Automated Cleanup Processing for Extracting Bibliographic Data from Biomedical Online Journals

In Cheol Kim, Daniel X. Le, and George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894
{ickim, daniel, gthoma}@mail.nih.gov

ABSTRACT

An R&D division of the National Library of Medicine (NLM) has developed the *Web-based Medical Article Records System* (WebMARS) to create citations from online biomedical journals. This paper presents one important part of this system, the automated cleanup module that extracts bibliographic information from HTML-formatted text based on a rule-based approach. A learning scheme comparing the output of the cleanup module to the verified processing result is newly introduced to create and update cleanup rules automatically, thereby minimizing the manual effort for rule setting and improving the performance of the cleanup processing. Experimental results show that the proposed automated cleanup module can effectively detect and extract the bibliographic data of interest from HTML-formatted online journal articles using relevant rules identified through the learning process.

Keywords: Automated cleanup processing, Online biomedical journal, Bibliographic information, Rule-based approach, Learning process, MEDLINE® database, National Library of Medicine.

1. INTRODUCTION

In recent years, with the rapid expansion and utilization of the Internet and Web technologies, increasing numbers of journal publishers provide online journals to their subscribers. As more online journals become available, effective retrieval of information and access to these resources have become increasingly important in the research fields of online document analysis, text data mining, the named entity task, etc [1][2][3].

To create citations from online biomedical journals for the MEDLINE database, the Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine (NLM) has developed an automated system, the *Web-based Medical Article Records System* (WebMARS) to download and analyze Web-based journal articles, and to extract bibliographic information [4][5]. This system consists of several modules: downloading Web document articles (WebPageCollection), zoning and labeling HTML-formatted text (WebLabeling), extracting and reformatting citation information (LabelCleanUp), verifying the processing results (WebReconcile) by human operator, and uploading the citation records to the MEDLINE database (Upload). This paper introduces a learning scheme to improve the performance of the LabelCleanUp module.

The current LabelCleanUp module employs a rule-based approach for extracting important bibliographic data from the labeled text zones created by the WebLabeling module: i.e., by journal-specific rules for extracting the title, author, pagination, and affiliation, and rules that are not journal-specific for identifying grant number, databank accession number, support, e-mail, and zip code. The problem in the current rule-based cleanup processing is that rules for new journal titles must be manually formulated, thereby requiring considerable labor and long processing times. Moreover, these journal-specific rules have to be updated manually whenever the article layout format is changed by the journal publisher.

In order to overcome the drawbacks of such manual rule setting required in the current module, we propose an automated LabelCleanUp module in which both journal-specific and non-journal-specific rules for new journals and/or new formats are automatically created and updated through an automated learning process. The learning process compares the outputs of the WebReconcile, LabelCleanUp, and WebLabeling modules to extract: (1) journal-specific delimiters for processing the title, author, pagination, and affiliation zones, and (2) non-journal-specific formats for detecting grant numbers and databanks. This learning strategy is attractive because it can minimize the manual effort required for rule setting and updating, and improve the performance of cleanup processing.

As an alternative approach, a dynamic updating technique [6] that automatically obtains an optimal feature from a combination of multiple features extracted from several previous issues in the same journal can be considered. However, this method requires a considerable degree of computational effort to estimate optimal features from all possible combinations of multiple previous features, while our proposed approach can extract an optimal feature or rule through a single learning process.

In experiments on a test set of HTML-formatted text zones obtained from online biomedical journals, we demonstrate the effectiveness of the automated cleanup process by comparing its accuracy of detecting citation information with that of the current cleanup process. The organization of the rest of this paper is as follows. In Section 2, we provide a detailed description of the proposed automated LabelCleanUp module including learning procedure for extracting cleanup rules. Section 3 provides the experimental results of cleanup processing and error analysis. Conclusions are presented in Section 4.

