一種抽取並組織網路上基因相關資料的方法

作者：陳鍾誠
地址：臺北市羅斯福路 4 段一號
台灣大學資訊工程所
電話：0938707315
服務機關：台灣大學資訊工程研究所
E-mail : johnson@turing.csie.ntu.edu.tw
專長：XML 全文檢索、自然語言處理
學歷：台灣大學資訊工程研究所博士

作者：高成炎
地址：臺北市羅斯福路 4 段一號
台灣大學資訊工程所
電話：(02) 23625336 轉 509
服務機關：台灣大學資訊工程研究所
E-mail : cykao@csie.ntu.edu.tw
專長：生物資訊學
學歷：美國威斯康辛大學（麥迪生校區）計算機科學博士

中文摘要

生物基因的相關資料目前大都可由全球資訊網上取得，例如、美國國家衛生研究院的生物科技資訊中心 (NCBI) 就儲存了大量的基因相關資料，這些資料會透過網頁形式被輸出到網路上，使用者可利用瀏覽器來閱讀這些資料，然而 這些資料常常散落在各個網頁輸出介面上，無法有效整合成單一的資料倉儲，導致許多研究者使用者無法有效的查詢出想要的資料，本論文提出一個基於欄位填充機制的資料整合架構，該架構可以由網路上抓取想要的網頁，並將這些網頁整合為單一的知識架構以便瀏覽與查詢，以方便生物研究者進行研究。

關鍵字

資訊抽取、資料整合、樣稿、知識架構、生物資訊、基因

A system integrates data about gene on Web into ontology

Chung Chen Chen
johnson@turing.csie.ntu.edu.tw

Cheng Yan Kao
cykao@csie.ntu.edu.tw

Dept. of Computer Science and Information Engineering
National Taiwan University, Taiwan

Abstract

The data about gene are distributed stored in several database on Internet. For example, National Center for Biotechnology Information (NCBI) collects several large-scale data sources about gene, each data source was exported to Web by an query interface.

However, researchers on biology need an integrated data-warehouse instead of several data sources. In this paper, we propose a method to integrate several data sources into a data warehouse based on a frame-based ontology framework. A slot-filling mechanism based on the framework is developed to extract data from web pages and then organized into frames. These data are stored into a data warehouse, and a query interface was built for user to query data from the data warehouse.

Keywords
Information Extraction, Data integration, Templates, Ontology, Bioinformatics, Gene

簡介

自從全球資訊網於 1994 年開始普及之後，越來越多的基因相關資料開始公布於網路上，然而、對於生物學研究者而言，常常苦於無法有效管理並運用這些大量的資料，目前、以人工的方式從網頁上剪貼資料並加以整理常常是生物學研究者所必須做的事，當然也有些生物學家會雇用程式設計師來撰寫程式以組織這些資料。

基因的相關資訊通常儲存在數個資料庫中 每個資料庫都描述了部分的資訊，例如 NCBI 中的 Gene Bank 與 SwissProt 都包含了基因的基本資訊，另外 基因標記的相關資訊則可在 NCBI 的 UniSTS 資料庫中找到，而基因的調控路徑又可在 BioCarta 與 KEGG 等資料庫中找到，使用者通常可以經由網頁的介面對這些資料庫進行查詢，然而 生物學研究者有各式各樣不同的需求，這些網站常常無法完全符合使用者的需求。

基因相關資料的整合是生物學研究上一個重要的工具，為了有效整合基因相關資料，研究者已發展出許多方法，方法之一是建立一個查詢代理人以擷取特定網站的網頁，然後從這些網頁中抽取出特定資料，並將這些資料重新組織以方便使用者查詢與瀏覽。

資訊抽取 (Information Extraction) 是基因資料整合系統的關鍵技術，傳統的資訊抽取方法通常是用來處理純文字的自然語言格式[DeJong,1982] [Hobbs et. al, 1996]，這些方法具有相當大的彈性但卻無法完全實用，因為自然語言中包含了太多無法預期的狀況，使得這些資訊抽取的方法雖然彈性卻無法完全處理各種狀況。

全球資訊網的興起吸引了許多人開始投入了網頁資訊抽取的研究領域，大多數的研究所關心的問題是如何以自動學習的方法以學習網頁的抽取規則，然後利用這些規則以抽取出所要的資訊 這些用來抽取網頁資訊的規則被稱為“封裝器”，已有許多研究者發展出可以由一群網頁中自動學習出封裝器的方法[Kushmerick,

2000][Hsu and Dung 1998][Muslea, 1999]，然而、這些方法有下列缺點，首先是需要輸入一整群的網頁，而非單一網頁，其次是無法自動給出每個抽取欄位的名稱，必需要靠使用者為每個抽取欄位指定名稱，這對使用者而言是相當耗時的工作，這兩個缺點阻礙了資料整合系統的發展。

