

For example, creating a map layer for gravestone rubbings of Tang dynasty.

- 应用图层的生成可以保证充分利用WebGIS提供的各种查询及可视化工具

The application map layers assure the system to make full use of various query and visualization tools provided by WebGIS.

基于Web-GIS的查询功能图

Web-GIS-based Retrieval Functions

The result can be displayed on the map and in the list at the same time, and the two styles are associated (When a user chooses an item in the list, the corresponding location is highlighted.)

基于WebGIS的查询特点

Characteristics of WebGIS Retrieval

- 查询结果在一个单独的窗口显示，可以根据用户的需要放大或缩小

The result is displayed in a single window, which can be enlarged or shrunked on users' demands

- 用户只需在返回的元数据简单信息列表选择相应资源的ID，即可返回完整信息

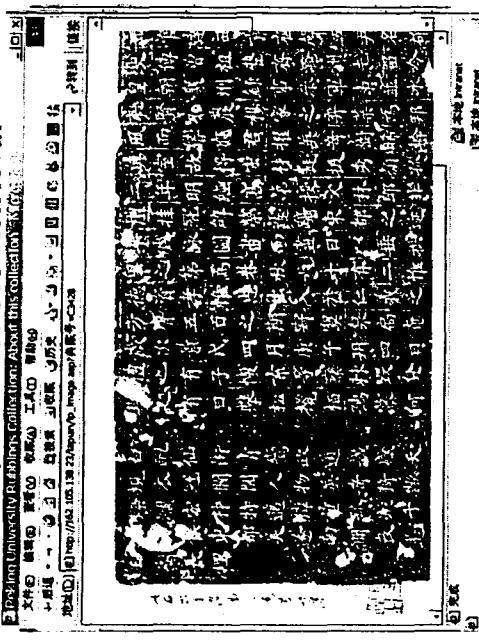
Clicking ID field in the result list, users can see the intact metadata

- 查询结果采用地图显示与列表显示两种方式，且这两种方式是联动的（当用户在列表中选择时，地图上的相应地点高亮闪烁）

The result can be displayed on the map and in the list at the same time, and the two styles are associated (When a user chooses an item in the list, the corresponding location is highlighted.)

基于WebGIS的查询演示

Example of WebGIS Retrieval



历史地图的选择 Choosing Historical Maps

用户可以用鼠标拉动时间条，系统根据历史地图与年代的对应关系，选择合适的历史地图，作为查询条件的一部分，也为结果显示的背景。

A user can use a mouse to drag the time bar, as a result the system will choose the appropriate map according to the correspondent relation between maps and historical periods.

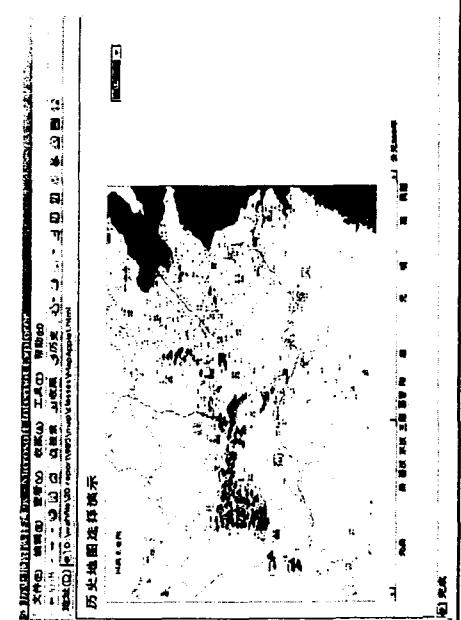
用户也可以直接在下拉式列表框中选择相应的朝代，系统调用相应的地图。

A user can also choose the dynasty in a multi-choice list, and then the system will use the appropriate map.

时间条与下拉式列表框是联动的

The two methods to choose the map are associated.

历史地图的选择演示 Example of Choosing Historical Maps



概念分类导航式浏览

Hierarchical Concept Navigation

使用概念分类，引导用户对元数据进行浏览，找到用户感兴趣的那类资源对源应的元数据，进而用户可以查找资源的详细信息

Concept hierarchy can guide a user to browse the metadata, and find the metadata that meet his interests, then to find the detailed information of the resources

使用多种概念分类层次：金石类型，朝代

A user can use various concept hierarchies, such as inscriptions types and dynasties.

