Template Mining for Information Extraction from Digital Documents

Gobinda G. Chowdhury

Abstract

With the rapid growth of digital information resources, information extraction (IE)—the process of automatically extracting information from natural language texts—is becoming more important. A number of IE systems, particularly in the areas of news/fact retrieval and in domain-specific areas, such as in chemical and patent information retrieval, have been developed in the recent past using the template mining approach that involves a natural language processing (NLP) technique to extract data directly from text if either the data and/or text surrounding the data form recognizable patterns. When text matches a template, the system extracts data according to the instructions associated with that template. This article briefly reviews template mining research. It also shows how templates are used in Web search engines—such as Alta Vista—and in meta-search engines—such as Ask Jeeves—for helping end-users generate natural language search expressions. Some potential areas of application of template mining for extraction of different kinds of information from digital documents are highlighted, and how such applications are used are indicated. It is suggested that, in order to facilitate template mining, standardization in the presentation and layout of information within digital documents has to be ensured, and this can be done by generating various templates that authors can easily download and use while preparing digital documents.

Gobinda G. Chowdhury, Division of Information Studies, School of Applied Science,
Nanyang Technological University, N4#2a—32, Nanyang Avenue, Singapore 639798
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 182-208

Information Extraction and Template Mining

Information extraction (IE), the process of automatically extracting information from natural language texts, is gaining more and more importance due to the fast growth of digital information resources. Most work on IE has emerged from research into rule-based systems in natural language processing. Croft (1995) suggested that IE techniques, primarily developed in the context of the Advanced Research Projects Agency (ARPA) Message Understanding Conferences (MUCs), are designed to identify database entities, attributes, and relationships in full text. Gaizauskas and Wilks (1998) defined IE as the activity of automatically extracting pre-specified sorts of information from short natural language texts typically, but by no means exclusively, newswire articles. Although works related to IE date back to the 1960s, perhaps the first detailed review of IE as an area of research interest in its own right was by Cowie and Lehnert (1996). However, a detailed review dividing the literature on IE into three different groups—namely, the early work on template filling, the Message Understanding Conferences (MUCs), and other works on information extraction—has recently been published by Gaizauskas and Wilks (1998).

Template mining is a particular technique used in IE. Lawson et al. (1996) defined template mining as a natural language processing (NLP) technique used to extract data directly from text if either the data and/or text surrounding the data form recognizable patterns. When text matches a template, the system extracts data according to instructions associated with that template. Although different techniques are used for information extraction and knowledge discovery—as described by Cowie and Lehnert (1996), Gaizauskas and Wilks (1998), and Vickery (1997)—template mining is probably the oldest information extraction technique. Gaizauskas and Wilks (1998) reported that templates were used to extract data from natural language texts against which "fact retrieval" could be carried out in the Linguistic String Project at New York University that began in the mid-1960s and continued into the 1980s (reported by Sager, 1981). Numerous studies have been conducted, though most of them are domain-specific, using templates for extracting information from texts. This article briefly reviews some of these works. It also shows how templates are used for information retrieval purposes in major Web search engines like AltaVista (http://www.altavista.com). This discussion proposes that template mining has great potential in extracting different kinds of information from documents in a digital library environment. To justify this proposition, this article reports some preliminary tests carried out on digital documents, more specifically on some articles published in the *D-Lib Magazine* (http://www.dilib.org/dilib).

WORKS ON TEMPLATE MINING

Template mining has been used successfully in different area:

- extraction of proper names by Coates-Stephens (1992), Wakao et al. (1996), and by Cowey and Lehnert (1996);
- extraction of facts from press releases related to company and financial information in systems like ATRANS (Lytinen & Gershman, 1986), SCISOR (Jacobs & Rau, 1990), JASPER (Andersen, et al., 1992; Andersen & Huettner, 1994), LOLITA (Costantino, Morgan, & Collingham, 1996), and FIES (Chong & Goh, 1997);
- abstracting scientific papers by Jones and Paice (1992);
- summarizing new product information by Shuldberg et al. (1993);
- extraction of data from analytical chemistry papers by Postma et al. (1990a, 1990b) and Postma and Kateman (1993);
- extraction of reaction information from experimental sections of papers in chemistry journals by Zamora and Blower (1984a,1984b);
- processing of generic and specific chemical designations from chemical patents by Chowdhury and Lynch (1992a,1992b) and by Kemp (1995); and
- extraction of bibliographic citations from the full texts of patents by Lawson et al. (1996).

Template mining has largely been used for extraction of information from news sources and from texts in a specific domain. Gaizauskas and Wilks (1998) reported that applied work on filling structured records from natural language texts originated in two long-term research projects: The Linguistic String project (Sager, 1981) at New York University and the research on language understanding and story comprehension carried out at Yale University by Schank and his associates (Schank, 1975; Schank & Abelson, 1977; Schank & Riesbeck, 1981). The first research was conducted in the medical science domain, particularly involving radiology reports and hospital discharge summaries, while the second research led to many other research works in the early 1980s that used the principles and techniques of IE to develop practical applications such as the FRUMP system developed by De Jong (1982). FRUMP used a simplified version of SCRIPTS, proposed by Schank (Schank, 1975; Schank & Abelson, 1977; Schank & Riesbeck, 1981), to process text from a newswire source to generate story summaries.

ATRANS (Lytinen & Gershman, 1986), another IE system, was soon developed and commercially applied. ATRANS used the *script* approach (Schank & Abelson, 1977; Schank & Riesbeck, 1981) for automatic processing of money transfer messages between banks. Another successful application of IE has produced a commercial online news extraction system called SCISOR (Jacobs & Rau, 1990) that extracts information about corporate mergers and acquisitions from online news sources. JASPER

(Andersen et al., 1992; Andersen & Huettner, 1994) was another IE system developed for fact extraction for Reuters. JASPER uses a template-driven approach and partial analysis techniques to extract certain key items of information from a limited range of texts such as company press releases. LOLITA (Costantino, Morgan, & Collingham, 1996) is a financial IE system that uses three pre-defined groups of templates designed according to a "financial activities approach," namely, company related templates, company restructuring templates, and general macroeconomic templates. In addition, the user-definable template allows the user to define new templates using natural language sentences. Chong and Goh (1997) developed a similar template-based financial information extraction system, called FIES, that extracts key facts from online news articles.

Applications of template mining techniques for automatic abstracting can be traced back to 1981 when Paice (1981) used what he called *indicator phrases* (such as "the results of this study imply that . . . ") to extract topics and results reported in scientific papers for generating automatic abstracts. Paice continued his work to improve on this technique and for resolving a number of issues in natural language processing (see, for example, Jones & Paice, 1992; Paice & Husk, 1987). Shuldberg et al. (1993) described a system that digests large volumes of text, filtering out irrelevant articles and distilling the remainder into templates that represent information from the articles in simple slot/filler pairs. The system consists of a series of programs each of which contributes information to the text to help determine which strings constitute appropriate values for the slots in the template.

Chemical and patent information systems have been the prominent areas for the application of templates for IE. TICA (Postma et al., 1990a, 1990b; Postma & Kateman, 1993) used templates to extract information from the abstracts of papers on inorganic tritimetric analysis. The parsing program used in TICA followed an expectation-driven approach where words or groups of words expect other words or concepts to appear. Zamora and Blower (1984a,1984b) developed a system that automatically generates reaction information forms (RIFs) from the descriptions of syntheses of organic chemicals in the *Journal of the American Chemical Society.* The techniques explored in the semantic phase of this work include the use of a *case grammar* and *frames* (Schank & Abelson, 1977; Schank & Riesbeck, 1981) to map the surface structure of the text into an internal representation from which the RIFs can be formed. Following the same methodology, Ai et al. (1990) developed a system that generates a summary of all preparative reactions from the experimental sections of the *Journal of Organic Chemistry* papers. This work identified seven sequences of events that were used for building templates for the text of an experimental paper.

Chowdhury and Lynch (1992a, 1992b) developed a template-based method for converting to GENSAL (a generic structure language developed at the University of Sheffield) those parts of the Derwent Documentation Abstracts that specify generic chemical structures. Templates for processing both the variable and multiplier expressions, which predominate in the assignment statements in the Derwent Documentation Abstracts, were identified for further processing. As part of this research, Chowdhury (1992) also conducted a preliminary discourse analysis of European chemical patents that identified the common patterns of expressions occurring in different parts of patent texts. This work prompted further research (Kemp, 1995; Lawson et al., 1996) leading to the use of template mining in the full text of chemical patents. Lawson et al. (1996) reported their work using the template mining approach to isolate and extract automatically bibliographic citations to patents, journal articles, books, and other sources from the full texts of English-language patents.

There is also some work that examines the development of specific tools and techniques for information extraction using templates. For example, Sasaki (1998) reported an ongoing project on building an information extraction system that extracts information from a real-world text corpus such as newspaper articles and Web pages. As part of this project, an inductive logic programming (ILP) system has been developed to generate IE rules from examples. Gaizauskas and Humphreys (1997) described the approach taken to knowledge representation in the LaSIE information extraction system, particularly the knowledge representation formalisms, their use in the IE task, and how the knowledge represented in them is acquired. LaSIE first translates individual sentences to a quasi logical form and then constructs a discourse model of the entire text from which template fills are derived.

