Effective techniques for automatic extraction of Web publications

A.C.M. Fong
S.C. Hui and
H.L. Vu

The authors

A.C.M. Fong works at the Institute of Information and Mathematical Sciences of Massey University, Auckland, New Zealand.
S.C. Hui is an Associate Professor and **H.L. Vu** is a Research Student, both at the School of Computer Engineering at Nanyang Technological University, Singapore.

Keywords

Internet, Research, Electronic publishing, Content analysis

Abstract

Research organisations and individual researchers increasingly choose to share their research findings by providing lists of their published works on the World Wide Web. To facilitate the exchange of ideas, the lists often include links to published papers in portable document format (PDF) or Postscript (PS) format. Generally, these publication Web sites are updated regularly to include new works. While manual monitoring of relevant Web sites is tedious, commercial search engines and information monitoring systems are ineffective in finding and tracking scholarly publications. Analyses the characteristics of publication index pages and describes effective automatic extraction techniques that the authors have developed. The authors' techniques combine lexical and syntactic analyses with heuristics. The proposed techniques have been implemented and tested for more than 14,000 Web pages and achieved consistently high success rates of around 90 percent.

Electronic access

The research register for this journal is available at
http://www.emeraldinsight.com/researchregisters

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1468-4527.htm

Online Information Review
Volume 26 · Number 1 · 2002 · pp. 4–18
© MCB UP Limited · ISSN 1468-4527
DOI 10.1108/14684520210418347

Introduction

The World Wide Web (WWW) is fast becoming the preferred medium for information transfer, particularly among members of the scientific community. Increasingly, scholarly publications are made available online by professional societies and academic publishers so that subscribers may download the relevant publications. Research organisations and individual researchers generally provide information on their research activities by listing their research works on their Web pages. While allowable by copyright laws, their research index Web pages also provide links to papers in portable document format (PDF) or Postscript (PS) format. Otherwise, links are often provided to the relevant publishers' Web pages or to preprint articles. This facilitates the exchange of ideas and sometimes creates new research opportunities and collaboration.

Since research organisations and individuals frequently update their index Web pages to include additions to their collections of research works, interested researchers at other organisations have to monitor and track progress of new findings regularly. While it is possible for members of a closely-knit research community to notify each other of new findings either regularly or whenever new findings become available, much of the information may not be relevant to the recipients. More importantly, most interested researchers still need actively to seek information that they require by closely monitoring relevant Web pages manually, which is both tedious and inefficient.

Commercial search engines are ineffective in finding or monitoring publication index Web pages as they generally return commercial Web sites in response to a search query. Many Web sites that are important to researchers are omitted in the process. Also, most commercial search engines do not index document files that are in PDF or PS format. It is therefore necessary to develop techniques for automatic extraction of, first of all, research index Web pages and, second, research papers available on the Web. However, a number of difficulties must be resolved. The first relates to the formats adopted by different researchers and

Refereed article received 11 October 2001
Approved for publication 10 November 2001

organisations for listing their publications on the Web. Also, publication information may be mixed with other information, such as biographies of individuals, on the same Web page. In this paper, we describe techniques and algorithms that have been implemented and tested to provide effective and automatic extraction of relevant scholarly publication information from a large selection of Web pages. The remainder of the paper is as follows. First, we give an overview of related techniques and systems. Next, we present a thorough analysis of publication citation Web pages. Effective extraction techniques and algorithms are then presented. We then describe the implementation of a system that provides a personalised monitoring service. This is followed by a performance evaluation. Finally, we present our conclusions.

Related techniques and systems

We now present a survey of current techniques and systems developed for finding and tracking information available on the WWW. It will then become apparent why we need to develop specific techniques for extracting research-related index Web pages and scholarly publications available on the Web.

Information analysis and extraction techniques

Information analysis and extraction techniques are used to analyse the content of retrieved documents, isolate specific text elements, and extract relevant information from the elements to form a database (Cowie and Lehnert, 1996). Currently, these techniques may be classified as lexicon-based, syntax-based, heuristic-based or machine learning-based.

A lexicon-based approach can be applied to extract information with contents identifiable by some common keywords, phrases, proper names, etc. Lexicon-based systems generally maintain a lexicon or list of word roots, including individual keywords and phrasal units, proper names, abbreviations, numbers, and codes that commonly appear in a particular information domain. They also include some affixes that can help recognise variations from main roots. For example, proper names, such as "IEEE" or "IEE", can be used to identify specific professional bodies, and unknown names can be predicted based on organisation or company designators (e.g. "Institute", "Laboratory", "Association", "Partners", "Corp.", "Inc." and "Ltd"). FRUMP (DeJong, 1982) is a system that uses the lexicon-based technique. Using a newswire network as its data source, FRUMP extracts news stories by applying simple relevance-matching scripts based mainly on keywords.

The syntax-based approach relies on data presented in certain predefined formats, or when the document structure is marked up using a set of predefined tags. For example, URL addresses can be easily recognised as they start with the prefix "http://". SRA (Pandit and Kalbag, 1997), WIRE (Aggarwal *et al.*, 1998), TetraFusion (Crimmins and Smeaton, 1999) and Jedi (Huck *et al.*, 1998) are examples of syntax-based systems.

In addition to lexical and syntactic knowledge, heuristic-based systems make use of additional rules to improve the extraction results. Heuristics are generally drawn from common conceptual knowledge and regular patterns observed from data sources. For example, CiteSeer uses a heuristic that the title of a publication is often written using the largest font size in the header of a document (Lawrence *et al.*, 1999a). Another heuristic, which is used to parse the citations in a reference list, is based on the "invariants first" principle. The system tries to identify the format of the first citation; the same format is then assumed across all citations.

Machine learning techniques, especially algorithms developed in natural language processing (NLP), may be applied to analyse free-texts in natural language (Kushmerick, 1999). SCISOR (Jacob and Rau, 1990) is a prototype that selects and analyses stories about corporate mergers and acquisitions from Dow Jones online financial service. The system incorporates a natural language processing component, which combines two analysis strategies, language-driven interpretation and expectation-driven processing, to identify the desired information from selected stories.

Web monitoring systems

A number of services are available to monitor the Web. Some of these systems combine several agents to provide a full-featured service from discovery, retrieval to filtering and delivery of information to users.

However, most monitoring and tracking services provided are commercially oriented. Therefore, there is not much information on their design and implementation available in the open literature.

In general, monitoring and tracking services can be classified into the following three types:

(1) *Personalised information monitoring services (PIMS)*. These systems provide monitoring of personalised information such as headline news, stock quotes, weather, and sports news. Service providers deliver updated information to a personalised Web interface provided. My Yahoo! (Yahoo!, 2001) and Infogate (Infogate, 2001) are two examples of PIMS.

(2) *News scanning and clipping services (NSCS)*. These services provide periodical scanning, filtering and clipping of interested news articles from a predefined set of information sources. Examples of NSCS include CyberAlert (Ultitech, 2001), WebClipping (AllResearch Inc., 2001), NewsHound (Knight Ridder, 2001), NewsPage (Individual.com, 2001) and CRAYON (NetPresence, 2001).

(3) *Web site monitoring and tracking services (WMTS)*. These provide monitoring and tracking of the Web site(s)/page(s) the user is interested in. They inform the user according to the updates made to them. Internet Scrapbook (Sugiura and Koseki, 1998), which supports Web page monitoring service, is an example of WMTS.

