Automated Medical Citation Records Creation for Web-Based On-Line Journals

Daniel X. Le, Loc Q. Tran, Joseph Chow
Jongwoo Kim, Susan E. Hauser, Chan W. Moon, George R. Thoma
National Library of Medicine, Bethesda, MD20894
daniel_le@nlm.nih.gov

Abstract

With the rapid expansion and utilization of the Internet and Web technologies, there is an increasing number of on-line medical journals. On-line journals pose new challenges in the areas of automated document analysis and content extraction, database citation records creation, data mining, and other document related applications. New techniques are needed to capture, classify, analyze, extract, modify, and reformat Web-based document information for computer storage, access, and processing. At the National Library of Medicine (NLM) we are developing an automated system, temporarily code-named WebMARS for Web-Based Medical Article Record System, to create citation records for the MEDLINE® database. The system downloads and classifies Web document articles, parses and labels the article contents, extracts and reformats the citation information from the article, presents the entire citation to operators for reconciling (validation), and uploads the citation records to the MEDLINE database.

1. Introduction and background

NLM's MEDLINE database contains about 11 million bibliographic records from over 4,300 journals and it adds 40,000 citation records every month. Currently, the majority of MEDLINE citation records [1] are entered by either manual and labor-intensive keyboarding, or scanning and optical character recognition (OCR) input using Medical Article Records System (MARS) [2, 3]. Since an increasing number of journal publishers are using the Internet and the World Wide Web to provide their subscribers with access to on-line journal issues, we see the benefit of having a system that can handle these Web journal issues. Two obvious advantages of such a Web-based document analysis and content extraction approach compared to either manual keyboard entry or the scanning/OCR approach are saving labor costs and improving performance (speed and accuracy). Instead of manually keying document information, or scanning and OCR-converting paper documents, Web journals can be downloaded and stored as HTML or PDF files and then processed. Since most on-line journals are available in either HTML or PDF file format, the amount of additional typing required to validate and correct typographical errors and other problems is reduced.

To take advantage of journal information available over the Internet, we are developing an automated system temporarily code-named WebMARS to create citation records for the MEDLINE database from on-line journals. This system: (1) downloads and classifies Web document articles as abstract, full text, PDF or image files, (2) converts PDF files into HTML files, if necessary, (3) parses HTML files to create text zones, and labels these text zones, (4) extracts, modifies, and reformats the citation information in labeled text zones to be reconciled (validated) by an operator, and (5) finally uploads the citation records to another NLM database for indexing by experts.

2. System overview

WebMARS consists of seven servers and two types of operator workstations: *download and classification*, and *reconciling.* A high-level diagram of the WebMARS system is shown in Figure 1. All workstations and servers are networked via a LAN and communicate through the WebMARS database server.

Briefly, the WebMARS system works as follows. The operator downloads the Web pages of articles, and the PDF or HTML files are sent to the file server. The *PDF to HTML files conversion server* performs text conversion. The *article zones creation and labeling server* extracts text zones from the HTML files, and labels these zones as belonging to "title", "author", "affiliation", "abstract", “pagination”, “databank accession number”, “grant number”, etc. The *article zone contents modification server* extracts, modifies, and reformats the text according to MEDLINE database conventions. The *MEDLINE article zones creation and Web-based reconciling workstation* selects and combines the appropriate portions of zone contents to create records that are available for validation and reconciling. The reconciling workstation operator then checks, proofreads, and corrects any remaining problems in the records including character and symbol validation, citation reformatting, field label selection and confirmation. The *SGML-based MEDLINE citation records creation server* generates text zones directly from journal publisher SGML information, if available, that can be used to further improve labeling process. Finally, the *MEDLINE citation records uploading server* uploads the completed citation records to other NLM systems for later indexing.

3. System servers and workstations

Each subsystem of the WebMARS system is described below.

3.1 Web files collection and classification workstation

This workstation downloads and classifies Web-based files as abstract, full text, PDF or image files [4]. It consists of two processes: *downloading* and *classification*. The first is based on *WinInet* software tool and a combination of the *Breadth First Search* algorithm and the *Constraint Satisfaction* method to traverse the Web page’s links, recognize and generate the successors of the downloading pages. The second relies on the contents of the hyperlinks (name and address) to classify the files.