Guarino (1997) argued that the task of information extraction can be seen as a problem of semantic matching between a user-defined template and a piece of information written in natural language. He further suggested that the ontological assumptions of the template need to be suitably specified and compared with the ontological implications of the text. Baralis and Psaila (1997) argued that the current approaches to data mining usually address specific user requests, while no general design criteria for the extraction of association rules are available for the end-user. To solve this problem, they have proposed a classification of association rule types that provides a general framework for the design of association rule mining applications and predefined templates as a means to capture the user specification of mining applications.

Although numerous research projects have been undertaken, and some are currently ongoing, Croft (1995) suggested that the current state of information extraction tools is such that it requires a considerable investment to build a new extraction application, and certain types of information are very difficult to identify. However, Croft further commented that extraction of simple categories of information is practical and can be an important part of a text-based information system. This article highlights some potential areas of application of template mining in a digital library environment.

Use of Templates in Web Search Engines

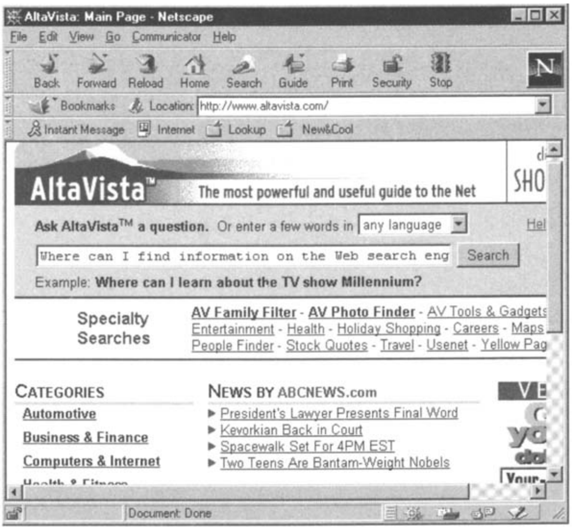
Gaizauskas and Wilks (1998) suggested that there is a contrast between the aims of information extraction and information retrieval systems in the sense that IR retrieves relevant documents from collections, while IE extracts relevant information from documents. However, the two are complementary, and their use in combination has the potential to create powerful new tools in text processing and retrieval. Indeed, IE and IR are equally important in the electronic information environment, particularly the World Wide Web, and templates have been used both for IR and IE. Many applications of template mining mentioned above handle digital texts available on the Web, while search engines use templates to facilitate IR.

Search engines are one of the most essential tools on the Internet—they help find Web sites relating to a particular subject or topic. Search engines are basically huge databases containing millions of records that include the URL of a particular Web page along with information relating to the content of the Web page supplied in the HTML by the author. A search engine obtains this information via a submission from the author or by the search engine doing a "crawl" using "robot crawlers" of the Internet for information. The most popular search engines include: AltaVista, Excite, Hotbot, Infoseek, Lycos, Webcrawler, Yahoo, and so on.

Some search engines use templates to help end-users submit natural language queries used by search engines to conduct a search on specific topics. Two small sets of tests were conducted to see how this is done in a large search engine—AltaVista—and in a meta search engine—Ask Jeeves. The following section shows how these search engines use templates for natural language query formulation in their interfaces.

Use of Templates in Alta Vista

The Alta Vista search engine (http://www.altavista.com) helps users find information on the Web. One interesting feature of this search engine is that a user can enter one or more search terms/phrases or can type a natural language statement such as "What is the capital of Alaska?" or "Where can I find quotations by Ingmar Bergman?" Taking the second option, a simple query statement, "Where can I find information on Web search engines?" was typed in the specified box of the Alta Vista search interface (see Figure 1). Along with the results, Alta Vista came up with two templates that contain natural language sentences related to the search topic (see Figure 2). By clicking on the box at the end of the statement "How do I (Internet skill)?" the system shows a box containing various options (Figure 3), any of which can be chosen to complete the sentence, the default one being "search through ALL web sites." By choosing this, or any other option from the box, a user can formulate a sentence-like query such as: "How do I search through all Web sites?" or "How do I learn HTML?" or "How do I use the Internet as a telephone?" and so on.



AltaVista: Main Page - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks
Instant Message Internet Lookup New&Cool
AltaVista™
The most powerful and useful guide to the Net
Search
Specialty Searches
AV Family Filter - AV Photo Finder - AV Tools & Gadgets
Entertainment - Health - Holiday Shopping - Careers - Maps
People Finder - Stock Quotes - Travel - Usenet - Yellow Pag
CATEGORIES
Automotive
Business & Finance
Computers & Internet
NEWS BY ABCNEWS.com
President's Lawyer Presents Final Word
Kevorkian Back in Court
Spacewalk Set For 4PM EST
Two Teens Are Bantam-Weight Nobels

Location: http://www.altavista.com/

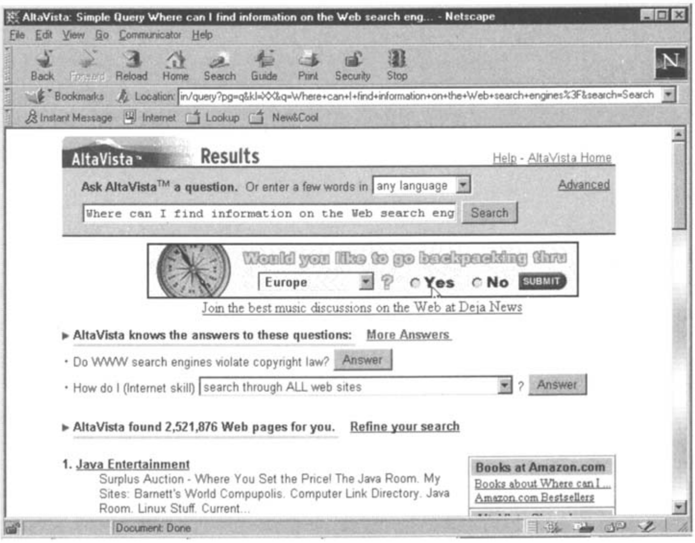
Ask AltaVista™ a question. Or enter a few words in any language

Where can I find information on the Web search eng

Example: Where can I learn about the TV show Millennium?

Document: Done

Figure 1. The Search Interface of Alta Vista.



AltaVista: Simple Query Where can I find information on the Web search eng... - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks
Instant Message Internet Lookup New&Cool
AltaVista™ Results
Help - AltaVista Home
Advanced
Search
Would you like to go backpacking thru
Europe
○ Yes ○ No SUBMIT
Join the best music discussions on the Web at Deja News
Answer
Books at Amazon.com
Books about Where can I ...
Amazon.com Bestsellers

Location: in/query?pg=q&kl=XX&q=Where+can+I+find+information+on+the+Web+search+engines%3F&search=Search

Ask AltaVista™ a question. Or enter a few words in any language

Where can I find information on the Web search eng

▸ AltaVista knows the answers to these questions: More Answers

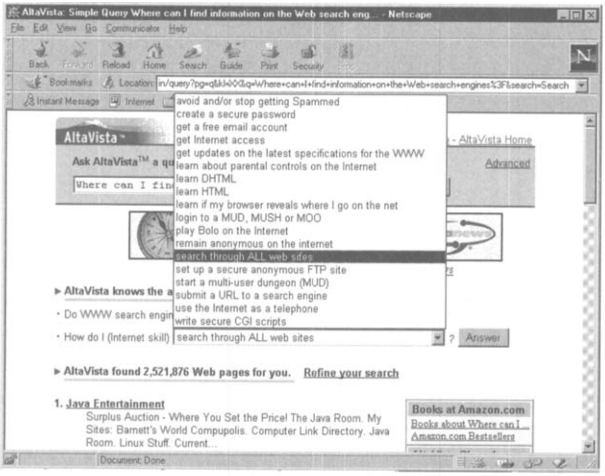
- Do WWW search engines violate copyright law? Answer
- How do I (Internet skill) search through ALL web sites ? Answer

▸ AltaVista found 2,521,876 Web pages for you. Refine your search

1. Java Entertainment
Surplus Auction - Where You Set the Price! The Java Room. My Sites: Barnett's World Compupolis. Computer Link Directory. Java Room. Linux Stuff. Current...

Document: Done

Figure 2. Output of a Simple Search.


AltaVista: Simple Query Where can I find infomation on the Web search eng... - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: in/query?pg=q&kl=XX&q=Where+can+I+find+infomation+on+the+Web+search+engines%3F&search=Search
Instant Message Internet
AltaVista
- AltaVista Home
Advanced
? Answer
Document: Done

Ask AltaVista™ a qu

Where can I fin

- avoid and/or stop getting Spammed
- create a secure password
- get a free email account
- get Internet access
- get updates on the latest specifications for the WWW
- learn about parental controls on the Internet
- learn DHTML
- learn HTML
- learn if my browser reveals where I go on the net
- login to a MUD, MUSH or MOO
- play Bolo on the Internet
- remain anonymous on the internet
- search through ALL web sites
- set up a secure anonymous FTP site
- start a multi-user dungeon (MUD)
- submit a URL to a search engine
- use the Internet as a telephone
- write secure CGI scripts

▸ AltaVista knows the a

· Do WWW search engin

· How do I (Internet skill) search through ALL web sites

▸ AltaVista found 2,521,876 Web pages for you. Refine your search

1. Java Entertainment
Surplus Auction - Where You Set the Price! The Java Room. My Sites: Barnett's World Compupolis. Computer Link Directory. Java Room. Linux Stuff. Current...