Some other systems such as Mind-it (NetMind Technologies, 2001), eWatch (Wavo, 2001) and CyberScan (2001) provide a combination of the above-mentioned services.

Discussion

Relatively simple analysis techniques, such as lexicon and syntax-based methods, are inadequate for most applications (Cowie and Lehnert, 1996). More advanced systems often combine several techniques to improve extraction results. The design of SCISOR (Jacob and Rau, 1990), for example, combines machine-learning techniques with lexical, syntactical and heuristic analysis. We have also mentioned a number of information monitoring systems. However, they are not really useful for extracting and monitoring research information from Web pages. Also, those services provided are commercially oriented.

Instead, we need specific techniques for extracting and monitoring research-oriented Web publications. Since most publication indexes are currently presented in Hypertext Markup Language (HTML) format, we focus on techniques for analysing and extracting publication information from HTML formatted contents. A syntax-based approach can be applied to HTML index pages, in which the desired information can be partly identified by locating HTML tags. In addition, we make the assumption that the most significant research works that are indexed on the Web are available in English. This is reasonable because major commercial databases such as INSPEC (IEE, 2001) also focus on English language publications. At any rate, keywords in other languages could also be incorporated into the list of keywords. Next, common keywords, such as "publication", can be used to detect research-related information. So, lexical analysis based on keyword search is also useful. Although the lexicon of well-known journal and author names may be used, it is not very effective as proper names can be cited in many different ways in different index pages. Due to a wide variation in the formats of publication index pages and citation blocks, we employ heuristics to help locate and extract the desired information. The advantage of using a combination of approaches is that we do not necessarily rely on well-composed HTML formatted pages, or other similarly strict requirements, for accurate extraction. However, advanced machine learning techniques, such as the NLP algorithms, are not useful for extracting citation information of scholarly publications, which are usually cited in a technical and abridged style.

Unlike digital libraries like CiteSeer (Bollacker *et al.*, 2000; Lawrence *et al.*, 1999b) or Rosetta (Bradshaw *et al.*, 2000), our aim is to provide a personalised monitoring service that evolves autonomously after initial human intervention. Manual input has been found effective in this context, as demonstrated by Yahoo!, the Institute for Scientific Information (ISI, 2001) and concept-based relevance feedback (Chang and Hsu, 1999). According to Bradford's law (Garfield, 1979), most of the

significant scientific results are carried by a core of important journals (Testa, 2001). A generalisation of this principle is that most researchers have a fairly good idea of where to find important research works relevant to them. Thus, we allow users to define Web sites or pages of their interest to explore. Thereafter, our techniques perform automatic retrieval and monitoring of relevant publications.

Citation standards and publication index Web pages

Citation standards for scholarly literature

A number of citation standards have been developed. In Table I we present several citation styles commonly used in scholarly literature. These include American Psychological Association (APA), Modern Languages Association (MLA), American Medical Association (AMA), Turabian and Chicago (Brandes, 2001; Walker, 2001a, b; Steward, 2001). The APA standard is often used for psychology, education, and other social sciences. MLA is used for literature, arts and humanities. AMA is mainly used for medicine, health and biological sciences. Turabian and Chicago are used for all academic subjects (Delaney, 2001).

Although an electronic document may stylistically resemble a print publication, the physical characteristics inherent in printed publications may not appear in the electronic form. Therefore, elements may be omitted or altered when electronic sources are cited. In fact, the standards for citing electronic resources are not fully established. However, several recommendations, such as ISO 690-2 (ISO, 2001) and Web Extension to APA (Land, 2001), have been developed based on

Table I Common citation styles for scholarly publications

Style	General syntax	Example	Notes
APA	[authors' names] [date of publication] [title of the work] [title of periodical] [volume, page, version, series] [place of publication]	Di Rado, A. (1995, March 15). Trekking through college: Classes explore modern society using the world of Star Trek, Los Angeles Times, p. A3	Use only the initials of the authors' first (and middle) names Date of publication in parentheses
AMA	[authors' names] [title of the work] [title of periodical] [place of publication] [date of publication] [volume, page, version, series]	1. Di Rado, A. Trekking through college: Classes explore modern society using the world of Star Trek, *Los Angeles Times*, March 15, 1995: A3	Items are listed numerically Use initials of authors' first and second names with no space
Chicago	[authors' names] [date of publication] [title of the work] [title of periodical] [volume, page, version, series] [place of publication]	Di Rado, Alicia. 1995. Trekking through college: Classes explore modern society using the world of Star Trek. *Los Angeles Times*, 15 March, sec. A, p. 3.	
MLA	[authors' names] [title of the work] [title of periodical] [place of publication] [date of publication] [volume, page, version, series]	Di Rado, Alicia. "Trekking through College: Classes Explore Modern Society Using the World of Star Trek.'' Los Angeles Times, 15 Mar. 1995: A. p. 3	Title is put in quotation marks Abbreviate the names of all months except May, June, and July
Turabian	[authors' names] [date of publication] [title of the work] [title of periodical] [volume, page, version, series] [place of publication]	Di Rado, Alicia. 1995. Trekking through college: Classes explore modern society using the world of Star Trek. *Los Angeles Times*, 15 March, A3	

standards for printed material. Table II illustrates some styles used in citing electronic online sources.

In publication index Web pages, publications are often cited according to a common style, similar to those used for electronic sources. However, since users can sometimes click on corresponding document hyperlinks to download the papers, the URL address of those links may not be explicitly listed as required by the citation standards. We have found APA and MLA to be most popular in citing online documents.

Publication index Web pages

Publication index Web pages list research works published by individual researchers, research groups or organisations. As shown in Figure 1, a section of the page is used to list references to publications. We call this section a "citation block" and each individual reference in the section is called a "citation item". Common sub-fields of a reference, such as the paper title and author(s), are called "citation attributes". Citation blocks, items and attributes are the basic citation elements of a publication index page. In general, an index page contains one or more citation blocks. Each citation block contains one or more citation items, and each item is represented by several citation attributes.

Publication index pages may be dedicated or mixed. As shown in Figure 2(a), a dedicated page contains only publications information. It is usually used by a research organisation with a large number of research publications. Figure 2(b) illustrates a mixed index page, which is common among individual researchers.

We have analysed numerous Web pages of research organisations and individual researchers to gain insight into the characteristics of publication index pages. The results of 100 of these are summarised in Table III. Our goal is to develop techniques based on our observations. In general, there are a high degree of similarity and pronounced trends among the research pages in several aspects, but there are also marked differences in others. We found that dedicated index pages outnumbered mixed index pages by a factor of 4:1. As expected, the average number of blocks per page was small. However, the maximum was found to be 22. On average, each block contained 20 items. We also found that over 83 percent were classified as online blocks. Overall, the 100 pages yielded a total of 5,409 items.

A publication index page usually contains indicative keyword(s) such as "publication" in a prominent location such as the Web page's HTML title or the citation block's header. We found that 100 percent of the mixed index pages used indicative keyword(s) in their citation block header, and 98.8 percent of the dedicated index pages used indicative keyword(s) in their title/URL. "Publication" and "paper" were the two most commonly used indicative keywords found in 67.5 percent and 26.3 percent of index pages we analysed. Also, we found three common citation block layouts as illustrated in Figure 3. Their frequencies of occurrence were 61.8 percent, 10.9 percent and 27.3 percent for the list format, table format and plain text format, respectively.