The downloading process shown in Figure 2 selects the Web address of the start page, and sends a request to the Web server to download that page. During download, the links of the current downloading page are classified, and the children are generated and added to the end of the open list (consisting of addresses of pages waiting to be downloaded). Then the successfully downloaded page is classified and saved to a directory in the file server corresponding to its type, and the address of the start page is added as a node to the closed list (consisting of addresses of pages that are successfully downloaded). The open list now contains all of the successor links of the downloaded page that are waiting to be downloaded. The downloading process continues selecting the first node of the open list to download. If the download is successful, the node is removed from the open list and added to the closed list. Otherwise, the node is_put into the revisited list (consisting of addresses that failed during the downloading process, and are to be revisited later). The children of the successfully downloaded page are determined by the classification process and added to the end of the open list. The entire process is repeated until the open list is empty.

The classification process validates the hyperlinks to make sure that they have satisfied the predefined constraints for saving the downloaded Web page and generating its children. Since the contents of the hyperlinks (name and address) consist of useful information about file types, these can be used to determine the appropriate storage for saving the documents and the satisfied child links of the currently downloading page. The downloaded Web pages are saved in different directories in the file server corresponding to their types. A hyperlink, which satisfies the constraints to be a potential successor of the current downloading web page, is generated as a child node and added to the open list. Otherwise, it would be eliminated.

3.2 PDF to HTML files conversion server

Occasionally, publishers offer certain Web journal articles as PDF files rather than as HTML files. Since the WebMARS requires the HTML file format, this conversion server subsystem is used to convert downloaded PDF files into HTML files with available commercial software.

3.3 Article zones creation and labeling server

This server subsystem first parses and creates text zones for each journal article based on its HTML file, and then labels these text zones. The subsystem consists of three modules: the *automated zoning module*, the *automated labeling module*, and the *automated updating module*.

The automated zoning module relies on HTML tags to parse and then to create text zones for each article of a journal issue.

The automated labeling module labels these text zones using a combination of statistics and fuzzy rule-based technology. Statistics are used to generate membership functions for the fuzzy rules-based method. Features and fuzzy rules derived for the automated labeling module are based on an analysis of the HTML layouts of journals. Both geometric and non-geometric features are considered here. Geometric features of a zone are based on location and order of appearance. Non-geometric features are derived from contents of zone, statistics, and font characteristics. Since text zones are characterized by the words they contain, word matching is an important operation in this module. For example, a zone has a higher probability of being labeled as “affiliation” when it contains words related to country, city, and school names. Also, a zone located between the words “abstract” and “keywords” has a higher probability of being labeled as “abstract” than other labels. For this purpose, fourteen word lists have been created, and the Ternary Search Tree algorithm [5] is used as a search engine for the word matching.

The automated updating algorithm is based on a difference analysis between text zones generated by this subsystem and corresponding text zones corrected by operators during the reconcile stage. Statistical labeled text zone information obtained from this analysis can be used to improve the labeling accuracy.

3.4 Article zone contents modification server

This subsystem extracts, modifies, and reformats text in labeled zones according to MEDLINE database conventions. It consists of two modules: the *label cleanup* module and the *cleanup coach* module. The first prepares bibliographic record information for the reconcile operator, while the second operates on post-reconcile data to collect statistics in order to develop new rules to improve the performance of the first module.

The label cleanup module removes, replaces and/or processes HTML tags and special characters in labeled zones, and reformats the “title”, “author”, “affiliation” and “abstract” zones. Figure 3 shows examples of author and abstract before and after this process. The label cleanup module relies on rules derived by analyzing the page layout for each journal. Generic rules cover situations that occur in every journal, and for journals for which we have no a priori data. In addition, there will be rules specific for a particular journal. For unknown tags or confusing text, the results of any processing will be enclosed in special symbols such as curly brackets, { }, and the enclosed characters will be highlighted to alert reconcile operators.

Since rules implemented in the label cleanup module are used to prepare labeled zone contents for reconciliation by operators, the final corrected zone contents can be used to improve the module’s performance. Given that a journal issue has been reconciled, the cleanup coach module automatically extracts differences between the final correct labels and the corresponding output from the label cleanup module. These differences may be analyzed manually or automatically to modify existing rules, or to infer additional rules for the label cleanup module.

3.5 MEDLINE article zones creation and Web-based reconciling workstation

This workstation selects and combines the appropriate portions of zone contents to create MEDLINE citation records for reconciling. The reconciling operator then checks, proofreads, searches, and corrects any remaining problems in the citation records including character and symbol errors, reformatting citations, capturing extra fields if necessary (pagination, data accession number, grant number, etc.), text zone labels selection and validation.

3.6 MEDLINE citation records uploading server

This subsystem creates an XML file based on NLMCommon DTD [1] for completed citation records of a journal issue and uploads the file to other NLM systems for later indexing.