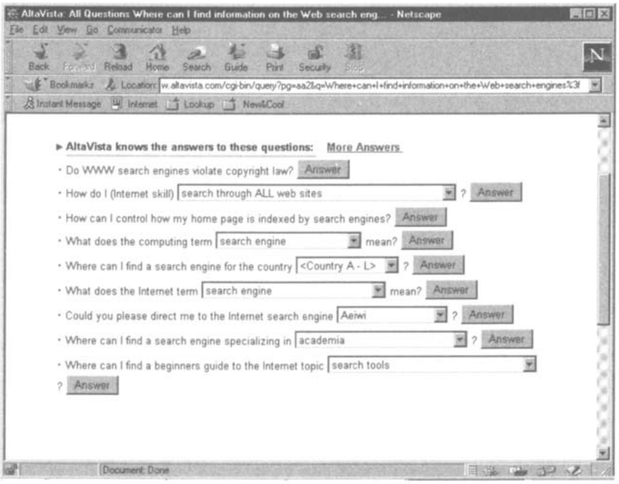
Books at Amazon.com
Books about Where can I...
Amazon.com Bestsellers

Figure 3. Options for the "(Internet skill)" Statement Slot of the Template in Figure 2.

The above examples show that the search engine uses templates and various options for the "(Internet skill)" slot in the template "How do I (Internet skill)." By clicking on the "More answers" option (see Figure 2), a user can get more such templates (see Figure 4). It may be noted that for many slots, such as "computing term," "Internet term," "search engine," and so on (see Figure 4), there are various options that can be displayed by clicking on the appropriate box. Some of these options are shown in Figure 5. Thus, the search engine uses various templates and provides options to fill in the slots to prepare sentence-like queries. Once a user prepares a search sentence by choosing the appropriate option from the drop-down box and clicks on the "Answer" button, the system conducts a search and fetches the relevant hits. However, it may be noted that the format of the query templates, and quite obviously the contents of the box showing the various options for the slots, vary from query to query. For example, when the system was asked "When will the next World Cup Football Games be held?" the templates that came up on the output screen were different (see Figures 6, 7, and 8). However, the system does not always come up with natural language query templates. For example, when the query, "What is a hurricane?" was given, the system simply produced a list of hits and no templates (see figure 9).

Use of Templates in Ask Jeeves

Ask Jeeves (http://www.askjeeves.com) is a meta-search engine that represents a model of an application using knowledge management techniques in order to better organize disparate information sources



AltaVista: All Questions: Where can I find information on the Web search eng... - Netscape

w.altavista.com/cgi-bin/query?pg=aa2&q=Where+can+I+find+information+on+the+Web+search+engines%3f

▸ AltaVista knows the answers to these questions: More Answers

- Do WWW search engines violate copyright law? Answer
- How do I (Internet skill) search through ALL web sites ? Answer
- How can I control how my home page is indexed by search engines? Answer
- What does the computing term search engine mean? Answer
- Where can I find a search engine for the country <Country A - L> ? Answer
- What does the Internet term search engine mean? Answer
- Could you please direct me to the Internet search engine Aeiwi ? Answer
- Where can I find a search engine specializing in academia ? Answer
- Where can I find a beginners guide to the Internet topic search tools ? Answer

Document: Done

File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Instant Message Internet Lookup New&Cool
N

Figure 4. More Templates for the Query Shown in Figure 1.

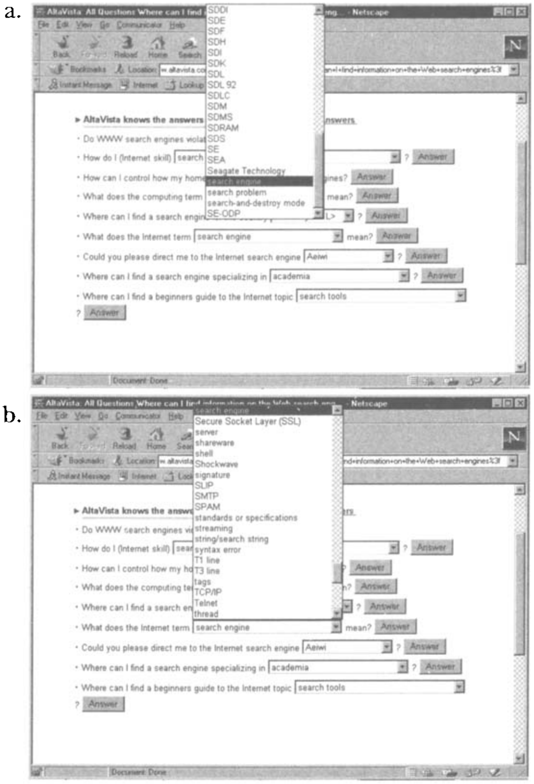
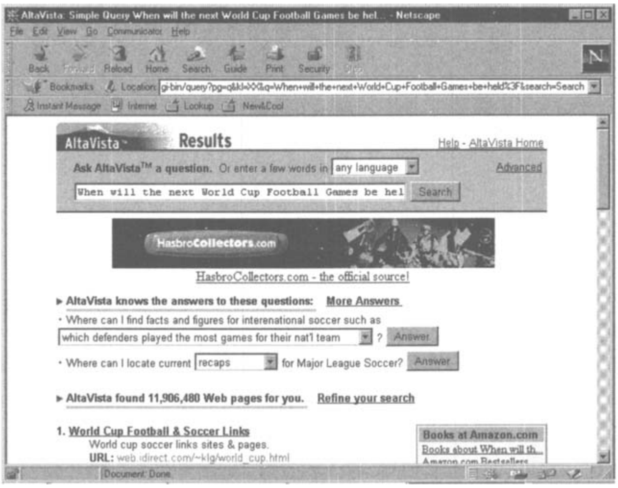


Figure 5. Options for the "computing term" (a) and "Internet term" (b) Slots in the Templates (in Figure 4).


File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Instant Message Internet Lookup New&Cool
Help - AltaVista Home
Advanced
Search
Answer
Document: Done

AltaVista: Simple Query When will the next World Cup Football Games be hel... - Netscape

gi-bin/query?pg=q&kl=XX&q=When+will+the+next+World+Cup+Football+Games+be+held%3F&search=Search

AltaVista Results

Ask AltaVista™ a question. Or enter a few words in any language

When will the next World Cup Football Games be hel

HasbroCollectors.com - the official source!

▸ AltaVista knows the answers to these questions: More Answers

· Where can I find facts and figures for interenational soccer such as which defenders played the most games for their nat'l team ?

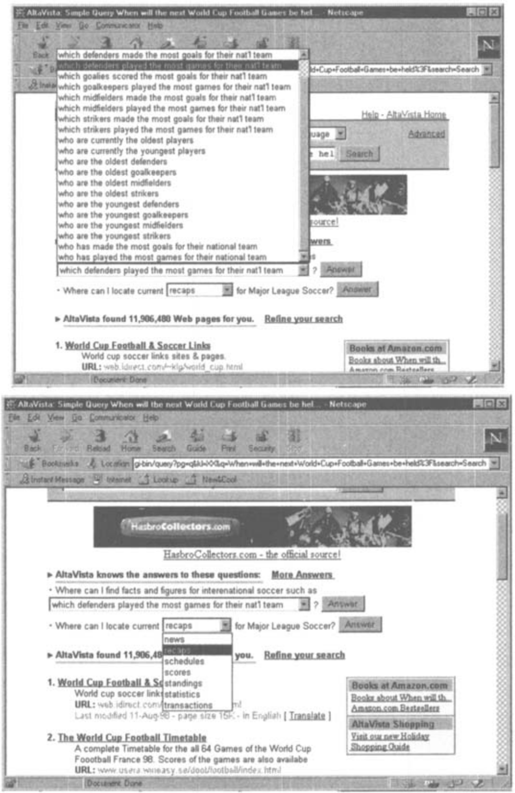
· Where can I locate current recaps for Major League Soccer?

▸ AltaVista found 11,906,480 Web pages for you. Refine your search

1. World Cup Football & Soccer Links
World cup soccer links sites & pages.
URL: web.idirect.com/~klg/world_cup.html

Books at Amazon.com
Books about When will th...
Amazon.com Bestsellers

Figure 6. Search Results of the Query "When Will the Next World Cup Football Games Take Place?"


File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Instant Message Internet Lookup New&Cool
Help - AltaVista Home
Advanced
Search
Answer
Document: Done

AltaVista: Simple Query When will the next World Cup Football Games be hel... - Netscape

which defenders made the most goals for their nat'l team
which defenders played the most games for their nat'l team
which goalies scored the most goals for their nat'l team
which goalkeepers played the most games for their nat'l team
which midfielders made the most goals for their nat'l team
which midfielders played the most games for their nat'l team
which strikers made the most goals for their nat'l team
which strikers played the most games for their nat'l team
who are currently the oldest players
who are currently the youngest players
who are the oldest defenders
who are the oldest goalkeepers
who are the oldest midfielders
who are the oldest strikers
who are the youngest defenders
who are the youngest goalkeepers
who are the youngest midfielders
who are the youngest strikers
who has made the most goals for their national team
who has played the most games for their national team

which defenders played the most games for their nat'l team ?

· Where can I locate current recaps for Major League Soccer?