2. AUTOMATED CLEANUP PROCESSING

The block diagram in Fig. 1 shows the overall workflow of the proposed automated LabelCleanUp module. Once the HTML-formatted text data of the first article of a new journal title is given, the cleanup process removes unnecessary HTML tags, extracts citation information using predefined default rules, and finally reformats the syntax of such information according to MEDLINE conventions (e.g., rearranging the order of first and last names for an author). The default cleanup rules are based on the most frequent existing rules or delimiters used in previous cleanup processing tasks. Next, the output of LabelCleanUp module is compared to that of WebReconcile module validated by a human operator. If these outputs are not identical, the learning procedure is then activated to identify the journal-specific delimiters and new formats. For the learning process, the outputs of WebReconcile and LabelCleanUp modules are assumed to be the target value and actual output, respectively. Finally, all extracted rules are saved in the WebMARS database for cleanup processing of the next journal articles.

Since labeled HTML-formatted text zones resulting from the zoning and labeling stages (WebLabeling) have different layout formats and the citation information of interest embedded in each zone is different, the proposed automated LabelCleanUp module employs zone-dependent learning procedures. That is, the learning processes for the contents of the author zone and the pagination zone (for example) are different. The detailed learning procedures for each zone are described below.

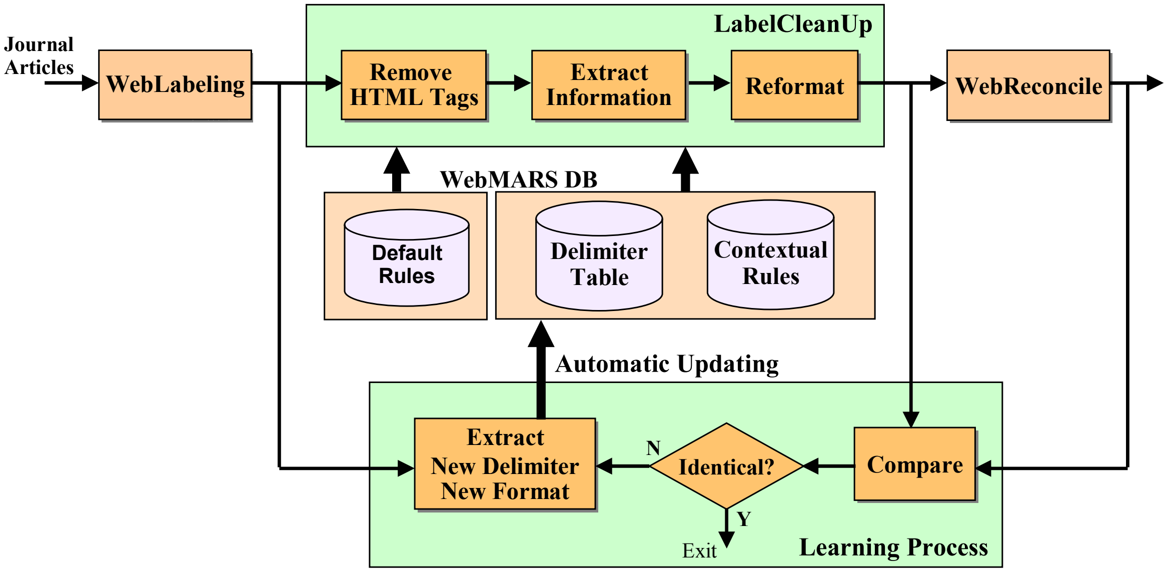

Fig. 1. Overall workflow of automated cleanup process.

Journal	Circulation Research			
Labeled text data (Author)	<STRONG> <NOWRAP> Mariana Mondrag&oacute;n-Palomino,<SUP>1</SUP><WBR> Blake C. Meyers,<SUP>2</SUP><SUP>,</SUP><A HREF="#FN3"><SUP>3</SUP></A><WBR> Richard W. Michelmore,<SUP>2</SUP> and<WBR> Brandon S. Gaut<SUP>1</SUP><SUP>,</SUP><A HREF="#FN4"><SUP>4</SUP></A> </NOWRAP> </STRONG>			
Reconcile output	Mondragón-Palomino, Mariana Meyers, Blake C Michelmore, Richard W Gaut, Brandon S			
Delimiters	Front	**<NOWRAP>**	Embedded	**<WBR>**