As shown in Table IV, further analysis revealed that virtually all the citation

Table II Examples of online electronic citations

Style	Example
APA	Lynch, T. (1996). DS9 Trials and Tribble-ations Review. Peoria, IL: Psi Phi: Bradley's Science Fiction Club. Retrieved October 8, 1997 from the World Wide Web: http://www.bradley.edu/campusorg/psiphi/DS9/ep/503r.html
AMA	Lynch, Tim. DS9 Trials and Tribble-ations Review. Psi Phi: Bradley's Science Fiction Club Web site. 1996. Available at: http://www.bradley.edu/campusorg/psiphi/DS9/ep/503r.html. Accessed October 8, 1997
Chicago	Lynch, Tim. 1996. DS9 Trials and Tribble-ations Review. In Psi Phi: Bradley's Science Fiction Club [online]. Peoria, IL: Bradley University, 1996 [cited 8 October 1997]. Available from World Wide Web: <http://www.bradley.edu/campusorg/psiphi/DS9/ep/503r.html>
MLA	Lynch, Tim. "DS9 Trials and Tribble-ations Review." Psi Phi: Bradley's Science Fiction Club. 1996. Bradley University. 8 Oct. 1997. <http://www.bradley.edu/campusorg/psiphi/DS9/ep/503r.html>
Turabian	Lynch, Tim. 1996. DS9 Trials and Tribble-ations Review [online]. Peoria, IL: Bradley University; available from http://www.bradley.edu/campusorg/psiphi/DS9/ep/503r.html; Internet; accessed 8 October 1997

Figure 1 Sample publication index page

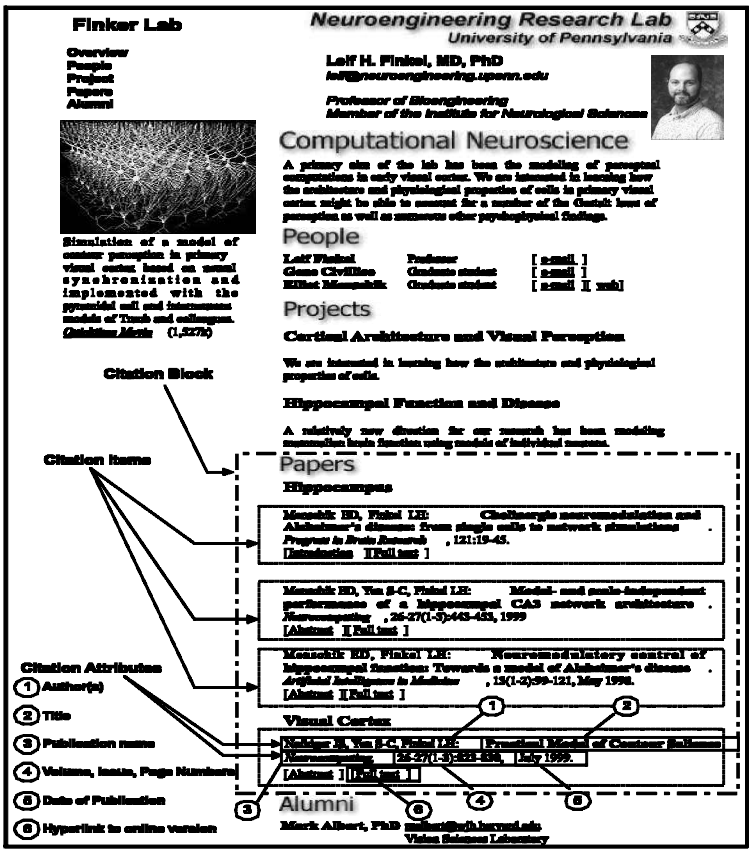
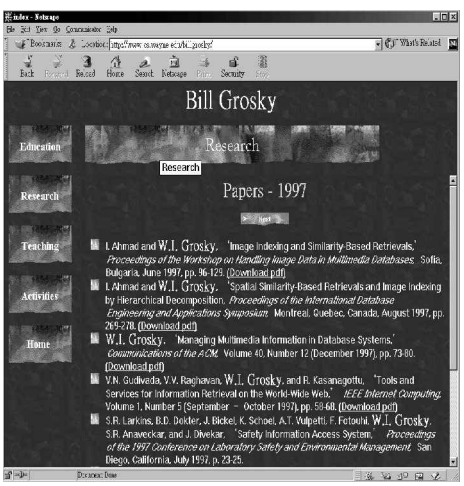
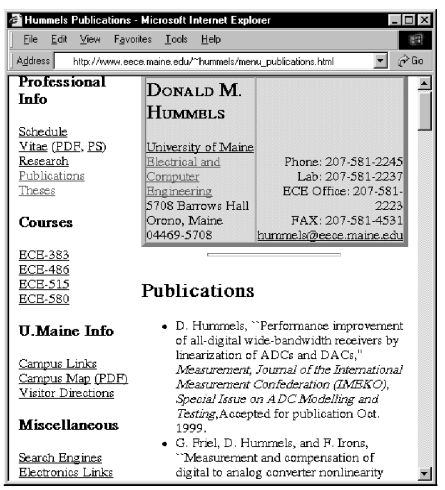


Figure 2 Index page types



(a) Dedicated page



(b) Mixed page

blocks used an author-first or title-first format.

Further, more than two-thirds of them used the author-title-other information format. "Other information" is a collective term that includes any combination of attributes such as publication name, series, pagination data, and publication date. These attributes may be arranged in different orders and some of them may be absent.

Extraction techniques and algorithms

Based on the results presented in the previous section, we now describe our techniques that automatically identify and extract citation

Table III Characteristics of publication index pages

	Maximum
Page statistics	
Total number of pages	100
Content type:	
Number of dedicated pages	81
Number of mixed pages	19
Number of blocks per page (average 2.75, minimum 1)	22
Number of items per page (average 54, minimum 1)	827
Citation block statistics	
Total number of blocks	275
Block type:	
Number of non-online blocks	46
Number of online blocks	229
Number of items per block (average 20, minimum 1)	449
Citation item statistics	
Total number of items	5,409

data of listed publications from the Web. There are two major processes involved: index page extraction/identification and publication extraction. The first process obtains publication index pages from among all Web pages presented. The second process then extracts the relevant information from those pages.