▸ AltaVista found 11,906,480 Web pages for you. Refine your search

1. World Cup Football & Soccer Links
World cup soccer links sites & pages.
URL: web.idirect.com/~klg/world_cup.html

Books at Amazon.com
Books about When will th...
Amazon.com Bestsellers

gi-bin/query?pg=q&kl=XX&q=When+will+the+next+World+Cup+Football+Games+be+held%3F&search=Search

HasbroCollectors.com - the official source!

▸ AltaVista knows the answers to these questions: More Answers

· Where can I find facts and figures for interenational soccer such as which defenders played the most games for their nat'l team ?

· Where can I locate current recaps for Major League Soccer?

news
recaps
schedules
scores
standings
statistics
transactions

▸ AltaVista found 11,906,48 you. Refine your search

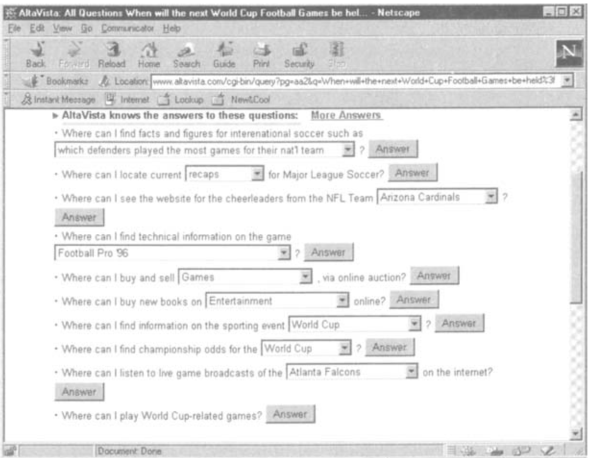
1. World Cup Football & Sc
World cup soccer link
URL: web.idirect.com/ ml
Last modified 11-Aug-98 - page size 15K - in English [Translate]

2. The World Cup Football Timetable
A complete Timetable for the all 64 Games of the World Cup Foootball France 98. Scores of the games are also availabe
URL: www.users.wineasy.se/doot/football/index.html

Books at Amazon.com
Books about When will th...
Amazon.com Bestsellers

AltaVista Shopping
Visit our new Holiday Shopping Guide

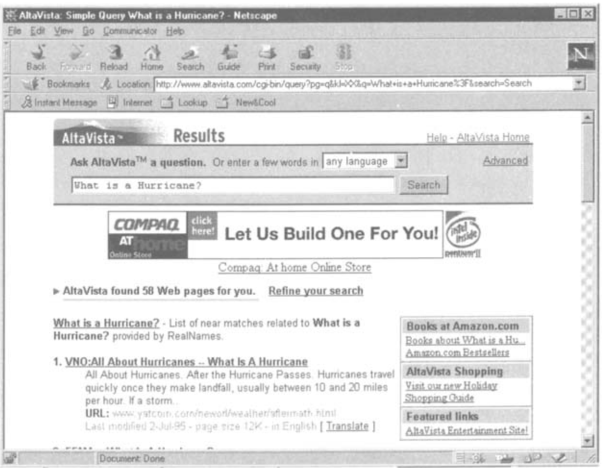
Figure 7. Options for the Various Templates.


AltaVista: All Questions When will the next World Cup Football Games be hel... - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: www.altavista.com/cgi-bin/query?pg=aa2&q=When+will+the+next+World+Cup+Football+Games+be+held%3f
Instant Message Internet Lookup New&Cool
Document: Done

▸ AltaVista knows the answers to these questions: More Answers

- Where can I find facts and figures for interenational soccer such as [which defenders played the most games for their nat'l team] ? Answer
- Where can I locate current [recaps] for Major League Soccer? Answer
- Where can I see the website for the cheerleaders from the NFL Team [Arizona Cardinals] ? Answer
- Where can I find technical information on the game [Football Pro '96] ? Answer
- Where can I buy and sell [Games] , via online auction? Answer
- Where can I buy new books on [Entertainment] online? Answer
- Where can I find information on the sporting event [World Cup] ? Answer
- Where can I find championship odds for the [World Cup] ? Answer
- Where can I listen to live game broadcasts of the [Atlanta Falcons] on the internet? Answer
- Where can I play World Cup-related games? Answer

Figure 8. More Templates for the Query Shown in Figure 6.


AltaVista: Simple Query What is a Hurricane? - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: http://www.altavista.com/cgi-bin/query?pg=q&kl=XX&q=What+is+a+Hurricane%3F&search=Search
Instant Message Internet Lookup New&Cool
AltaVista
COMPAQ AT home Online Store
click here!
Let Us Build One For You!
intel inside pentium II
Document: Done

AltaVista Results Help - AltaVista Home

Ask AltaVista™ a question. Or enter a few words in [any language] Advanced

[What is a Hurricane?] Search

Compaq: At home Online Store

▸ AltaVista found 58 Web pages for you. Refine your search

What is a Hurricane? - List of near matches related to **What is a Hurricane?** provided by RealNames.

1. VNO:All About Hurricanes -- What Is A Hurricane
   All About Hurricanes. After the Hurricane Passes. Hurricanes travel quickly once they make landfall, usually between 10 and 20 miles per hour. If a storm...
   URL: www.ystcom.com/neworl/weather/aftermath.html
   Last modified 2-Jul-95 - page size 12K - in English [Translate]

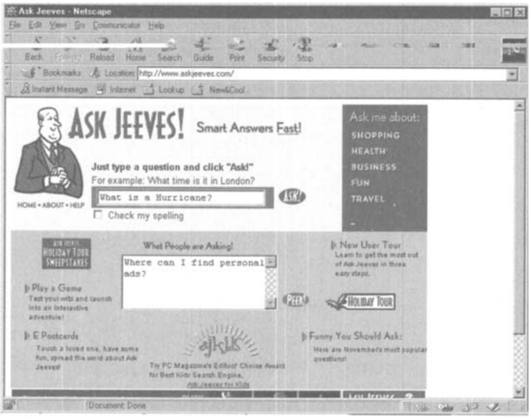
Books at Amazon.com
Books about What is a Hu...
Amazon.com Bestsellers

AltaVista Shopping
Visit our new Holiday Shopping Guide

Featured links
AltaVista Entertainment Site!

Figure 9. Output of the Query Shown in Figure 6 "What is a Hurricane?"

(Stanley, 1998). It draws on the expertise of experienced human Web searchers and encapsulates this expertise in a database so that it can be put to use by others. Various questions and their answers are manually selected by human editors who scan resources on the Web on a daily basis to build up a knowledge base of information about sites which might be used to answer common questions. The questions and the Web pages which answer them are then stored as a series of templates in the Ask Jeeves knowledge base, and keywords and concepts in a search string are matched against them in order to retrieve the questions and their corresponding Web sites (Stanley, 1998). Ask Jeeves was asked a simple question: "What is a hurricane?" (see Figure 10), and the system created a number of templates as shown in Figure 11.


Ask Jeeves - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Instant Message Internet Lookup New&Cool
ASK JEEVES!
HOME • ABOUT • HELP
ASK!
Ask me about:
SHOPPING
HEALTH
BUSINESS
FUN
TRAVEL
ASK JEEVES HOLIDAY TOUR SWEEPSTAKES
PEEK!
HOLIDAY TOUR
Document: Done

Location: http://www.askjeeves.com/

Smart Answers Fast!

Just type a question and click "Ask!"
For example: What time is it in London?

What is a Hurricane?

□ Check my spelling

What People are Asking!

Where can I find personal ads?

Play a Game
Test your wits and launch into an interactive adventure!

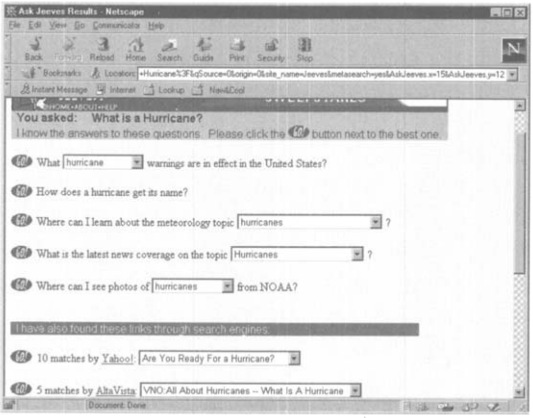
E Postcards
Touch a loved one, have some fun, spread the word about Ask Jeeves!

Try PC Magazine's Editors' Choice Award for Best Kids' Search Engine.
Ask Jeeves for Kids

New User Tour
Learn to get the most out of Ask Jeeves in three easy steps.

Funny You Should Ask:
Here are November's most popular questions!

Figure 10. Search Interface of Ask Jeeves.


Ask Jeeves Results - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Instant Message Internet Lookup New&Cool
HOME • ABOUT • HELP
Document: Done

Location: +Hurricane%3F&qSource=0&origin=0&site_name=Jeeves&metasearch=yes&AskJeeves.x=15&AskJeeves.y=12

You asked: What is a Hurricane?
I know the answers to these questions. Please click the GO! button next to the best one.

GO! What [hurricane] warnings are in effect in the United States?

GO! How does a hurricane get its name?

GO! Where can I learn about the meteorology topic [hurricanes] ?

GO! What is the latest news coverage on the topic [Hurricanes] ?

GO! Where can I see photos of [hurricanes] from NOAA?