(a)

Journal	Physiological Genomics			
Labeled text data (pagination)	<BR> <EM>Physiological Genomics</EM> 4:165-174 (2001)			
Reconcile output	165-74			
Delimiters	Front	**:**	Rear	**(**

(b)

Journal	Genome Research			
Labeled text data (Title)	<H2> <FONT COLOR=A70716 SIZE=-1>LETTER</FONT> <BR> Four-Hundred Million Years of Conserved Synteny of Human Xp and Xq Genes on Three Tetraodon Chromosomes </H2>			
Reconcile output	Four-hundred million years of conserved synteny of human xp and xq genes on three tetraodon chromosomes.			
Delimiters	Front	**<BR>**	Rear	**</H2>**

(c)

Fig. 2. Examples of labeled data, corresponding reconcile (validated) output, and delimiters for (a) Author, (b) Pagination, and (c) Title zones.

Learning process for extracting author delimiters

To extract author names from a given text zone, two types of journal specific delimiters are generally required: a front delimiter for detecting the start of author names and an embedded delimiter for separating each name. Once a new journal is given, such journal-specific author delimiters are defined through the following steps:

1) The cleanup process for extracting author names is performed using the predefined default delimiters.
2) The list of author names resulting from cleanup process is compared to that of the WebReconcile module, i.e., manual text verification. If these outputs are not identical, the learning process is activated.
3) Using a string matching technique, all author names included in the original text string generated from the

WebLabeling module can be localized. After removing unnecessary HTML tags and non-name text such as author titles (PhD, MD, etc.), superscripts, and subscripts, two types of delimiters are identified: front delimiter located at the beginning of the first author name, and embedded delimiter commonly found between one author name and another.

4) Finally, all extracted delimiters are saved in the WebMARS database for future cleanup processing.

Figure 2 (a) shows an example of an HTML formatted text string labeled as author zone, the list of author names validated by the WebReconcile module, and the front and embedded delimiters identified through the above learning process.

Learning process for extracting delimiters for pagination and other labeled zones

As for the author field, the cleanup process for the pagination zone relies on journal-specific delimiters. Thus, the learning procedure for this zone also focuses on extracting such delimiters.

1) Detect exact pagination position in the original text string using the validated pagination data through a string matching operation. Note that the formats of pagination data generated from the WebReconcile module is often different from those in the original string due to the reformatting needed in compliance with MEDLINE conventions. Thus a preprocessing step to restore the original format of this data needs to be performed before string matching.
2) Find the front delimiters commonly used to indicate the start of pagination such as “pp.”, “p.”, and “Pages”.
3) If such delimiters are not available, specific HTML tags or non-alphanumeric characters such as punctuation marks may be chosen as front delimiters. Moreover, contextual information such as format and position, as well as structure of the text string in a pagination zone will be utilized to improve detection accuracy.
4) Find a punctuation mark or other non-alphanumeric character located at the end of pagination, and define it as the rear delimiter.

A similar learning strategy is applied to other labeled zones such as title and affiliation to extract their journal-specific delimiters. Figure 2 (b) and (c) show examples of HTML-formatted text strings of pagination and title zones, and corresponding final cleanup results validated manually by the text verification operator. The front and rear delimiters are also given at the bottom of each figure.

Learning process for extracting non-journal-specific rules

Generally, bibliographic fields such as grant number, databank, zip code, and e-mail have characteristic formats that are not unique to journals or journal publishers. Therefore, the cleanup processing to extract such information from a given text is based on non-journal-specific rules created to identify their particular format. For example, an NIH grant number usually consists of 2 letters of a code name representing the administering organization (institute) followed by 5 or 6 numerals of a serial number. By recognizing the combination of an organizational code and serial number, we can detect a grant number in a given text. Furthermore, new formats for grant numbers can be recognized, should there be a newly created organization code or different number of digits in the serial number.