3.7 SGML-based MEDLINE citation records creation server

Some journal publishers also supply to NLM some portions of MEDLINE citation records such as “title”, “author”, “abstract”, “pagination” in SGML format. However, the records are not always complete, so additional effort is often required to capture missing information. Since the SGML information even if incomplete is usually correct, this server subsystem exploits such information by generating text zones directly from it. These text zones will be used to further improve the labeling and reconciliation processes.

4. Summary

This paper describes our WebMARS system designed to create medical citation records for the MEDLINE database directly from on-line journals. The advantages of WebMARS over either manual keyboard entry or the scanning/OCR approach are saving labor costs and improving performance. The performance of our WebMARS prototype has been evaluated using a sample of 28 medical journal Web sites. Preliminary evaluation results show the feasibility of our design to create MEDLINE citations directly and effectively using Web document pages.

5. References

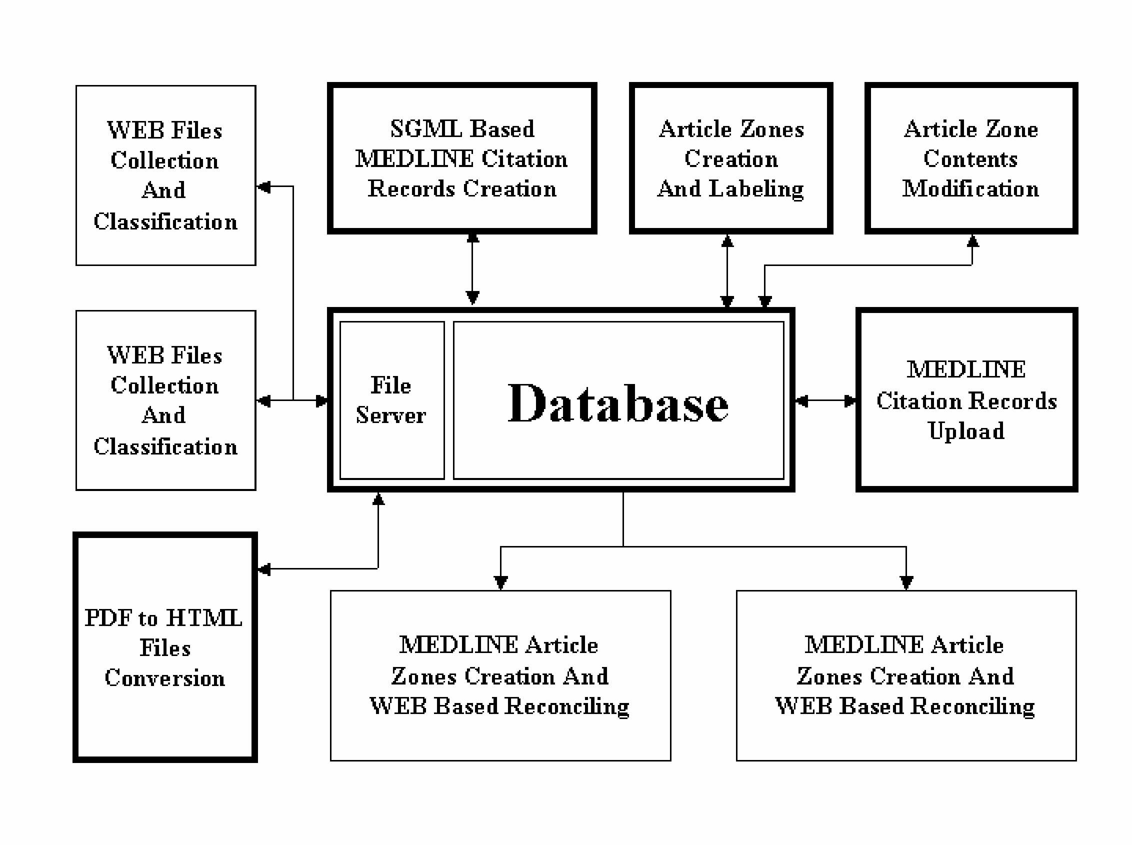
[1] http://www.nlm.nih.gov/bsd/licensee.html

[2] D. Le, J. Kim, G. Pearson, and G. Thoma, "Automated Labeling of Zones from Scanned Documents," *Proceedings 1999 Symposium on Document Image Understanding Technology*, pp.219-226, 1999.

[3] G. Thoma and D. Le, "Medical database input using integrated OCR and document analysis and labeling technology," *Proceedings 1997 Symposium on Document Image Understanding Technology*, pp. 180-181, 1997.

[4] L. Tran, C. Moon, D. Le, and G. Thoma, "Web Page Downloading and Classification," *The Fourteenth IEEE Symposium on Computer-Based Medical Systems*, July 2001. (Accepted for publication)

[5] J. Bentley and B. Sedgewick, Ternary Search Trees, *Dr. Dobb's Journal*, pp. 20-25, April 1998.