I have also found these links through search engines.

GO! 10 matches by Yahoo!: [Are You Ready For a Hurricane?]

GO! 5 matches by AltaVista: [VNO:All About Hurricanes -- What Is A Hurricane]

Figure 11. Output (Showing Templates) of the Query Shown in Figure 10.

Thus, the search engines use templates, though not for IE but for IR, more specifically to help users formulate a natural language query. However, the way these templates are created is quite interesting. Human experts conduct searches on the topics and, based on the search results, organize them into different groups. These groupings could be on different topics such as "computing terms," "Internet terms," "Internet skills," and so on. These are then created as slots in sentence-like queries, and the various options for the slots (each "Internet term," "Internet skill," and so on) are then presented in boxes for the end-user to select while searching.

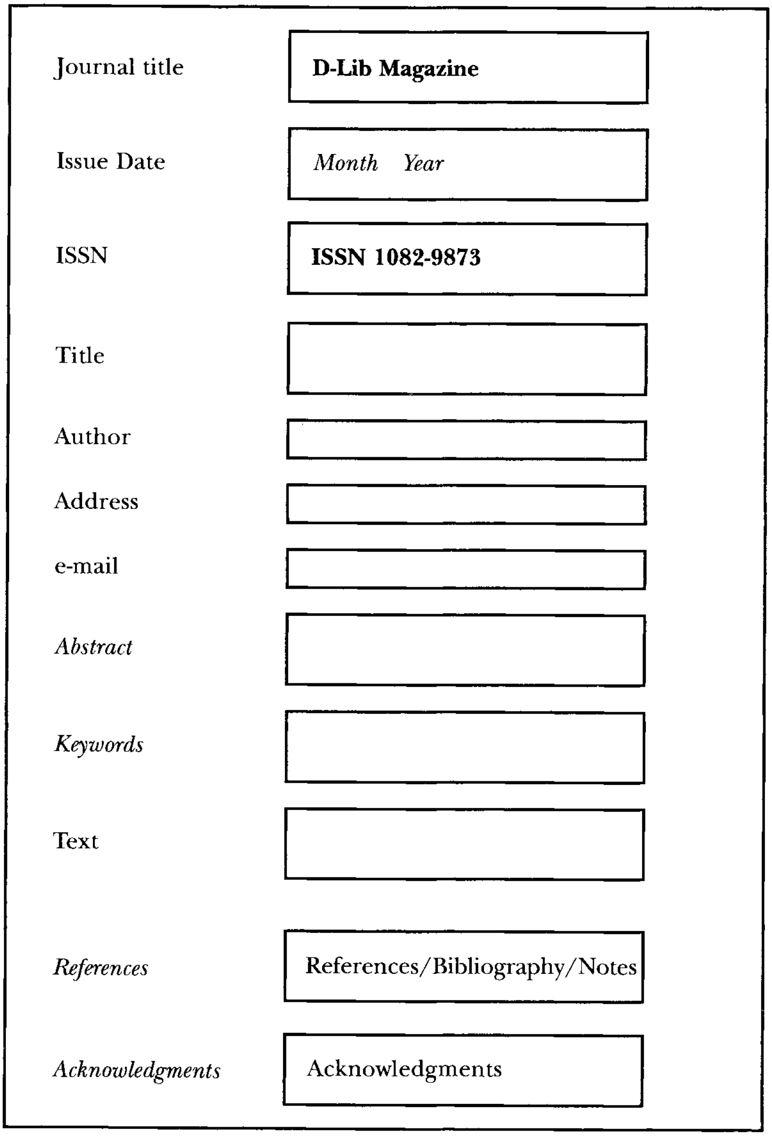
Template Mining in Digital Libraries

The British Library DL (digital library) Program (The British Library ..., 1997) defines digital library as the widely accepted descriptor for the use of digital technologies to acquire, store, conserve, and provide access to information and materials in whatever form it was originally published. The Stanford digital library working paper (Reich & Winograd, 1995) defines a digital library as a coordinated collection of services that are based on collections of materials, some of which may not be directly under the control of the organization providing a service in which they play a role. The contents of a digital library, being digital information sources, provide ample opportunities for applying template mining resulting in the extraction of valuable information in a number of areas in a digital library environment. Four areas of such an application have been identified and how this is done is discussed in the following sections.

Automatic Creation of Citation Databases of Digital Documents

Recently there have been some works on citation studies in the Web environment. Almind and Ingwersen (1998) introduced the concept of "Webometrics"—i.e., the application of informetric methods to the WWW. They have argued that citation analysis on the WWW has not been tested in practice. In another publication, Ingwersen and Hjortgaard (1997) discussed the advantages and disadvantages of using the WWW for informetric analysis and examined the pitfalls of online informetric analysis using the ISI (Institute of Scientific Information) files.

A quick and simple examination of some issues of the electronic journal *D-Lib Magazine* (http://www.dilib.org/dilib) revealed that template mining can be used to develop citation databases automatically from the online articles. Such databases may contain information somewhat similar to the ISI databases, such as the citing author, address of the citing author, title of the citing article, keywords, and so on as well as the authors, titles, and bibliographic details of the cited articles. The simple template mining approach may be used to extract the information for each of these fields in the citation database that can later be used for various citation analysis and other purposes. Analysis of the articles (they are called stories in *D-Lib Magazine*) published in the nine 1998 issues of *D-Lib Magazine* (January, February, March, April, May, June, July/August, September, and October) revealed a general structure of the articles (see Figure 12).



Journal title	**D-Lib Magazine**
Issue Date	*Month Year*
ISSN	**ISSN 1082-9873**
Title	
Author	
Address	
e-mail	
Abstract	
Keywords	
Text	
References	References/Bibliography/Notes
Acknowledgments	Acknowledgments

Figure 12. General Structure of Articles (called Stories) in *D-Lib Magazine.*

Figure 12 provides a general idea of the different kinds of templates that can be generated to extract information from the various sections of each article. Text that appears in bold in Figure 12 shows that the concerned text is constant—i.e., it appears in each article. Similarly, empty boxes indicate that texts appear there to indicate value for the given slot—e.g., title, author's name, and so on. Texts in some slots appear in a specific format, for example, in the "Issue Date" slot, the particular issue appears in the format "Month Year"—e.g., "February 1998." In some slots, the heading varies. For example, the heading used for references is usually "References" but the headings "Bibliography" or "Notes" are also used. This preliminary study has shown that, although this is the general structure of the articles in *D-Lib Magazine*, some articles may not have some of the slots—i.e, "Keywords," "Abstract," "References," or "Acknowledgments."

The values for some of the slots remain constant while, for most of the slots, they vary from one article to the other. For example, in the "Author" and "Address" slots, different patterns have been noticed. Articles may be written by only one author, by two authors, or by more than two authors. Again, when more than one author is involved, they may have the same or different addresses. Articles also differ in terms of layout for writing the authors' names and addresses: while in most cases they appear vertically, one after the other, in some cases they appear horizontally, one after the other. The general structure of these slots is shown in Figure 13.

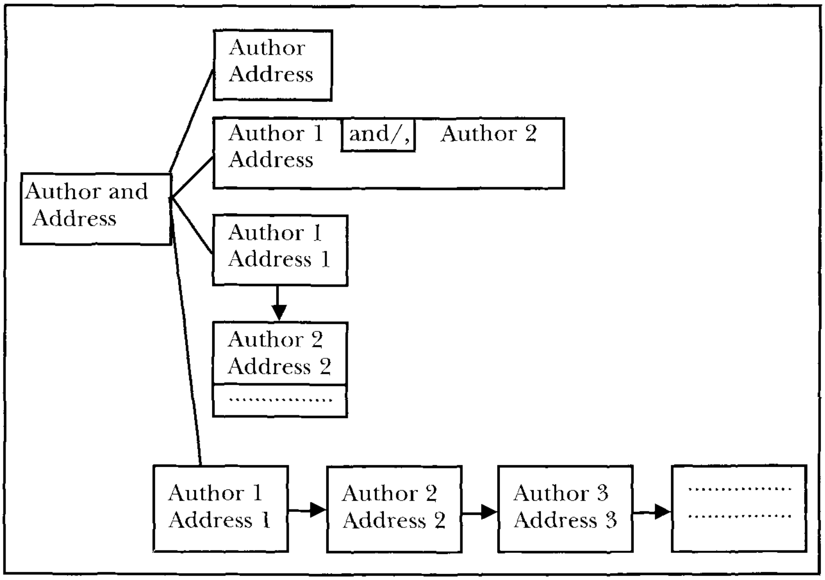
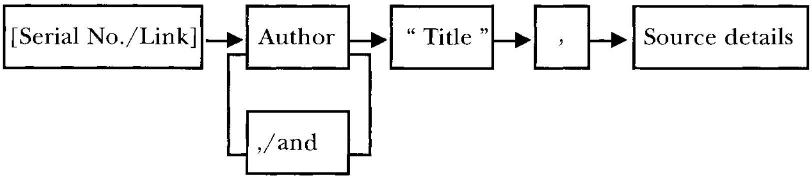


Figure 13. Template for the Author and Address Information.