Extraction and identification of index pages

Figure 4(a) shows the index page extraction process for Web site monitoring. It attempts to extract all related Web pages from a monitored Web site and store them in a link buffer. It checks the supplied URL by exploring the hyperlinks that connect Web pages to each other. Whenever a new Web

Figure 3 Popular citation block formats

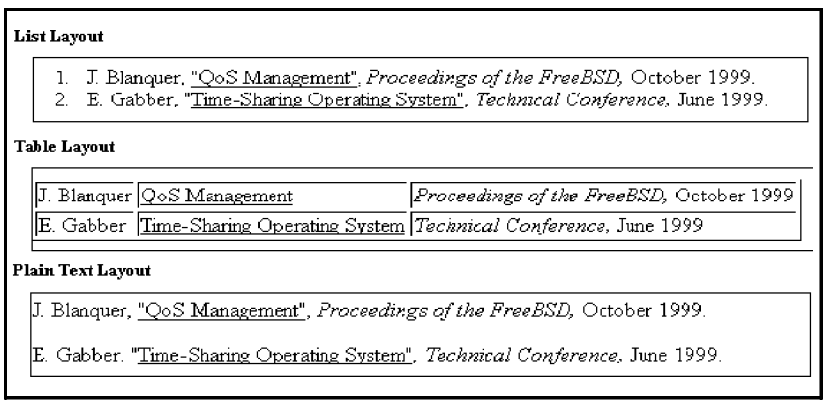
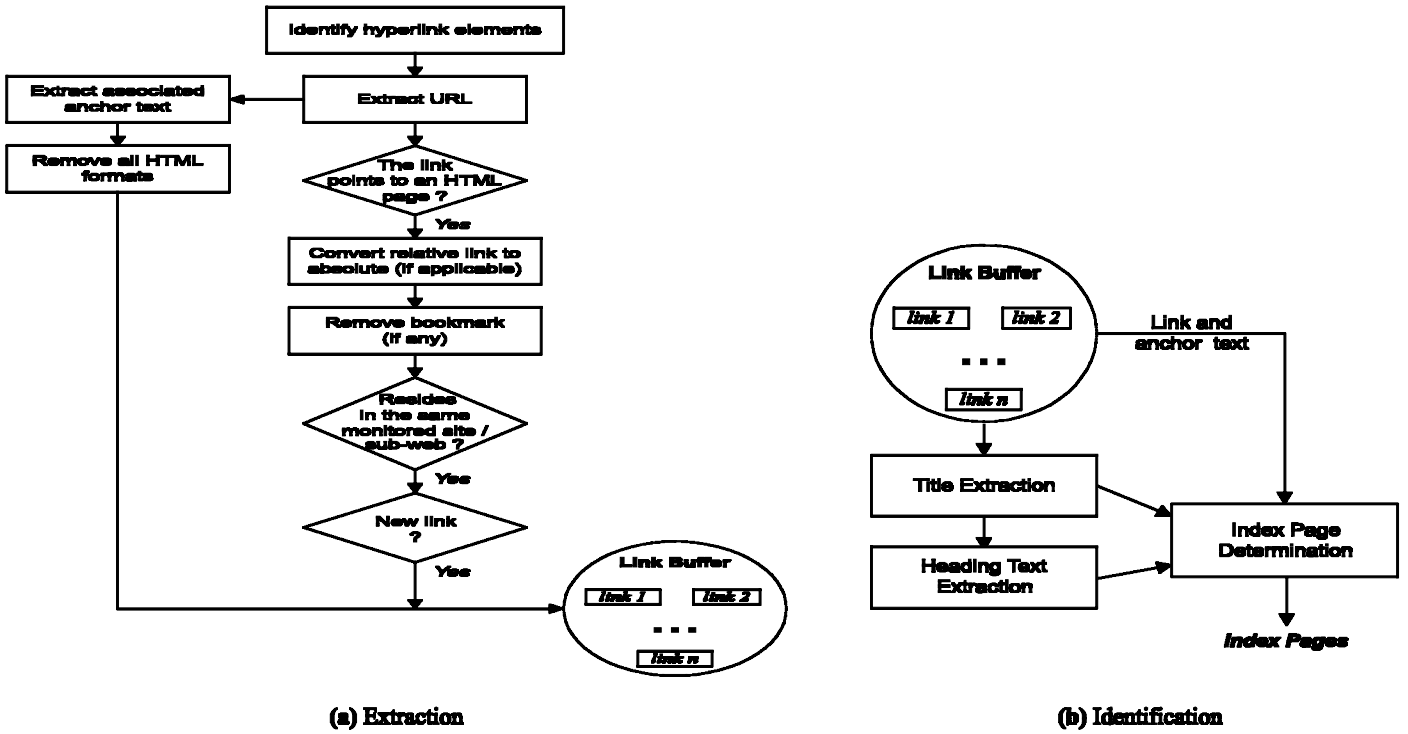


Table IV Citation formats

Type	Format	Example	Number of citation blocks (total = 275)	Percentage
Author first	Author, title, other information	J. Blanquer, and B. Ozden, "Resource Management for QoS in Eclipse", Proceedings of the FreeBSD 1999 Conference, California, October 1999, pp. 560-567	192	69.8
	Author, title	J. Blanquer, and B. Ozden, "Resource Management for QoS in Eclipse".	4	1.5
Title first	Title, author, other information	"Resource Management for QoS in Eclipse". J. Blanquer, and B. Ozden, Proceedings of the FreeBSD 1999 Conference, California, October 1999, pp. 560-567	9	3.3
	Title, author	"Resource Management for QoS in Eclipse", J. Blanquer, and B. Ozden	7	2.5
	Title, other information	"Resource Management for QoS in Eclipse". Proceedings of the FreeBSD 1999 Conference, California, October 1999, pp. 560-567.	14	5.1
	Title only	"Resource Management for QoS in Eclipse".	47	17.1
Others	Special formats	148K, "rmqe.ps". 12/1999. J. Blanquer, and B. Ozden. "Resource Management for QoS in Eclipse". Proceedings of the FreeBSD 1999 Conference, California, pp. 560-567	2	0.7

Figure 4 Web page extraction/identification



page is found, its content is analysed and all hyperlinks embedded in the page are extracted. However, only those links to the pages that reside on the same Web site or sub-Web with the main Web page are saved in the link buffer.

Next, index page identification inspects each of the Web pages from the link buffer and determines if it is a publication index page. Figure 4(b) shows the three sub-processes involved:
(1) title extraction;
(2) heading text extraction; and
(3) index page determination.

Index page determination identifies publication index pages based on the following elements: the Web page's title, URL, heading text and anchor text of the link from the parent page. Table V shows the rules used to test if a Web page is an index page. The rules are listed in descending order of importance.

A page is very likely to be a publication index page if it conforms to at least one primary rule and at least one secondary rule. In particular, we consider the following when determining the relative likelihood of finding an index page:
- number of rules satisfied;
- relative importance of the rules;
- relative importance of indicative keywords contained. For example, "publication" and "paper", which are most commonly used, carry a higher weight than other keywords.

Publication extraction

Publication extraction begins when publication index pages have been identified. After successful extraction, the relevant information is stored in a Web publication database for subsequent retrieval. Publication extraction comprises three processes:
(1) pre-processing;
(2) extraction; and
(3) post-processing.

