Automated Labeling of Bibliographic Data Extracted From Biomedical Online Journals

Jongwoo Kim*, Daniel X. Le, and George R. Thoma
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland 20894

ABSTRACT

A prototype system has been designed to automate the extraction of bibliographic data (e.g., article title, authors, abstract, affiliation and others) from online biomedical journals to populate the National Library of Medicine's MEDLINE® database. This paper describes a key module in this system: the labeling module that employs statistics and fuzzy rule-based algorithms to identify segmented zones in an article's HTML pages as specific bibliographic data. Results from experiments conducted with 1,149 medical articles from forty-seven journal issues are presented.

Keywords: HTML, Online journals, Labeling, Fuzzy Rule-Based Algorithm, Statistics, WebMARS, NLM

1. INTRODUCTION

The Lister Hill National Center for Biomedical Communications, an R&D division of the National Library of Medicine developed the Medical Article Records System (MARS) to automatically extract bibliographic data from paper-based biomedical journals. MARS consists of subsystems for scanning, OCR, page segmentation (zoning), labeling[1-2], and lexical analysis combined with biomedical lexicons. As a production system, MARS is used to routinely process hundreds of journal articles a day. Since biomedical journals are increasingly appearing in electronic form available from publishers' websites, we see an opportunity to offset some of the disadvantages of scanning and OCR necessary for processing paper journals by going directly to the online journals. We have developed a prototype system called WebMARS[3] for this purpose. A key subsystem in WebMARS is the WebLabeling module (WLM) that uses statistics and fuzzy rule-based technology[4-8] to identify segmented regions (zones) as useful bibliographic data such as title, author names, affiliation, abstract, rubric, email, zip code, pagination, grant number, databank, and corporate author.

Section 2 provides a brief system overview, Section 3 presents the basic structure of the WebLabeling module, Section 4 describes features used in this module, and Section 5 describes a fuzzy rule-based algorithm. Experimental results and conclusion are in Sections 6 and 7.

2. OVERVIEW OF WEBMARS

In WebMARS, a module named WebPageCollection downloads articles of recent journal issues from websites, and saves the articles in the WebMARS database. The WebLabeling module (WLM) then divides HTML-formatted articles into several zones, and labels each zone as one of ten important items of bibliographic data: article title, author names, affiliation, abstract, rubric, e-mail, zip code, pagination, databank accession number, and grant number. The LabelCleanUp module removes unnecessary tag information, and reformats each zone according to conventions followed in MEDLINE fields. The WebReconcile module allows an operator to check the results as a final verification step, and finally the verified data is uploaded by the Upload module to the MEDLINE® database.

*kimj@mail.nlm.nih.gov; phone 1 301 435-3227; fax 1 301 402-0341; archive.nlm.nih.gov

3. BASIC STRUCTURE OF THE WEBLABELING MODULE

The WLM consists of three sub modules: zoning, labeling, and updating. The zoning module divides HTML-formatted articles of a journal issue into several zones using tag information. The labeling module labels the zones as one of the important labels such as title, author, affiliation, abstract, etc. The updating module collects statistical and non-statistical information from the labeling results, and updates information related to the journal and important label zones in the WebMARS database. In this paper, we will focus on the labeling module.

4. FEATURES USED IN THE LABELING MODULE

Most features and rules derived for labeling are based on an analysis of the HTML-formatted articles (files which have extension ".html"). Both geometric and non-geometric features are used. A geometric feature is based on the zone order in an article. For example, the title zone is followed by the author, affiliation, abstract and other zones. Non-geometric features are derived from zone contents, tag information, and font characteristics. Since the label of a zone is frequently recognized by the words it contains, word matching is an important function in the automated labeling module. For example, a zone has a higher probability of being labeled as "affiliation" when it has a high proportion of words representing 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. Seventeen word-list tables are collected as shown in Table 1, and the Ternary Search Tree algorithm[9] is used as the search engine for the word matching.

The Title, Author, Affiliation, and Abstract tables are collected from articles that were processed by MARS from May 1997 to January 2001, and the words with top 10% of the highest frequency in each table are used in this experiment. TitleSpecial, KeyAffiliation, and AbstractSpecial tables are subsets of Title, Affiliation, and Abstract tables, respectively, and the words in the tables are collect manually.

The ZipCode table has 164 pairs of state names with the first two digits of U.S. zip codes. Examples are: "(MD, Maryland, 20)" and "(MO, Missouri, 65)". The GrantNumber table has pairs of U.S. government institute names and their Grant codes, such as "(NIH, AA)" and "(USPHS, DK)".