Similarly, the contents of the "References" slot vary significantly depending on the type of item referred to. Some articles have only print versions while others are available only in electronic form. Therefore, the citations relating to the sources vary: there may be journal title, volume number, and so on or there may only be the URL or both the bibliographic details and the URL. Although the general pattern of citations as they appear at the end of each article appears to be quite simple (see Figure 14), and therefore amenable to template mining, a closer look at the references indicates that there are a number of irregularities there. For example, different authors use different citation styles and, sometimes, within the same article, authors follow different citation styles for similar types of material (see Figure 15). If these irregularities are sorted out, the template matching technique can be used to extract relevant information, including URLs, from the citations. There are two ways to sort out these irregularities and to ensure a standard citation style. The first one could be to impose rigorous editorial practice, but that would be more expensive and time consuming, causing delays in publication. The second, and less expensive, approach may be to prepare templates for each type of citation and make them available online. Authors can download and make use of these templates for preparing the list of references. This will ensure a standard citation style. The same practice may be followed for the other parts of the article, and eventually the whole structure of the articles can be standardized, thereby facilitating the use of template mining.



Sample References from D-Lib Magazine

[Chen et al., 1996] Hsinchun Chen, Chris Schuffels, and Rich Orwig, "Internet Categorization and Search: A Machine Learning Approach," Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, Volume 7, Number 1, Pages 88-102, 1996.

[5] Harnad E., Print Archive and Psycoloquy and BBS Journal Archives <http://www.princeton.edu/~harnad>

[11] Trudi Bellardo Hahn, Text Retrieval Online: Historical Perspective on Web Search Engines, pp. 7-10, Bulletin of the American Society for Information Science, April/May, 1998.

Figure 14. Simple Structure of References in the *D-Lib Magazine*.

[ieeePC97]
Bacon J, Bates J and Halls D, Location oriented multimedia, IEEE Personal Communications 4(5), pp 48-57, October 1997.

[Phelps and Wilensky, 1996b] Thomas A. Phelps and Robert Wilensky, "Multivalent Documents: Inducing Structure and Behaviors in Online Digital Documents", Proceedings of Hawaii International Conference on System Sciences '96 (Best Paper Award, Digital Documents Track).

[Salton, 1989]Gerard Salton, Automatic Text Processing, Addison-Wesley, 1989.

[Weiss et al., 1996]Ron Weiss, David Gifford et al., "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering", Proceedings of the Seventh ACM Conference on Hypertext, March 1996, Washington, DC.

[Weiss and Indurkhya, 1993] S. Weiss and N. Indurkhya, "Optimized rule induction," IEEE Expert 8, 6, 61-69.

[Yahoo]Yahoo [http://www.yahoo.com/]

[Zhu et al., 1997] Quan Zhu et al., "Searching for Parts and Services on the Web", Proceedings of International Symposium on Research, Development, and Practice in Digital Libraries, Nov. 18 - 21, 1997, Tsukuba, Japan.

[1] Tom Sanville of the OhioLINK consortium has noted in public presentations a correlation between ease of access and use and the amount of use.

[2] TULIP: The University Licensing Program <http://www.elsevier.nl/locate/tulip>

[3] Red Sage: Final Report <http://www.springer-ny.com/press/redsage>

[5] Harnad E. Print Archive and Psycoloquy and BBS Journal Archives <http://www.princeton.edu/~harnad>

[9] JSTOR http://www.jstor.org

[Arms 1995] R. Kahn and R. Wilensky, A Framework for Distributed Digital Object Services, (May 1995).

[Bearman 1998] D. Bearman and J. Trant, Authenticity of Digital Resources:Towards a Statement of Requirements in the Research Process, D-LIB on-line magazine, (June 1998).

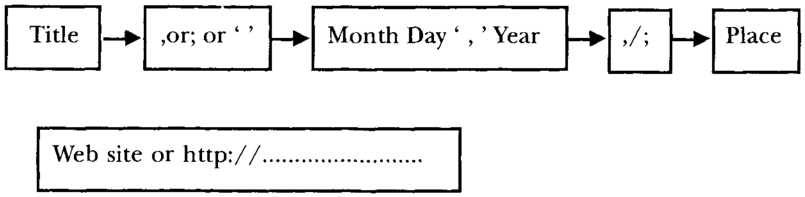
C. Lynch et al., A White Paper on Authentication and Access Management Issues in Cross-organizational Use of Networked Information Resources,

(Note: The underlines in the last three references indicate that they are hyperlinked to the respective URLs)

Figure 15. Sample References from Different Issues of *D-Lib Magazine* Showing the Lack of Standards in the Citation Style.

Automatic Extraction of Information from News Items in Electronic Journals

Electronic journals often contain news items, and important information can be extracted for the creation of databases or for any other use by a simple template mining approach. A scan through the pages of the nine issues of *D-Lib Magazine* revealed that there is a section called "Goings on" that provides information on conferences/seminars/workshops, and so on under the heading "Clips & Pointers." A scan through the items appearing under the heading "Goings on" in all the 1998 issues of *D-Lib Magazine* revealed that the seminar/conference/workshop announcements follow a general pattern as shown in Figure 16. This shows that a simple template can extract information about the various forthcoming seminars, conferences, and so on. Particular information, such as the place, date, Web address, and so on, can also be extracted by such templates. However, once again such a template mining approach calls for a standard format and layout. Templates can also be generated to extract further information such as specific topics, contact address, deadlines, and so on.



Note: Sometime in the text the phrase 'Web site' appears that has a hyperlink and sometimes the URL appears; either of these can be used for template mining

Figure 16. Simple Templates for Conference/Seminar/Workshop Announcements.

Automatic Identification of Funding/Sponsoring Agencies for Research

A scan through the "Acknowledgment" section in the articles revealed that they contain information about the funding/sponsoring agency's name, address, grant number, and so on. A study of the various articles appearing in the electronic journals can help generate a pattern, and thereby appropriate templates, that will be able to extract the relevant information for further use. Again, in order to standardize the practice of providing these items of information, the editorial board of the journal can make appropriate templates available to the authors.

Information Extraction Using Metadata and Template Mining

Metadata are data about data. However, this definition is too simple and does not tell us their characteristic features. A better definition has been provided by Dempsey and Heery (1998) according to whom metadata are data associated with objects which relieve their potential users of having to have full advance knowledge of the existence of characteristics. Younger (1997) defined metadata as documentation about documents and objects; they describe resources, indicate where they are located, and outline what is required in order to use them successfully. There are several metadata schemes created by library and information professionals over the years, the most prominent ones being the MARC formats, the AACR2 catalog formats, subject headings lists (such as the LCSH), and classification schemes such as LC, DDC, UDC, and so on. Each of these schemes is constructed by experts in the field from an understanding of the specific domains, information resource needs, and the requirements for describing documents. While these schemes have been used for bibliographic access and control for decades, there remains the question of how to catalog and index materials available on the Internet using these schemes. This has given rise to a thought that electronic documents need to be *self-indexed* (as opposed to the assignment of cataloging and indexing tags and value added by cataloging and indexing agencies or library staff). However, it is obvious that, in order for the documents to be self-indexed, a core set of metadata elements must be identified, and each creator of electronic documents should be able to implement it in the record that he creates. With this objective, a simple resource description set of data has emerged—the Dublin Core (http://purl.org/metadata/dublin_core). The Dublin Core metadata set prescribes fifteen elements, namely (http://www.lub.lu.se/cgi-bin/nmdc.pl): title, creator, subject (keywords, controlled vocabulary, and classification), description (abstract and content description), publisher, contributor (other than the creator), date, type (category of the resource), format (HTML, Postscript, etc.), identifier (URL, string or number used to identify the resource), source (from which this resource is derived), language, relation (with other resources), coverage (spatial and/or temporal characteristics of the resources), and rights (link to a copyright notice, etc.).

Weibel (1995,1996) suggested that, in order to enable information creators to apply metadata, a mechanism for embedding the data within HTML documents had to be established. The Dublin Core looks at one aspect of metadata—i.e., the simple description—but, as Dempsey and Heery (1998) suggested, there is a need for more complex description for particular specialist domains. In 1996, there was a conference organized by UKOLN and OCLC to examine the various metadata issues including the Dublin Core. This meeting gave rise to a proposal, the Warwick Framework proposal (Lagoze, 1996) (named after the place of the conference), calling for an architecture for the interchange of various metadata packages.

Metadata are an important tool for resource discovery from digital documents. They not only help users locate the required information resources, they also help in the examination and selection (or rejection) of the retrieved items. Various fields, such as subject, description, language, sources, and so on, can provide necessary information for examining the relevance of the retrieved resources. Web search engines, as an aid during the examination phase, generally construct an information surrogate to display the search hits. These surrogates generally consist of the URL, title of the Web page, and some summary text that is derived with the aid of some heuristics (Lagoze, 1997).

Lagoze (1997) argued that, recognizing the limitations of the current search engines, researchers are now actively pursuing both standards for descriptive surrogates for networked objects and methods for associating surrogates with those objects. Such studies aim at developing a number of surrogate templates that would facilitate the resource discovery process. The template mining approach can be used effectively for the resource discovery process. The metadata tags, such as subject and description in the Dublin Core metadata format, for example, can be extended and be made more structured by specifying various templates. The following section briefly describes how this can be achieved.