Pre-processing

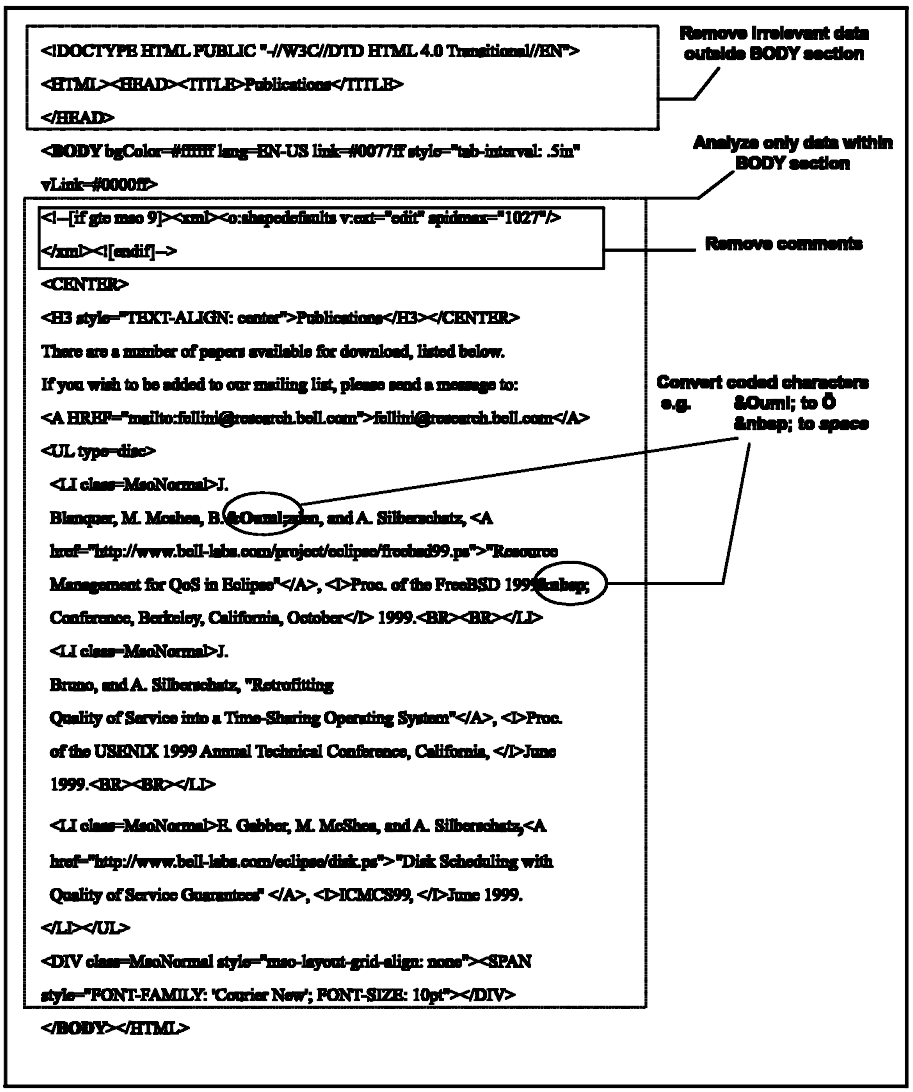
Figure 5 shows an example of pre-processing. Since citation information is contained in the body section of an index page, only this section is analysed. Also, comments based on the HTML tags "*!–*" and "*–>*" are removed. In addition, special characters such as "* *", and "*"*" are often coded in the HTML source file. These are converted into their intended formats in pre-processing.

The steps involved in extraction are summarised in Figure 6. Page content type identification determines if an index page is dedicated or mixed. Since the positions of indicative keywords in dedicated and mixed

Table V Rules for determining index pages

Group	Rules
Primary	The page's title contains indicative keywords
	The page contains indicative keywords in heading style
Secondary	The anchor text contains indicative keywords
	The page's file name contains indicative keywords
	The page contains indicative keywords in bold text
	The page's directory path contains indicative keywords
	The page contains a layout often used for structured citation blocks (i.e. HTML lists, table or block quotes)

Figure 5 Pre-processing

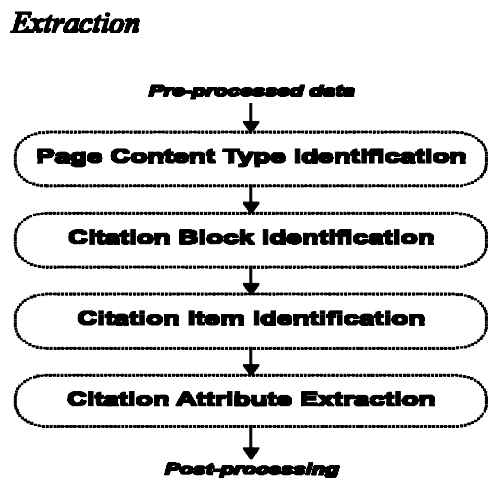


pages are different, we use an indicative keyword search approach to determine the content type of the index page. Next, citation block identification looks for potential citation blocks based on the knowledge of layout formats and the use of indicative keyword search.

Functional tags are HTML tags embedded in a citation block to define the block elements and control the layout display. Four kinds of functional tags are defined: block start tag and block end tag mark the beginning and end of each citation block; item start tag and item end tag define individual citation items in a citation block. Citation blocks can be classified as structured or unstructured based on functional tags.

As shown in Figure 7(a), three different functional tags (block start, block end and item start) are always used in a structured

Figure 6 Extraction steps



block. An unstructured block (Figure 7(b)) uses the same tags to mark the beginning and the end of the block, and to define the block items. Figure 8 illustrates our algorithm for identifying citation blocks. The algorithm is applied repeatedly to locate all citation blocks for each page.

The third extraction step is citation item identification. As shown in Figure 9, citation items can be deduced based on the layout format in which the items are presented. For example, for structured blocks using HTML lists, citation items are usually identified by the standard list item tags <LI> (for unordered and ordered lists) or <DT> (for definition list). The corresponding item end tags </LI> and </DT> can also be found. If these standard list tags are not found, the algorithm looks for paragraph break tags <P> and line break tags
.

In the final step of extraction, citation attributes are extracted. Figure 10 illustrates the citation attribute extraction process. Hyperlink information including URL and anchor text is identified first. Then, all HTML formats are removed and only the plain text is retained. This is then divided into three separate components: author, title and other information. Publication date, pagination information and publication name can be extracted from the other information components. Sometimes, the publication date may also be found in the author component. Finally, the online document link identifies the hyperlink that points to the online version of the document. Other irrelevant links (if any) are removed.

Post-processing

Since extraction relies heavily on heuristics, index page structures can affect the extracted results. Citation blocks may contain non-citation data such as brief introductions, notes or return links to main pages, etc. If these data are presented in the same layout format as citation information, they could be misidentified as citation items. Post-processing attempts to identify and remove false citation information in five steps:

(1) Title is the only mandatory attribute of citations. Items without a title are discarded.
(2) In addition, there is always at least one other attribute available. Therefore, non-online items with no other attribute(s) are removed.
(3) If the author component contains unlikely words such as “download”, “link