3. EXPERIMENTAL RESULTS

In experiments, we applied the proposed automated LabelCleanUp module to the cleanup task for extracting title, author names, pagination, and grant number from labeled text zones, and compared its detection accuracy to that of the current LabelCleanUp module with default rules. The experimental results are shown in Table 1. In the case of the title zone, the automated cleanup module achieves 98.98% detection accuracy for the test data set collected from 295 articles from 10 different journal issues. This performance is about 13% better than that of the current LabelCleanUp module based on default title delimiters. Similarly, cleanup results for other labeled zones show that the detection accuracy can be significantly improved by employing the automated LabelCleanUp module.

Through error analysis, we found that most cleanup errors caused by the proposed module were generated during the cleanup processing for the first article of each journal issue for which the default delimiter is applied. However, such errors are found to be reduced effectively by using the new rules obtained after performing learning process based on the cleanup output and corresponding reconcile output of the first journal article. Thus we conclude that journal-specific delimiters and new formats relevant for extracting citation information of interest such as title, author, pagination, and NIH grant number can be identified automatically through the aforementioned learning procedure, thereby improving cleanup performance.

Table 1. Experimental results of cleanup processing for extracting title, author names, pagination, and grant number

Labeled zone	Issues	Articles	Current cleanup (%)	Automated cleanup (%)
Title	15	295	85.42	98.98
Author	15	312	68.59	96.47
Pagination	10	301	60.80	98.01
Grant number	16	94	75.53	97.87

4. CONCLUSIONS

In this paper, we have described an automated cleanup module, one of the key components of *Web-based Medical Article Records System* (WebMARS), a system to extract and reformat bibliographic information data embedded in HTML-formatted online journal article using a rule-based approach. The proposed automated LabelCleanUp module employs a learning scheme that compares the outputs of the WebReconcile, LabelCleanUp, and WebLabeling modules to identify relevant cleanup rules automatically, thereby minimizing the manual effort for rule setting and improving cleanup performance.

Through a series of experiments on HTML-formatted text data obtained from real online biomedical journal articles, we found that the proposed automated LabelCleanUp module delivers better performance in terms of its detection accuracy than the current LabelCleanUp module with default rules. These experiments demonstrate that journal-specific delimiters and new formats suitable for extracting citation information of interest such as title, author, pagination, and NIH grant number can be identified accurately through the learning procedure.

Future work is planned to include multiple delimiters to handle certain changes of text format within a given journal that are not covered by the current single delimiter-based approach.

5. REFERENCES

[1] S. Mukherjea, L. V. Subramanium, G. Chanda, S. Sankararaman, R. Kothari, V. Batra, D. Bharadwaj, and B. Srivastava, “Enhancing a Biomedical Information Extraction System with Dictionary Mining and Context Disambiguation,**”** **IBM Journal of Research and Development** , vol. 48, no. 5/6, pp. 693-701, 2004.

[2] J.C. French and A.L. Powell, “Using Clustering for Creating Authority Files,” **Journal of the American Society for Information Science**, vol. 51, no. 8, pp. 774-786, 2000.

[3] D.M. Bikel, R. Schwartz, and R.M. Weischedel, “An Algorithm That Learns What’s in a Name,” **Machine Learning**, Vol. 34, No.1-3, pp. 211-231, 1997.

[4] D.X. Le and G.R. Thoma, “Automated Article Links Identification for Web-based Online Medical Journals,” **Proc. 8th World Multiconference on Systemics, Cybernetics, and Informatics**, vol. 5, pp. 462-466, 2004.

[5] J. Kim, D.X. Le, and G.R. Thoma, “Automated Labeling of Bibliographic Data Extracted from Biomedical Online Journals,” **Proc. SPIE, Document Recognition and Retrieval**, vol. 5010, pp. 47-56, 2003.

[6] S. Mao S, J. Kim, and G.R. Thoma, “A Dynamic Feature Generation System for Automated Metadata Extraction in Preservation of Digital Materials,” **Proc. 1st Int’l Workshop on Document Image Analysis for Libraries**, pp. 225-232, Palo Alto, CA, January 2004.