方便不善于提出查询条件的用户

Browsing is convenient to those users who can not express their information needs.

2. 基于OAIMH的元数据提供 Metadata Provider based on OAI

- OAI: Open Archive Initiative

An international organization aims to provide a method to solve the interoperability problem.
- OAIMH:
 - OAI元数据采集协议
 - OAI protocol for Metadata Harvesting
 - 数据提供者与服务提供者模式: 后者通过发请求从前者获取元数据
 - It divides its participants into two classes, data provider and service provider, and the latter gets metadata from the former by submitting the requests to the former

概念分类导航式浏览演示

Example of Hierarchical Concept Navigation

The screenshot shows a hierarchical tree structure for concept navigation. The root categories are 现代 (Modern), 古代 (Ancient), 新石器时代 (Neolithic Age), and 先秦两汉 (Pre-Qin and Han). Under 现代, there are 学科 (Discipline) and 文化 (Culture). Under 古代, there are 文字 (Characters), 象形文字 (Pictographic Characters), and 其他 (Others). Under 新石器时代, there are 陶器 (Ceramics), 石器 (Stones), and 其他 (Others). Under 先秦两汉, there are 金文 (Jinwen), 铜器 (Bronze Ware), and 其他 (Others).

基于OAIMH的元数据提供 (续1) Metadata Provider based on OAIMH(cont.1)

- 数据提供者使用OAIMH中规定的格式, 返回服务提供者需要的元数据

According to the schemas that OAIMH sets down, the data provider returns the metadata that the service provider needs.

- 数据提供者必须实现对以下六种请求的响应:

The data provider must implement the responses to the following six requests.

基于OAIMH的元数据提供 (续2) OAIMH-based Metadata Provider(cont.2)

- Identify: 给出数据提供者的标识(Listing the repository's administrative information)
- ListMetadataFormats: 列出元数据格式(Listing the metadata formats that the repository supports)
- ListSets: 列出元数据的分类(Listing the grouping information of metadata in the repository.)
- ListRecords: 列出多条元数据记录(Listing multiple metadata records)
- ListIdentifiers: 列出元数据标识(Listing the identifiers of metadata records)
- GetRecord: 获得一条元数据记录(Getting one metadata record)

基于OAIMH的元数据提供演示(续3) Example of Metadata Provider(cont.3)

Identify

Please choose the OAI request you need:

- Identity Enter> [See!](#)
- ListMetadataFormats Enter> [See!](#)
- ListsSets Enter> [See!](#)
- ListIdentifiers Enter> [See!](#)
- GetRecord Enter> [See!](#)
- ListRecords Enter> [See!](#)

[HOW TO GET THE METADATA]

④ 先执行

ListRecords Verb ListRecords Verb

ListRecords

Please choose the OAI request you need:

- Identity Enter> [See!](#)
- ListMetadataFormats Enter> [See!](#)
- ListsSets Enter> [See!](#)
- ListIdentifiers Enter> [See!](#)
- GetRecord Enter> [See!](#)
- ListRecords Enter> [See!](#)

[HOW TO GET THE METADATA]

④ 先执行

Identify的使用 Identify Verb

Index of OAI

Please choose the OAI request you need:

- Identity Enter> [See!](#)
- ListMetadataFormats Enter> [See!](#)
- ListsSets Enter> [See!](#)
- ListIdentifiers Enter> [See!](#)
- GetRecord Enter> [See!](#)
- ListRecords Enter> [See!](#)

[HOW TO GET THE METADATA]