Figure 7 Citation blocks in HTML code

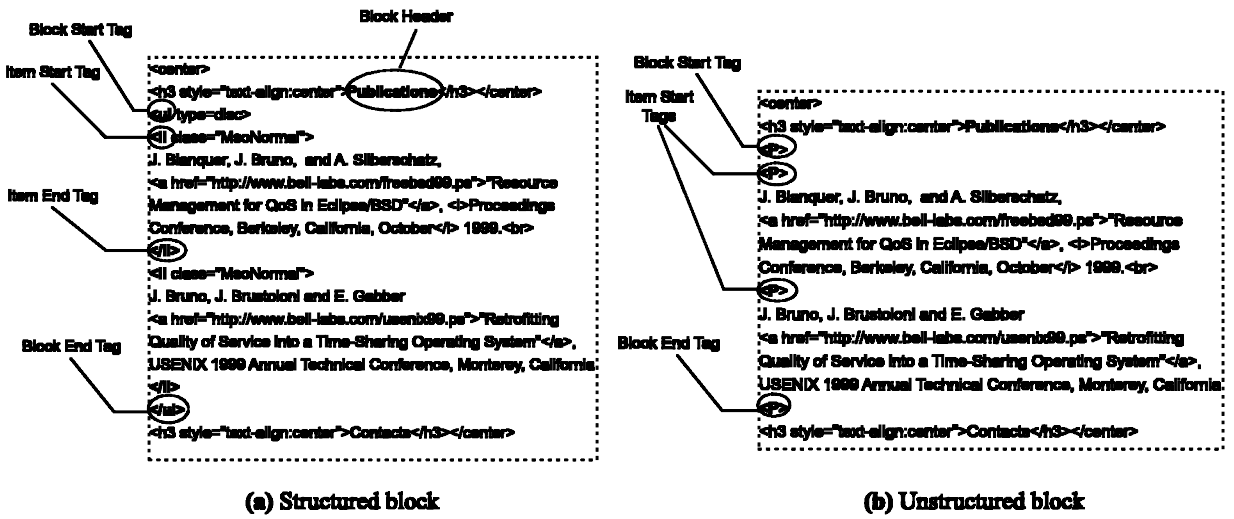


Figure 8 Algorithm for identifying citation blocks

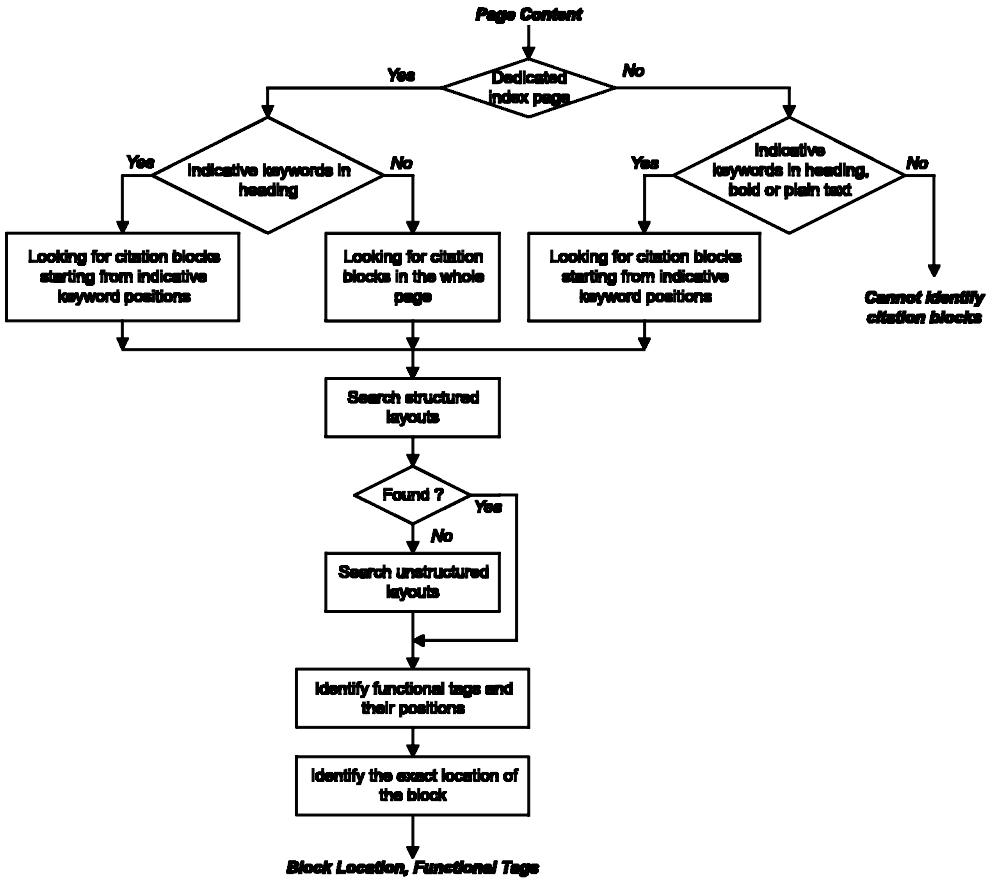
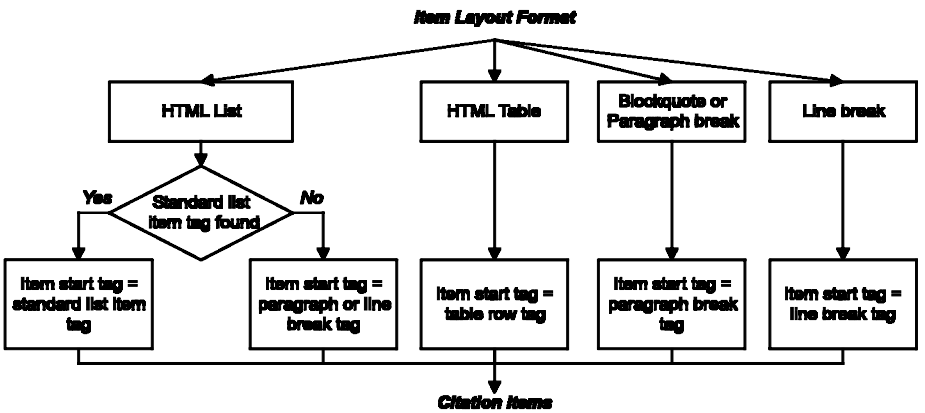


Figure 9 Algorithm for identifying citation items



to", "copyright", "@", "return to home page; and "back to", etc., the extracted item is discarded.

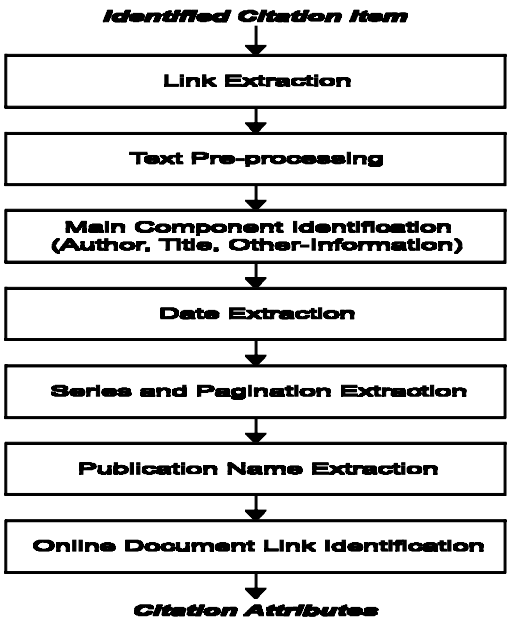
(4) If most (90 percent) citation items in a block have a pair of double quotes, then we assume double quotes are used for titles. So, items in that block without double quotes are removed.

(5) We found that the date attribute has a high consistency among citation items. Also, extraction errors often occur at the boundaries of citation blocks, where irrelevant data are misclassified. So, if most (90 percent) extracted citations in a block have publication dates, items extracted at the beginning and/or end of that block without the date attribute are removed.

System implementation

We have implemented a system known as PubWatcher that provides personalised

Figure 10 Extraction of citation attributes



research publications retrieval and monitoring services. Figure 11 shows the user interface of PubWatcher implemented using Java applets and can be launched from any standard Web browser. The user may then specify Web sites to be monitored as shown in Figure 12. Each user can therefore maintain a personalised set of Web sites. By default, updates are activated daily at night to process monitored Web sites specified by the user.