Use of the Principles of Pre-Coordinate Indexing Systems in Building Subject Templates

Information retrieval systems have used two different types of indexing systems, namely, pre-coordinate indexing and post-coordinate indexing (for detailed discussions, see Lancaster, 1998; Foskett, 1996). Pre-coordinate indexing systems—classification schemes like the Dewey Decimal Classification (DDC) (Dewey, 1996), Universal Decimal Classification (UDC, 1985), and Colon Classification (CC) (Ranganathan, 1965), and so on that use artificial notations or subject indexing systems such as relational indexing (Farradane, 1980a, 1980b), PRECIS (Austin & Dykstra, 1984), and POPSI (Bhattacharyya, 1981) that use natural language terms—represent the content of an information resource by synthesizing the various components of the subject and organizing these in a specific order. On the contrary, post-coordinate indexing systems do not rely on the a priori relations or organization of the constituent search terms; rather, retrieval is performed by searching through each individual term and then the output is generated based on the coordination of the terms at the retrieval stage; for example, according to the principles of set theory. Most of the modern day IR systems, including the search engines, follow the principles of post-coordinate indexing systems. However, it is proposed that, for resource discovery from digital libraries or from the Web, one can use the principles of pre-coordinate indexing systems. The following paragraph suggests a simple approach to this.

The basic tenet of this approach is that pre-coordinate indexing systems help indexers generate subject index entries that represent the subject matter of the document concerned. This approach has been successfully used in libraries for organizing library materials on the shelves based on classification numbers. These class numbers are created according to the principles of pre-coordinate indexing systems, and the notations represent the content of the document concerned. Similarly, this approach has been used in preparing alphabetical subject index entries for documents in national bibliographies and in other bibliographic databases. One major drawback of these pre-coordinate indexing systems is that it is largely human dependent because human indexers need to analyze the documents and prepare the subject statements, which are then manipulated by computers for generating multiple entries or are used to prepare the class number. In other words, this process involves a significant amount of human expertise and time and therefore is a slow and expensive process. As a result, adopting this approach is almost impossible with a large collection and is impossible in the Web environment with millions of documents.

The aforementioned problem could be solved if this task was accomplished by the author or the generator of information resources. If authors can somehow indicate the key concepts treated in the documents, along with an indication of the appropriate categories where they belong, then subject statements or simple surrogates can be prepared automatically. A template can be provided to the generator of information resources—e.g., an author of an article or a report may fill in the various slots with appropriate information. Such templates can be generated using the various categories proposed in pre-coordinate indexing systems—e.g., the fundamental categories of Ranganathan (1967), the nine relational categories proposed by Farradane (see for discussion, Farradane, 1980a, 1980b; Chowdhury, 1989), the various role operators proposed in PRECIS by Austin (Austin & Dykstra, 1984), the various categories proposed by Bhattacharyya (1981) in POPSI, and so on. The author or the generator of the information resource should be able to understand the connotation of each category and thus should be able to fill in the template according to the semantic content of the concerned information resource. For example, the author of a document entitled "Internet as a Tool for Management of OPACs in the Libraries in Singapore" is given a simple template created according to the PRECIS Role Operators (see Austin & Dykstra, 1984; Chowdhury, 1995). Now the author can fill in the slots in the template according to the role of each term in the given document as shown in Figure 17. Note that the template shown in Figure 17 is not the full implementation of PRECIS and therefore does not show all PRECIS operators; this is rather indicative of the kind of application that one can build according to the principles of any pre-coordinate indexing system. Which system is the best, and yet easy for an author to understand and apply, is a matter for further research.

Location: Singapore

Key system/Object of transitive action/Agent of intransitive action/Effect of action: Libraries
Part/property: OPAC

Action, Discipline, etc.: Management

Agent/Performer of transitive action/Intake/Factor: Internet

Viewpoint:

Selected Instance, study region, sample population:

Form of Document, target user:

Figure 17. Simple Templates according to the PRECIS Role Operators.

Such templates, prepared according to the principles of any pre-coordinate indexing system or a modified version of that, can be used both for better retrieval and for the preparation of document surrogates. Better retrieval can be achieved because of the semantic values attached to the terms; this would help reduce the *false drops.* Document surrogates can be prepared automatically according to the prescribed rules of the concerned indexing system—e.g., according to the principle of generating the index entries in PRECIS (Austin & Dykstra, 1984).

Conclusion

The explosive growth in our capabilities to collect and store data over the past decades has given rise to a new field of study, called knowledge discovery in databases, that is concerned with the creation of a new generation of tools and techniques for automated and intelligent database analysis. Raghavan et al. (1998) suggest that KDD refers to the whole process in the path from data to knowledge and to use descriptive phrases for specific tasks in the process, such as pattern extraction methods, pattern evaluation methods, or data cleaning methods. Thus, simply speaking, KDD is the process of deriving useful knowledge from real-world databases through the application of pattern extraction techniques. The grand challenge of KDD is, therefore, to automatically process large quantities of raw data, identify the most significant and meaningful patterns, and present these as knowledge appropriate for achieving a user's goals.

This discussion has proposed that template mining, which is based on pattern recognition and pattern matching in natural language texts, can be used for extracting different kinds of information from digital documents or text databases. Citation databases can be built automatically based on the template mining approach, from digital documents, such as from articles in electronic journals. Template mining can also be used to extract different types of information, such as information about the funding/sponsoring agencies of research projects as they appear in the acknowledgments section of articles. Another application could be to identify inter-document links by tracing the hypertext links using the "http://...." template, and thus a network of articles in a specific domain can be built that would be useful for researchers in the subject concerned. This article has also indicated that template mining can be used to extract various items of news from digital documents such as in the extraction of conference information from electronic journals, and so on.

This article has also indicated that the template mining approach can be incorporated within the metadata format in order to facilitate better information retrieval and to enable automatic generation of document surrogates that would help the end-user filter the search output. This is very necessary, particularly in the WWW environment where a given search may retrieve several thousands, even millions, of records. This could be achieved by using the concept categorization and organization principles of the pre-coordinate indexing systems. Such pre-coordinate systems will improve the quality of the output of search engines that basically follow the principles of post-coordinate indexing systems. In other words, while post-coordinate indexing principles used in the search engines will retrieve digital information resources, the principles of pre-coordinate indexing systems may be used to filter them.

However, all the above-mentioned applications will be possible provided the digital information resources appear in standard format and layout. The first few applications of template mining mentioned above will require strict adherence to formats. This should not be too difficult, as authors are used to following author instructions issued by publishers/

editors of printed journals and publication houses. It should not be a big problem to implement the same in the electronic environment. The only requirement will be to formulate appropriate guidelines and, if possible, appropriate templates can be made available to the authors that could be used online while preparing the documents. For example, simple templates may be prepared for each type of publication, and authors may be required to just click on the type of document, such as journal article, conference paper, online sources, and so on, to get the appropriate slots to fill-in with data for author, title, source details, and so on. This would not only ensure consistency in the references but would also facilitate template mining applications.

For the last application proposed above, much more work needs to be done. It may be incorporated as an element in the metadata format, or may be added as a required field in any digital document, or embedded in HTML. However, some experiments must be conducted in order to determine which pre-coordinate indexing system will be more appropriate and yet easy for the authors to use while preparing the digital information sources. This may be kept at a simple level just by creating slots for each category of the chosen pre-coordinate indexing system or may be made more complex by incorporating the phase relations proposed by Ranganathan (1967, 1987) as well. However, with the increasing complexity of the system, it may be more difficult, and therefore more inhibiting, for the authors who are required to fill in the slots in the templates. Nevertheless, this job has to be done by the generators of digital information resources, otherwise it may be too difficult and expensive for any agency to analyze each digital document and fill in the slots as done in traditional libraries for classifying and indexing materials. Initially, it may seem an extra burden to the authors but, as we need to follow instructions from editors and publishers and follow HTML and similar standards while preparing hard copy and/or digital documents, we may have to do this in order to make our generated information more widely and easily available to potential users. Eventually, all these additional activities may be incorporated within popular software and, as we can now use editors for creating HTML documents rather than coding them by hand, authors may just fill in the templates as part of their document creation task. Detailed experiments are currently underway and, upon successful completion of this research, we expect to develop a model for template mining from digital information resources.

Acknowledgment

The author gratefully acknowledges the encouragement and support obtained while writing this article from Michael F. Lynch, Professor Emeritus, Department of Information Studies, University of Sheffield, Sheffield, England.

REFERENCES

Ai, C. S.; Blower, P. E.; & Ledwith, R. H. (1990). Extraction of chemical reaction information from primary journal text. *Journal of Chemical Information and Computer Sciences, 30*(2), 163-169.

Almind, T. C., & Ingwersen, P. (1998). Informetric analyses on the World Wide Web: Methodological approaches to "Webmetrics." *Journal of Documentation, 54*(4), 404-426.

Andersen, P. M.; Hayes, P. J.; Huettner, A. K.; Schmandt, L. M.; Nirenburg, I. B.; & Weinstein, S. P. (1992). Automatic extraction of facts from press releases to generate news stories. In *Third Conference on Applied Natural Language Processing* (31 March-3 April, 1992, Trento, Italy) (pp. 170-177). Morristown, NJ: Association of Computational Linguistics.

Andersen, P. M., & Huettner, A. K. (1994). Knowledge engineering for the JASPER fact extraction system. *Integrated Computer-Aided Engineering, 1*(6), 473-493.

Austin, D., & Dykstra, M. (1984). *PRECIS: A manual of concept analysis and subject indexing* (2d ed.). London, England: British Library.

Baralis, E., & Psaila, G. (1997). Designing templates for mining association rules. *Journal of Intelligent Information Systems, 9*(1), 7-32.