在本論文中，我們發展出一個可以由網頁中抽取欄位資料並加以組織的系統，該系統可用來幫助使用者管理並瀏覽所監控的網頁，該系統包含兩個組成元件 - 網頁抽取程式與樣稿填充程式，網頁抽取程式會監控使用者所指定的網站，並從網站中抓取網頁以抽取所要的資料，樣稿填充程式則將所抽取到的資料填入使用者所定義的樣稿中，以將網頁中所抽取到的資料重新組織成使用者所想要的樣子。

本系統使得生物學研究者可以方便的監控網站，並將網頁上的資料組織成所要的樣子，雖然使用者不需要自己寫程式來完成這些工作，但仍需要寫出“抽取規則”與“組織樣稿”，本系統會利用這些規則抽取出特定欄位，然後利用樣稿以組織這些欄位，事實上 這些規則與樣稿相當容易撰寫，即使是完全沒有寫過程式的生物學研究者也能輕易的完成規則與樣稿的建立工作，因此、本系統將可提供大部分的生物學研究者方便的使用。

系統架構

本系統包含一個網站監控程式 (Spider) 一個組織程式 (Organizer) 與一個暫存伺服器 (Proxy) 網站監控程式會監控特定網站並自動抓取所需要的網頁，組織程式會將這些網頁組織成使用者所想要的樣子以便瀏覽，暫存伺服器會將所抓取的網頁儲存起來以加快查詢速度，並在無法連接目標網站時做為該網站的備份之用，圖一顯示了本系統的架構。

由於組織程式是本系統中最關鍵的部分，因此本論文將會著重於組織程式所使用的方法上，該組織程式可以進一步分解成兩個子程式，包含網頁抽取程式與樣稿填充程式，網頁抽取程式利用使用者所建立的一組簡單的的抽取規則以從網頁中抽取出所要的資訊，樣稿填充程式則利用使用者所建立的樣稿，將所抽取的資料組織成為單一的網頁以便使用者瀏覽與管理，圖二顯示了該組織程式的資料流程圖。

在下一節中，我們將介紹如何從一群被監控的網頁中抽取出所要的資料，並將之組合成單一網頁的方法。

組織網頁的方法

在本節中，為了說明我們所使用的抽取與組織方法，我們使用 NCBI 網站中的 UniGene 與 LocusLink 資料庫的網頁作為範例，搭配實例與解說，以求能清楚說明本方法的理念。

假若有一個生物學研究者需要取得基因的名稱，標記與功能，因此需要整合 NCBI 中的 UniGene 與 LocusLink 資料庫，這些資料庫可由下列網頁的查詢介面取得：

UniGene 資料庫 ：http://www.ncbi.nlm.nih.gov/UniGene/
LocusLink 資料庫 ：http://www.ncbi.nlm.nih.gov/LocusLink/

若該研究者利用本系統以整合這兩個網頁，則必須先建立如圖三所示的抽取規則利用這些規則以從 UniGene 與 LocusLink 的網頁中抽取出存取編號 (Accession Number)、群組編號 (Cluster Number)、名稱 (Symbol)、標記 (Locus)、功能 (Function) 等等，然後再利用樣稿 (template) 將之組合成單一文件以便管理與瀏覽。

這些抽取規則的運作方法如下首先指定所欲抽取資料庫的網頁的存取網址，例如下列網址就指定了 UniGene 資料庫的存取方法。

http://www.ncbi.nlm.nih.gov/UniGene/query.cgi?ORG=Hs&TEXT=^ACCESSION^

其中的 ^ACCESSION^ 是一個參數，我們可將實際的存取號碼填入該網址的 ^ACCESSION^ 參數中，以取得實際的網頁，例如若我們將 T95289 填入 ^ACCESSION^ 參數中，則可取得下列網址的網頁。

http://www.ncbi.nlm.nih.gov/UniGene/query.cgi?ORG=Hs&TEXT=T59289

圖四顯示個該網頁中某些片段的內容，這個網頁是一個 HTML 形式的文件，該網頁會被送到抽取程式中以抽取指定的欄位，抽取欄位的方法是利用規則以從網頁中抽出所需要的字串，更明確的說是利用字串比對的方法以從網頁中以抽取出各個欄位的內容，以下是幾個抽取規則的範例。

Cluster^**CLUSTER**^
<B>^**SYMBOL**^</B>
LocRpt.cgi?l=^**LOCUS**^">

其中、被符號 ^ 所夾住的部份代表的是一個變數，第一條規則 Cluster^**CLUSTER**^ 可用來抽取出代表群組編號的變數^**CLUSTER**^，第二條規則<B>^**SYMBOL**^</B> 可用來抽取出代表基因名稱的變數 ^**SYMBOL**^ ，第三條規則 LocRpt.cgi?l=^**LOCUS**^"> 可以用來抽取出代表基因標記的變數 ^**LOCUS**^，這些抽取出來的欄位值會被紀錄在程式的相對應變數中等待被輸出。