WEB Files Collection And Classification
SGML Based MEDLINE Citation Records Creation
Article Zones Creation And Labeling
Article Zone Contents Modification
WEB Files Collection And Classification
File Server
Database
MEDLINE Citation Records Upload
PDF to HTML Files Conversion
MEDLINE Article Zones Creation And WEB Based Reconciling
MEDLINE Article Zones Creation And WEB Based Reconciling

Figure 1: WebMARS system diagram (Bold outlines show servers)

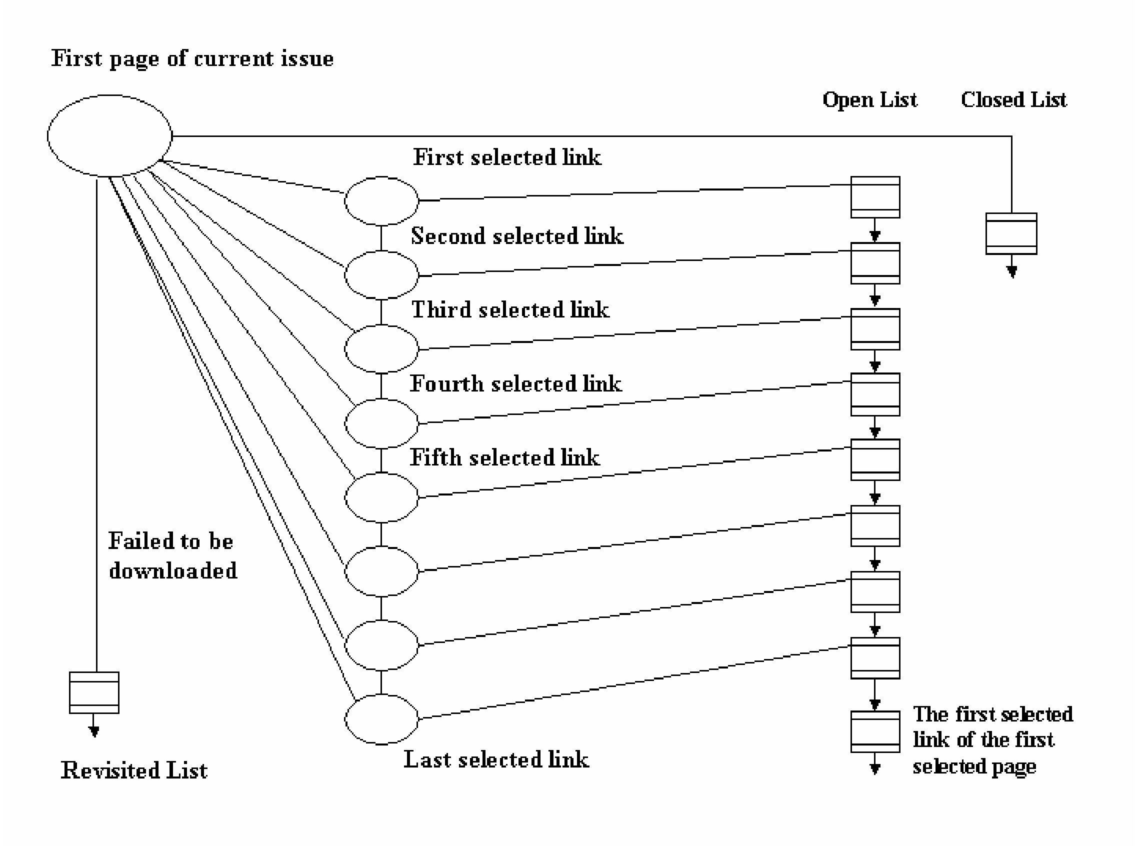


Figure 2: Dowloading process

Author label input:

<NOBR>Henry Abriel, MD, PhD</NOBR>
<NOBR>Michael V. Wehrens, MSc</NOBR>

Author label output:

Abriel H
Wehrens MV

Abstract label input

<I>Background</I>&#151;Multiple mutations of <I>SCN5A</I>, the gene that<SUP> </SUP> encodes the human Na<SUP>+</SUP> channel <IMG SRC="/math/agr.gif" ALT="{alpha}" BORDER="0">-subunit, are linked to 1 form…

Abstract label output

BACKGROUND: Multiple mutations of SCN5A, the gene that encodes the human Na(+) channel alpha-subunit, are linked to 1 form

Figure 3: Examples of author and abstract before and after the label cleanup process