④ 先执行

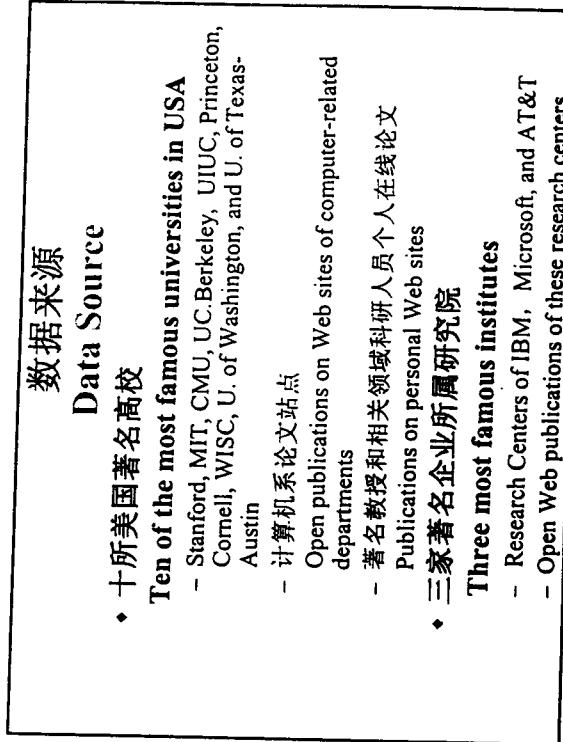
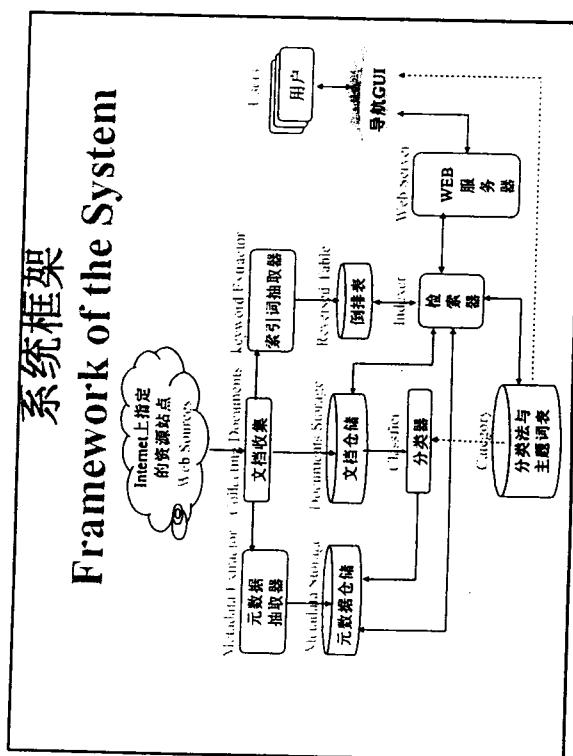
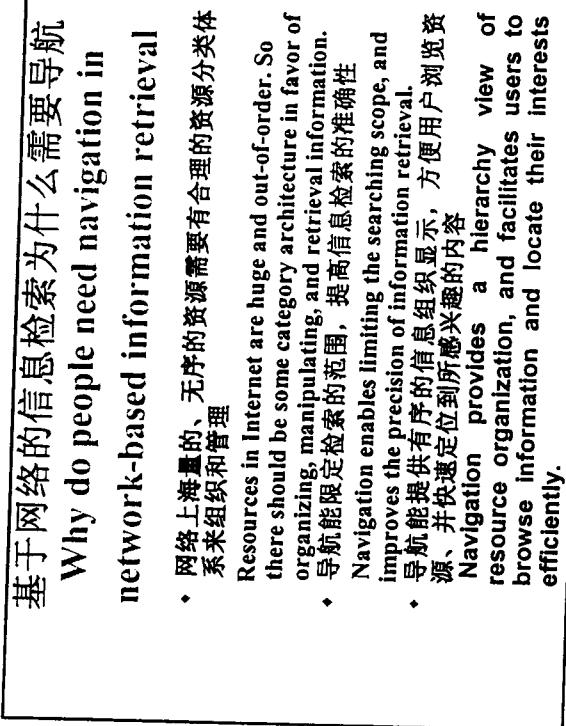
进一步的研究方向 Further Research Direction

- 实现将GIS时空检索、概念层次导航式浏览与元数据结构查询相集成的元数据检索模式
- To integrate spatial-temporal retrieval, browsing based concept hierarchy navigation and metadata structure-based query
- 时空数据挖掘算法的应用，以获取一些历史地理文化知识
- Applying spatial-temporal data mining algorithms into the system to acquire historical geographic knowledge.
- 检索结果的多种可视化显示方式
- Displaying the retrieval result in varied visualization styles
- 基于OAIMH的元数据采集
- Metadata Harvesting based on OAIMH