本系統中所採取的欄位抽取方法是利用簡單的字串比對方法，例如 規則 Cluster^**CLUSTER**^ 可以用來與圖四網頁中的 UniGene Cluster Hs.59544 字串進行比對，比對的結果可以取得變數 ^**CLUSTER**^ **的值為** Hs.59544，同樣的方法也可以取得變數 ^SYMBOL^ 的值 ERCC1 與變數 ^**LOCUS**^ 的值 2067。

這些被抽取出來的值可以再次的被填入到網址中，以連鎖反應的方式不斷的抓取新的網頁 然後再利用該網頁的抽取規則進行欄位抽取的動作，直到沒有新的網頁被抓取回來為止，例如、當抽取規則取出 ^LOCUS^ 為 2067 後，這個值會再度被填入到 http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=^**LOCUS**^ 中以形成一個新的網址 http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=2067，然後該網址上的網頁 (一個 HTML 文件) 會被抓回來，接著再利用該網頁的抽取規則 Proteome Summary:^**FUNCTION**^</td> 以抽取欄位 ^FUNCTION^ 的值 Endonuclease with role in nucleotide excision repair; mouse ortholog is essential protein Ercc1，如此即完成了整個欄位抽取的過程。

在抽取過程完成之後，圖三中的所有變數都已具有指定的值，這些抽取出來的變數值會在樣稿填充程式中被填入一個由使用者所定義的樣稿中，以形成一個新的文件以便輸出，圖五顯示了一個樣稿的範例，在範例中、樣稿填充程式會將先前從 UniGene 與 LocusLink 兩個資料庫的網頁中所抽出來的變數值填入該樣稿中，以形成如圖六所示的輸出文件，如此即完成了整個網頁抽取與組織的過程。

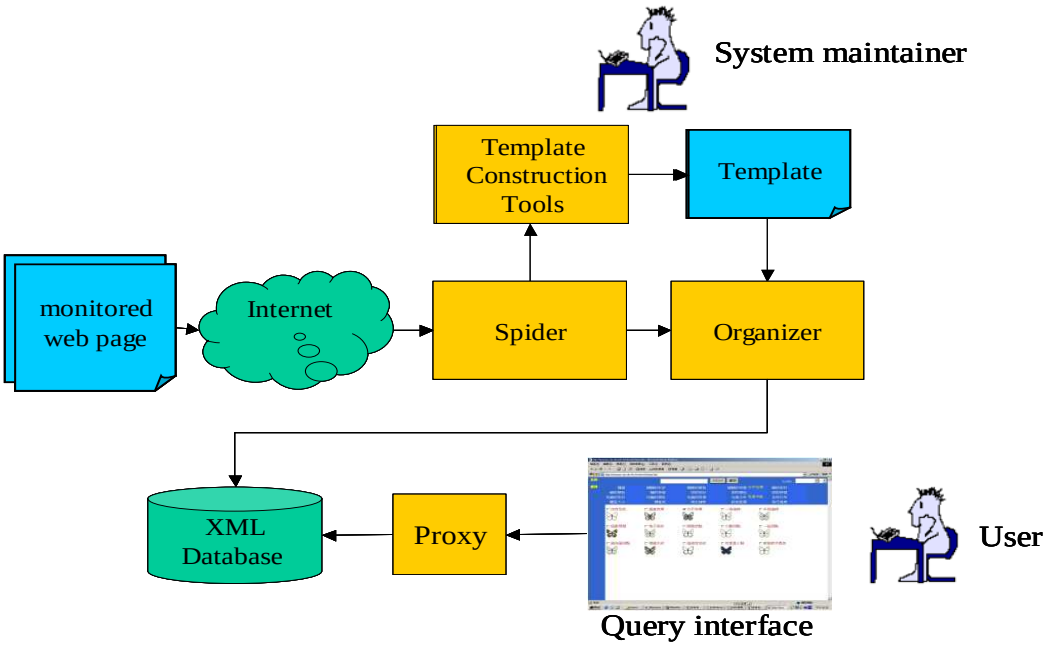
結論

本論文提出了一個基於欄位抽取與欄位填充機制的資料整合系統，其中欄位抽取的方法是利用簡單的字串比對方法，而欄位填充的方法則是利用字串取代的方法本方法雖然簡單，但是卻能有效達成資料整合的功能，降低生物學研究者在整合資料上所遭遇到的困難，本系統的使用者只要撰寫簡單的抽取規則與樣稿，即可將網路上的網頁資料重新組織，以符合使用者的需求，當然 當使用者所欲處理的網頁結構複雜時，所需寫的抽取規則也相對會較為複雜，但整體而言，本系統仍相當簡單且容易使用，未來我們將會為本系統設計一個使用方便的視窗介面，進