Performance evaluation

Based on the techniques and algorithms described, we implemented and tested the index page extraction/identification process and the publication extraction process summarised in Figure 13. The index page extraction/identification process is applied to both new and existing Web sites specified by the user. A comparator is used to check for necessary updates to the URL database. The publication extraction process is then applied to extract new publication data. A second comparator is used to check for necessary updates to the publication database. The final output is used to update the personal folder for each user and to send an e-mail notification to inform the user of new updates.

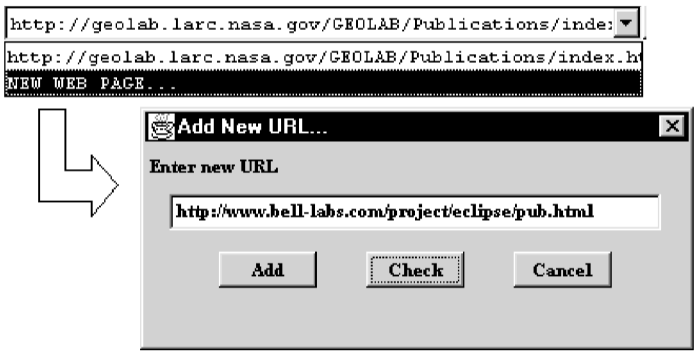
In our tests, we used ten Web sites (five from universities, and five from research centres/laboratories; see Appendix) with 14,115 Web pages to test the index page extraction/identification process, and we used the corresponding 225 index pages to test the publication extraction process.

Index page extraction/identification

Our index page extraction technique successfully extracted all 14,115 Web pages found in the test Web sites. The extracted Web pages were then parsed through the index page identification process, which

Figure 11 PubWatcher user-interface

Figure 12 Updating a personal record of Web sites to be monitored



correctly identified 96.1 percent of all the 225 index pages. Among the unsuccessful attempts, the number of false accepts outnumbered the number of false rejects by 2:1. The index page in Figure 14(a) contains only one keyword, "Publications", in the heading style and is falsely rejected. On the other hand, the page in Figure 14(b) was misidentified, as it satisfied the requirements of an index page. However, it does not contain any publications, but lists all the links connected to different publication index pages.

Publication extraction

We conducted tests on the effectiveness of the four steps involved in the extraction process. Our content type identification technique correctly identified all 59 mixed index pages and correctly identified 97.6 percent of the 166 dedicated pages. Next, we used the 221 correctly identified index pages to test our citation block identification technique, which successfully identified 96.8 percent of the 469 blocks contained in the pages. The 454 successfully identified blocks were then used to test our citation item identification technique. We found that 99.7 percent of the 7,782 citation items were correctly identified. Finally, we used the 7,758 correctly processed citation items to test our citation attribute extraction step, which achieved a success rate of 93.6 percent.

Conclusion

In this paper, we have outlined the motivation for providing a means of finding Web publications to promote the exchange of ideas among members of the research

Figure 13 Summary of the entire monitoring process

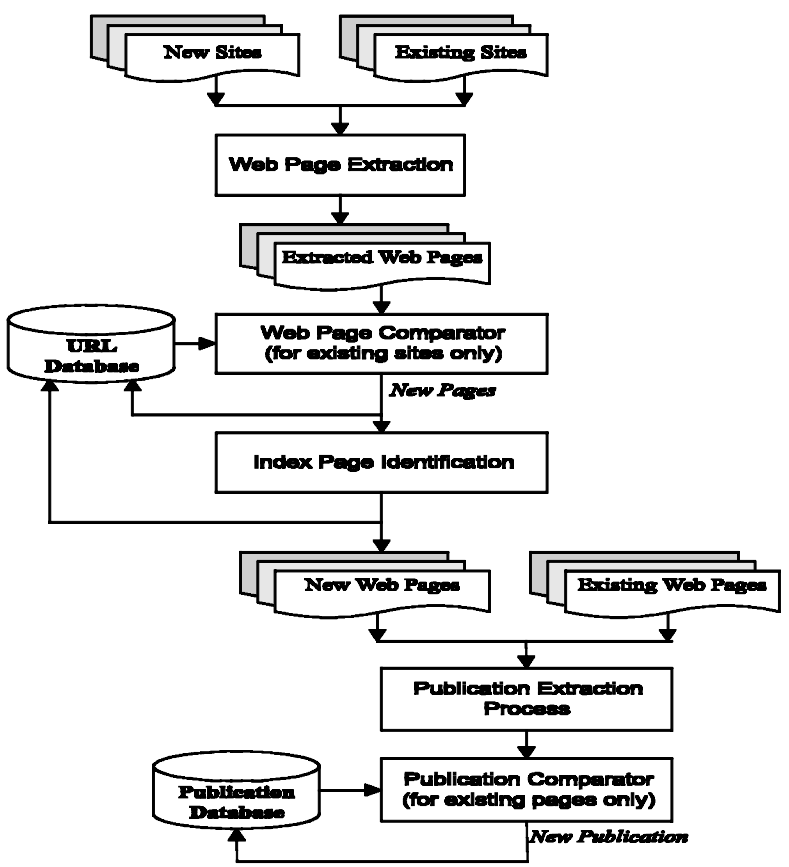
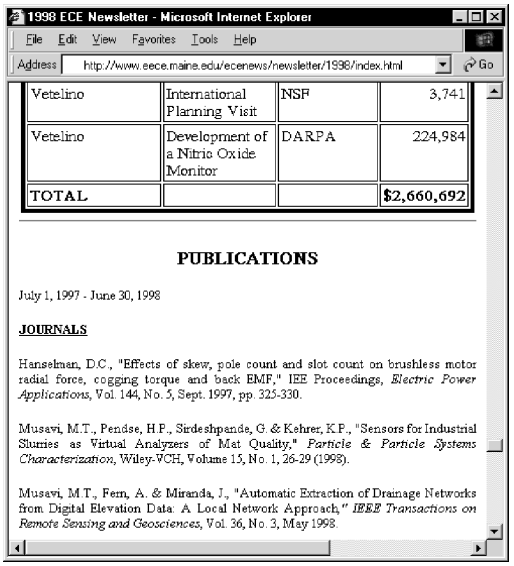
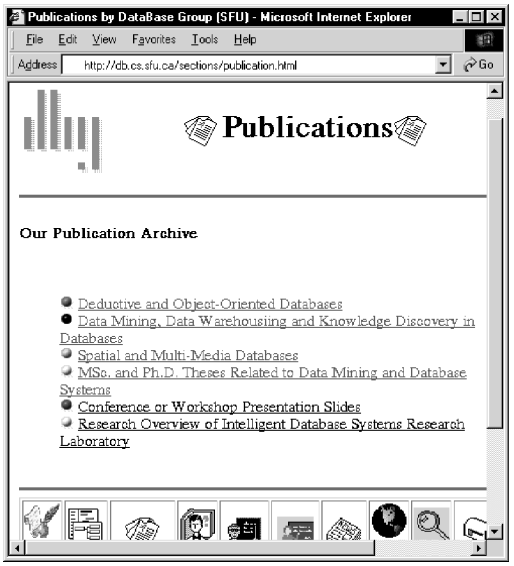


Figure 14 Examples of unsuccessful identifications



(a) False reject



(b) False accept

communities. In response to this, we have developed a series of techniques and algorithms that can perform the task of automatic monitoring, tracking and extraction of Web publications.