二. 科技文献导航系统

II. Knowledge Navigation for Scientific Documents

- 科技论文的元数据信息抽取
Metadata Extraction
 - 基于分类模型的导航
Category-based navigation
 - 导航、元数据和关键词相结合的检索手段
Document retrieval mixing navigation, metadata, and keyword



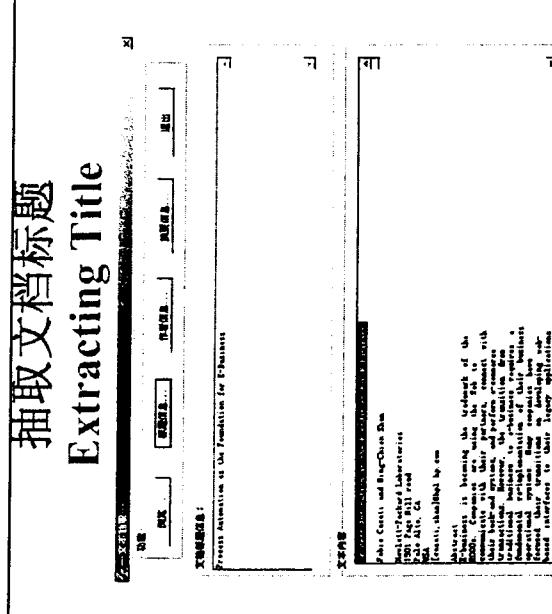
分类体系 Category

- ACM Computing Classification System
[1998 Version]

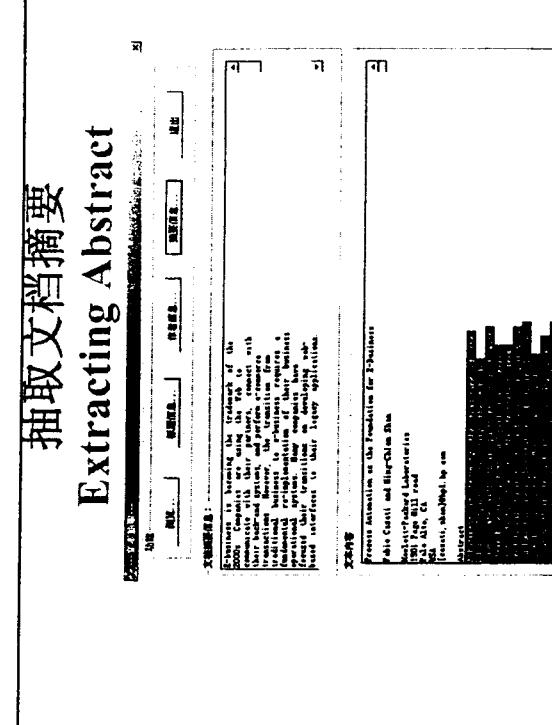
元数据信息抽取 Metadata Extraction from Documents

- PDF文档向TXT文档的转换;
Transform from PDF to TXT;
- 主要在TXT文档中抽取部分元数据，采用规则
匹配技术;
Rule-Based metadata extraction ;
- 抽取的内容包括标题信息、作者信息和摘要信
息等。
Metadata including title, author, abstract, etc.

抽取文档标题 Extracting Title



抽取文档摘要 Extracting Abstract



元数据信息抽取

Metadata Extraction from Documents

- PDF文档向TXT文档的转换;
Transform from PDF to TXT;
- 主要在TXT文档中抽取部分元数据，采用规则
匹配技术;
Rule-Based metadata extraction ;
- 抽取的内容包括标题信息、作者信息和摘要信
息等。
Metadata including title, author, abstract, etc.