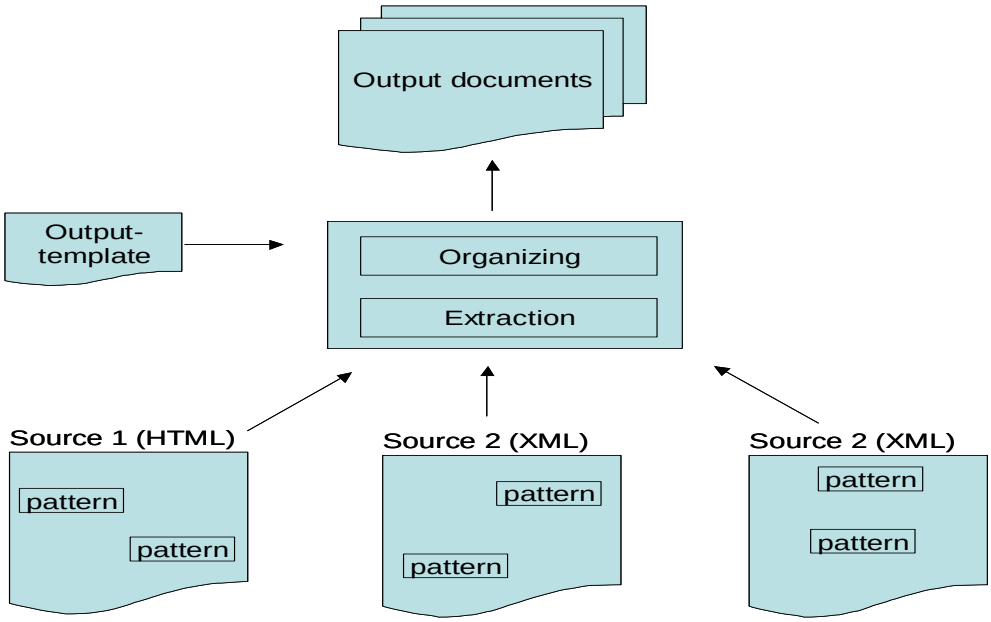
一步降低系統使用的困難度，以提供一個方便的網頁整合工具給生物學研究者使用。

參考文獻

1. DeJong; G.. ,1982 An Overview of the FRUMP System. In Strategies for Natural Language Processing, W.G.Lehnert & M.H.Ringle (Eds), Lawrence Erlbaum Associates, 1982, 149-176.
2. Hobbs, J. and Appelt, D. and Bear, J. and Israel, D. and Kameyama, M. and Stickel, M. and Tyson, M., 1996. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. In Finite State Devices for Natural Language Processing, MIT Press, 1996.
3. Hsu, C.N. and Dung, M.T., 1998 Generating finite-state transducers for semistructured data extraction from the web. Information Systems, 23(8):521-538, Special Issue on Semistructured Data, 1998.
4. Kushmerick, N. ,2000. Wrapper induction: Efficiency and expressiveness. Artificial Intelligence J. 118(1-2):15-68 (special issue on Intelligent Internet Systems).
5. Muslea, I. ,1999. Extraction Patterns for Information Tasks : A Survey. In AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.



圖一. 本資料整合系統的架構圖



圖二. 本系統中組織程式的資料流程圖

```
http://www.ncbi.nlm.nih.gov/UniGene/query.cgi?ORG=Hs&TEXT=^ACCESSION^
Cluster^CLUSTER^</b>
<B>^SYMBOL^</B>
LocRpt.cgi?l=^LOCUS^">
http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=^LOCUS^
Proteome Summary:</b>^FUNCTION^</td>
```

圖三. 用來抽取 UniGene 與 LocusLink 網頁中各欄位的規則

```
…
<b>UniGene Cluster Hs.59544</b>
<I>Homo sapiens</I><BR><br><B>ERCC1</B>
…
<A HREF="http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=2067">
…
```

圖四. UniGene 網頁中 HTML 文件的一些片段

```
<gene>
   <cluster>^CLUSTER^</cluster>
   <symbol>^SYMBOL^</symbol>
   <locus>^LOCUS^</locus>
  <function>^FUNCTION^</function>
</gene>
```

圖五. 用來整合 UniGene 與 LocusLink 的樣稿

```
<gene>
   <cluster>Hs.59544</cluster>
   <symbol>RCC1</symbol>
   <locus>2067</locus>
  <function>Endonuclease with role in nucleotide excision repair; mouse ortholog is
            essential protein Ercc1</function>
</gene>
```

圖六. 將所抽取的變數值填入樣稿後的結果