Our techniques correctly identified almost all of the publication index pages from among 14,115 test Web pages. Our four-step publication extraction process then successfully retrieved more than 90 percent of useful publication information. Unlike approaches that rely on neural networks, our algorithms do not require training/retraining to achieve useful results.

References

Aggarwal, S., Hung, F. and Meng, W. (1998), "WIRE – a WWW-based information retrieval and extraction system", *9th International Workshop on Database and Expert Systems Applications (DEXA'98)*, 26-28 August, Vienna.

AllResearch (2001), "WebClipping Service home page", http://www.webclipping.com/

Bollacker, K., Lawrence, S. and Giles, C. (2000), "Discovering relevant scientific literature on the Web", *IEEE Intelligent Systems*, Vol. 15 No. 2, pp. 42-7.

Bradshaw, S., Scheinkman, A. and Hammond, K. (2000), "Guiding people to information: providing an interface to a digital library using reference as a basis for indexing", *Proceedings of the International Conference Intelligent User Interfaces*, pp. 37-43.

Brandes, J. (2001), "Citing the World Wide Web in style: APA and MLA formats", Troy State University Florida Region Central Library, Troy State University, http://www.tsufl.edu/library/5/citation.htm

Chang, C.-H. and Hsu, C.-C. (1999), "Enabling concept-based relevance feedback for information retrieval on the WWW", *IEEE Trans. Knowledge and Data Engineering*, Vol. 11 No. 4, pp. 595-609.

Cowie, J. and Lehnert, W. (1996), "Information extraction", *Communications of the ACM*, Vol. 39 No. 1, pp. 80-91.

Crimmins, F. and Smeaton, A.F. (1999), "TetraFusion: information discovery on the Internet", *IEEE Intelligent Systems*, Vol. 14 No. 4, pp. 55-62.

CyberScan (2001), "Personalised Internet clipping services from CyberScan", http://www.clippingservice.com/

DeJong, G.F. (1982), "An overview of the FRUMP system", in Lehnert, W.G. and Ringle, M.H. (Eds), *Strategies for Natural Language Processing*, Erlbaum, Hillsdale, NJ, pp. 149-76.

Delaney, R. (2001), "Citation style for research papers", Davis Schwartz Memorial Library, http://www.cwpost.liunet.edu/cwis/cwp/library/workshop/citation.htm

Garfield, E. (1979), *Citation Indexing*, John Wiley & Sons, New York, NY.

Huck, G., Fanklauser, P., Aberer, K. and Neuhold, E. (1998), "Jedi: extracting and synthesizing information from the Web", *Proceedings of the Third IFCIS Conference on Cooperative Information Systems*, pp. 32-43.

Individual.com (2001), "Individual.com NewsPage Home page", http://www.individual.com/

Infogate (2001), "Infogate", http://www.infogate.com/

Institution of Electrical Engineers (IEE) (2001), "INSPEC, the leading English-language bibliographic information service", http://www.iee.org/Publish/INSPEC/ISI, Institute for Scientific Information (2001), http://www.isinet.com

International Organisation for Standardization (ISO) (2001), "ISO 690-2, Bibliographic references to

electronic documents", http://www.nlc-bnc.ca/iso/tc46sc9/standard/690-2e.htm

Jacob, P.S. and Rau, L.F. (1990), "Scisor: extracting information from online news", *Communications of the ACM*, Vol. 33 No. 11, pp. 87-97.

Knight Ridder (2001), "NewsHound Home page", http://www.newshound.com/

Kushmerick, N. (1999), "Gleaning the Web", *IEEE Intelligent Systems*, Vol. 14 No. 2, pp. 23-7.

Land, T. (2001), "Web extension to American Psychological Association Style (WEAPAS)", Rev. 1.6. http://www.beadsland.com/weapas/

Lawrence, S., Bollacker, K. and Giles, C.L. (1999a), "Indexing and retrieval of scientific literature", *Proceedings of the Eighth International Conference on Information Knowledge Management*, Kansas City, MO, pp. 139-46.

Lawrence, S., Giles, C.L. and Bollacker, K. (1999b), "Digital libraries and autonomous citation indexing'', *IEEE Computer*, Vol. 32 No. 6, pp. 67-71.

NetMind Technologies (2001), "Welcome to Mind-it", http://mindit.netmind.com/mindit.shtml

NetPresence (2001), "CRAYON home page", http://www.crayon.net/

Pandit, M.S. and Kalbag, S. (1997), "The selection recognition agent: instant access to relevant information and operations", *Proceedings of the 1997 International Conference on Intelligent User Interfaces*, Orlando, FL, pp. 47-52.

Steward, M.D. (2001), "Chicago style: main", Online Writing Center, University of Minnesota, http://rhetoric.agoff.umn.edu/Rhetoric/Student/Graduate/MStewart/ECD/cmain.htm

Sugiura, A. and Koseki, Y. (1998), "Internet scrapbook: automating Web browsing tasks by demonstration", *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*, San Francisco, CA, pp. 9-18.

Testa, J. (2001), "The ISI database: the journal selection process", http://www.isinet.com/whatshot/essays/-199701.html

Ultitech (2001), "CyberAlert: Internet monitoring and Web clipping service", http://www.cyberalert.com/

Walker, J.R. (2001a), "APA-style citations of electronic sources'', *The Columbia Guide to Online Style*, V. 1.0, http://www.cas.usf.edu/english/walker/apa.html

Walker, J.R. (2001b), "MLA-style citations of electronic sources'', *The Columbia Guide to Online Style*, V. 1.3, http://www.cas.usf.edu/english/walker/mla.html

Wavo (2001), "eWatch: accurate, comprehensive, trusted Internet monitoring", http://www.ewatch.com/

Yahoo! (2001), "My Yahoo! – personal home page", http://my.yahoo.com/

Appendix

The Web pages used in our tests are listed in Table AI.

Table AI Test Web sites

	Site	URL	Number of Web pages	Number of index pages
1	Database Group, Simon Fraser University	http://db.cs.sfu.ca/index.html	74	7
2	Projects of Bell Laboratories Innovations, Lucent Technologies Inc.	http://www.bell-labs.com/project/	815	13
3	Publications of Circuit Theory Laboratory, Helsinki University	http://www.aplac.hut.fi/publications/	306	1
4	Electrical and Computer Engineering Department, University of Maine	http://www.eece.maine.edu/	1,549	15
5	Abbas El Gamal Professor, Information Systems Laboratory, Stanford University	http://www-isl.stanford.edu/~abbas/	13	2
6	Information Technology Laboratory, National Institute of Standard and Technology, USA	http://www.itl.nist.gov/	9,952	77
7	GEOLAB Home page, NASA	http://geolab.larc.nasa.gov/	77	2
8	Ultrafast Laser Physics Laboratory, Institute of Quantum Electronics	http://www.iqe.ethz.ch/ultrafast/	99	5
9	Forestry Sciences Laboratory, Forest Service, USA	http://www.rtp.srs.fs.fed.us/	917	30
10	Electronic Visualization Laboratory, University of Illinois	http://www.evl.uic.edu/EVL/	313	3