Bhattacharyya G. (1981). Some significant results of current classification research in India. *International Forum on Information and Documentation, 6*(1), 11-18.

The British Library Research and Innovation Centre. (1998). *The British Library Digital Library Programme.* Retrieved December 16, 1998 from the World Wide Web: http://www.bl.uk/services/ric/diglib/digilib.html.

Chong, W., & Goh, A. (1997). FIES: Financial information extraction system. *Information Services & Use, 17*(4), 215-223.

Chowdhury, G. G. (1989). *Nature and applicability of Farradane's relational analysis with particular reference to the preparation of linear index entries.* Unpublished doctoral dissertation, Jadavpur University, Calcutta, India.

Chowdhury, G. G. (1992). *Application of natural language processing to chemical patents.* Unpublished doctoral dissertation, University of Sheffield, Sheffield, England.

Chowdhury, G.G. (1995). *PRECIS: A workbook.* Calcutta, India: IASLIC.

Chowdhury, G. G., & Lynch, M. F. (1992a). Automatic interpretation of the texts of chemical patent abstracts. Part 1: Lexical analysis and categorization. *Journal of Chemical Information and Computer Sciences, 32*(5), 463-467.

Chowdhury, G. G., & Lynch, M. F. (1992b). Automatic interpretation of the texts of chemical patent abstracts. Part 2: Processing and results. *Journal of Chemical Information and Computer Sciences, 32*(5), 468-473.

Coates-Stephens, S. (1992). *The analysis and acquisition of proper names for robust text understanding.* Unpublished doctoral dissertation, City University, London, England.

Costantino, M.; Morgan, R. G.; & Collingham, R. J. (1996). Financial information extraction using pre-defined user-definable templates in the LOLITA system. *Journal of Computing & Information Technology, 4*(4), 241-255.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Croft, W. B. (1995). What do people want from information retrieval?: The top 10 research issues for companies that use and sell IR systems. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/november95/11croft.html.

De Jong, G. (1982). An overview of the FRUMP system. In W. Lehnert & M. H. Ringle (Eds.), *Strategies for natural language processing* (pp. 149-176). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dempsey, L., & Heery, R. (1998). Metadata: A current view of practice and issues. *Journal of Documentation, 54*(2), 145-172.

Dewey, M. (1996). *Dewey decimal classification and relative index.* Albany, NY: Forest Press.

Farradane, J. E. L. (1980a). Relational indexing. Part 1. *Journal of Information Science, 1*(5), 267-276.

Farradane, J. E. L (1980b). Relational indexing. Part 2. *Journal of Information Science, 1*(6), 313-324.

Foskett, A. C. (1996). *The subject approach to information* (5th ed.). London, England: Library Association Publishing.

Gaizauskas, R., & Wilks, Y. (1998). Information extraction: Beyond document retrieval. *Journal of Documentation, 54*(1), 70-105.

Gaizauskas, R., & Humphreys, K. (1997). Using a semantic network for information extraction. *Natural Language Engineering, 3*(2/3), 147-169.

Guarino, N. (1997). Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. In M. T. Pazienza (Ed.), *Information extraction: A multidiciplinary approach to an emerging information technology international summer school* (No. 1299: Lecture Notes in Artificial Intelligence) (pp. 139-168). New York: Springer.

Ingwersen, P., & Hjortgaard C. F. (1997). Data set isolation for bibliometric online analysis of research publication: Fundamental methodological issues. *Journal of the American Society for Information Science, 48*(3), 205-217.

Jacobs, P., & Rau, L. F. (1990). SCISOR: Extracting information from on-line news. *Communications of the ACM, 33*(11), 88-97.

Jones, P. A., & Paice, C. D. (1992). A "select and generate" approach to automatic abstracting. In T. McEnery & C. D. Paice (Eds.), *Proceedings of the BCS 14th Information Retrieval Colloquium* (pp. 141-154). Berlin, Germany: Springer-Verlag.

Kemp, N. M. (1995). *The application of natural language processing to chemical patents.* Unpublished doctoral dissertation, University of Sheffield, Sheffield, England.

Lagoze, C. (1996). The Warwick Framework: A container architecture for diverse sets of metadata. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

Lagoze, C. (1997). From static to dynamic surrogates: Resource discovery in the digital age. *D-Lib Magazine.* Retrieved December 7, 1998 from the World Wide Web: http://www.dlib.org/dlib/june97/06lagoze.html

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice* (2d ed.). Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois.

Lawson, M.; Kemp, N.; Lynch, M. F.; & Chowdhury, G. G. (1996). Automatic extraction of citations from the text of English language patents: An example of template mining. *Journal of Information Science, 22*(6), 423-436.

Lytinen, S. L., & Gershman, A. (1986). ATRANS: Automatic processing of money transfer messages. In *AAAI '86* (The Fifth National Conference on Artificial Intelligence, August 11-15, 1986, Philadelphia, Pennsylvania) (pp. 1089-1093). Los Altos, CA: Morgan Kaufmann.

Paice, C.D. (1981). The automatic generation of literature abstracts: An approach based on the identification of self-indicating phrases. In R. N. Oddy, S. E. Robertson, C. J. Van Rijsbergen, & P. W. Williams (Eds.), *Information retrieval research* (pp. 172-191). London, England: Butterworths.

Paice, C. D., & Husk, G. D. (1987). Towards the automatic recognition of anaphoric features in English text: The impersonal pronoun "it." *Computer Speech and Language, 2*(2), 109-132.

Postma, G. J., & Kateman, G. (1993). A systematic representation of analytical chemical actions. *Journal of Chemical Information and Computer Sciences, 33*(3), 350-368.

Postma, G. J.; van der Linden, J. R.; Smits, J. R. M.; & Kateman, G. (1990a). TICA: A system for the extraction of data from analytical chemical texts. *Chemometrics & Intelligent Laboratory Systems, 9*(1), 65-74.

Postma, G. J; van der Linden, J. R.; Smits, J. R. M.; & Kateman, G. (1990b). TICA: A system of extraction of analytical chemical information from texts. In E. J. Karjalainen (Ed.), *Scientific computing and automation (Europe)* (pp. 407-414). Amsterdam: Elsevier.

Raghavan, V. V.; Deogun, J. S.; & Sever, H. (1998). Introduction (to the special issue on knowledge discovery and data mining). *Journal of the American Society for Information Science, 49*(5), 397-402.

Ranganathan, S. R. (1987). *Colon classification* (7th ed.). Bangalore, India: Sarada Ranganathan Endowment for Library Science.

Ranganathan, S. R. (1967). *Prolegomena to library classification* (3d ed.). Bangalore, India: Sarada Ranganathan Endowment for Library Science.

Reich, V., & Winograd, T. *Working assumptions about the digital library* (Stanford Digital Library Working Paper Feb. 23, 1995). Retrieved December 16, 1998 from the World Wide Web: http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1995-0006.

Sager, N. (1981). *Natural language information processing: A computer grammar of English and its applications.* Reading, MA: Addison-Wesley.

Sasaki, Y. (1998). Learning of information extraction rules using ILP-programming report. In *PADD98* (Proceedings of the Second International Conference on the Practical Application of Knowledge Discovery and Data Mining, London, 25-27 March 1998) (pp. 195-205). Blackpool, England: Practical Application.

Schank, R. C. (1975). *Conceptual information processing.* Amsterdam: North-Holland.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Lawrence Erlbaum.

Schank, R. C., & Riesbeck, C. K. (1981). *Inside computer understanding: Five programs plus miniatures.* Hillsdale, NJ: Lawrence Erlbaum.

Shuldberg, H. K; Macpherson, M.; Humphrey, P.; & Corely, J. (1993). Distilling information from text: The EDS template filler system. *Journal of the American Society for Information Science, 44*(9), 493-507.

Stanley, T. (1998). Ask Jeeves: The knowledge management search engine. *Ariadne, Vol. 17.* Retrieved December 11, 1998 from the World Wide Web: http://www.ariadne.ac.uk/issue17/search-engines/intro.html

UDC: BS 1000 International Medium Edition. English text. (1985). London, England: British Standards Institution.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Wakao, T.; Gaizauskas, R.; & Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *COLING '96* (The 16th International Conference on Computational Linguistics, August 5-9, 1996, Copenhagen, Denmark) (pp. 418-423). Copenhagen, Denmark: Center for Sprogteknologi.

Weibel, S. (1995). Metadata: The foundations of resource description. *D-lib Magazine.* Retrieved December 11, 1998 from the World Wide Web: http://www.dlib.org/dlib/July95/07Weibel.html.

Weibel, S. (1996). *A proposed convention for embedding metadata in HTML.* Retrieved December 10, 1998 from the World Wide Web: http://www.oclc.org/~weibel/html-meta.html.

Younger, J. A. (1997). Resources description in the digital age. *Library Trends, 45*(3), 462-487.

Zamora, E., & Blower, P. E. (1984a). Extraction of chemical reaction information from primary journal text using computational linguistic techniques: 1. Lexical and syntactic phases. *Journal of Chemical Information and Computer Sciences, 24*(3), 176-181.

Zamora, E., & Blower, P. E. (1984b). Extraction of chemical reaction information from primary journal text using computational linguistic techniques: 2. Semantic phase. *Journal of Chemical Information and Computer Sciences, 24*(3), 181-188.