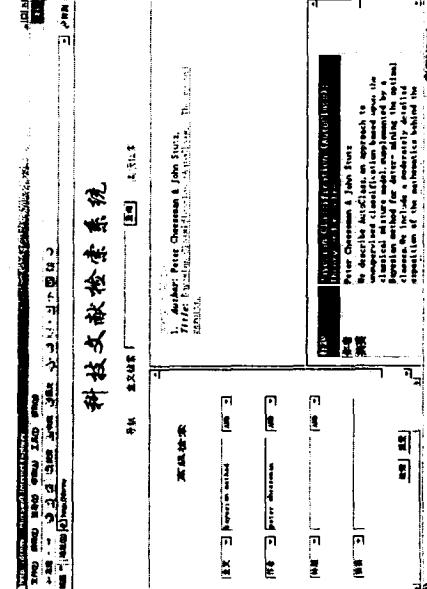
导航与检索 Navigation and Retrieval

- 采用树型层次结构显示分类，左边可选择类别，右边显示属于该类别的文档信息；
Representing category by tree in GUI;
- 提供对元数据字段的结构型检索和全文检索，满足用户对文档结构和内容的检索要求；
Providing navigation, metadata-based retrieval, and keyword-based retrieval;

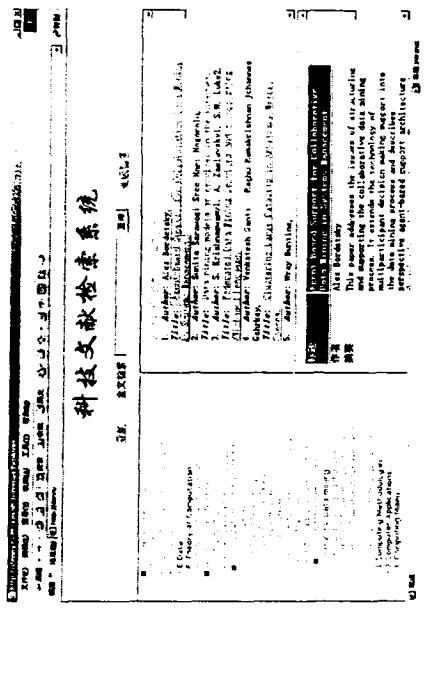
- 提供导航、元数据检索和全文检索相结合的检索模式，实现限定范围的检索。
Providing mixed retrieval.

字段检索

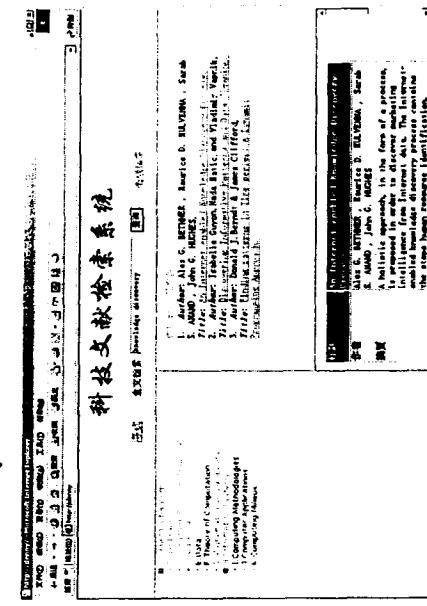
Metadata-based Retrieval



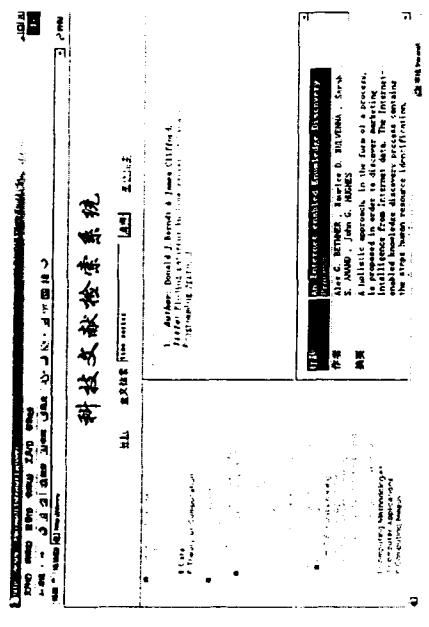
文献导航 Navigation



关键词检索 Keyword-based Retrieval



混合检索 Mixed Retrieval



深入研究 Further Research

- ◆ 资源的自动分类
Automatic resource classification
- ◆ 具体领域的应用与实践
Applications in specify domain
- ◆ 用户行为挖掘与主动服务
User behavior mining and active services
- ◆ 基于本体的信息检索技术
Ontology-based information retrieval technology

Thank you!

ENID

mzhang@db.pku.edu.cn