Sixty-three features are extracted from each zone in the HTML-formatted articles for the labeling module, some of which are shown in Table 2.

The Rubric table contains words describing the type of document: a Review, an Editorial, Letters to the Editor, and so forth. Such words frequently appear just above the title zone. This table is used to estimate Number_Rubric in a zone in the sixth row in Table 2. Title in the third row in Table 1 is also used to estimate Number_Title in a zone in the seventh row in Table 2.

5. FUZZY RULE-BASED LABELING ALGORITHM

5.1 Membership function generation

Statistics are used to make membership functions for a fuzzy rule-based algorithm. Membership functions for title, author, affiliation, abstract, number of words, and order of the label zone are estimated for each important label. Eight different journal issues consisting of 214 articles are selected to generate membership functions. In the case of title, all title zones are collected from 214 articles, and histograms of Ratio_Title, Ratio_SumAuthor, Ratio_Affiliation, Ratio_Abstract, Number_Word, and Order_Zone for title are estimated. For other labels, the same methods are used to estimate six histograms for each label.

Since 214 articles are insufficient to produce smooth histogram distribution, a smoothing operator (averaging operator with size eleven) is used to make the histograms smooth, and the smoothed histograms are normalized so that the maximum value is 100.

Figure 1 shows the procedure of generating the membership function of Ratio_Affiliation for affiliation label. Figure 1(a) shows the histogram of data collected from the 214 articles. The horizontal axis indicates Ratio_Affiliation, and the vertical axis indicates number of zones. Figure 1(b) shows the normalized histogram after smoothing the histogram in Figure 1(a) fifty times with the averaging operator. Figure 1(c) shows membership function of Ratio_Affiliation for affiliation label. As shown in Figure 1(c), when a zone has Ratio_Affiliation more than 50%, membership value of the zone is 100.

Figure 2 shows all six-membership functions for the affiliation label. Figures 2(a)-(f) are membership functions of Ratio_Title, Ratio_SumAuthor, Ratio_Affiliation, Ratio_Abstract, Number_Words, and Order_Zone for affiliation label. When a zone has following values; Ratio_Title ≈ 10, Ratio_Author ≈ 0, Ratio_Affiliation ≥ 50, Ratio_Abstract ≈ 15, Number_Word ≈ 30, and Order_Zone ≈ 20, membership values of each function have higher values, and the zone has higher probability to affiliation label as shown in Figure 2.

Table 3 shows membership functions used in the labeling module. Each label has six membership functions. In the case of affiliation label, the membership function of Ratio_Title for affiliation label is described as $MF_{af,ti}$, and the membership function of Ratio_SumAuthor for affiliation label is described as $MF_{af,au}$ as shown in the fourth row in the table.

To estimate probability of a zone to affiliation label, six features (Ratio_Title, Ratio_SumAuthor, Ratio_Affiliation, Ratio_Abstract, Number_Word, and Order_Zone) are estimated from the zone to have membership values from six membership functions, and a defuzzification method is used to estimate the probability from the six membership values. The same method is used to estimate probabilities of title, author, and abstract labels.

5.2 Fuzzy rules

Four fuzzy rules are used to estimate the probability of a zone bearing a label such as title, author, affiliation, and abstract labels in a zone.

Rule Title: If {Ratio_Title is $MF_{ti,ti}$ and Ratio_SumAuthor is $MF_{ti,au}$ and Ratio_Affiliation is $MF_{ti,af}$ and Ratio_Abstract is $MF_{ti,ab}$ and Number_Word is $MF_{ti,wo}$} and Order_Zone is $MF_{ti,or}$, the zone belongs to Title label zone.

Rule Author: If {Ratio_Title is $MF_{au,ti}$ and Ratio_SumAuthor is $MF_{au,au}$ and Ratio_Affiliation is $MF_{au,af}$ and Ratio_Abstract is $MF_{au,ab}$ and Number_Word is $MF_{au,wo}$} and Order_Zone is $MF_{au,or}$, the zone belongs to Author label zone.

Rule Affiliation: If {Ratio_Title is $MF_{af,ti}$ and Ratio_SumAuthor is $MF_{af,au}$ and Ratio_Affiliation is $MF_{af,af}$ and Ratio_Abstract is $MF_{af,ab}$ and Number_Word is $MF_{af,wo}$} and Order_Zone is $MF_{af,or}$, the zone belongs to Affiliation label zone.

Rule Abstract: If {Ratio_Title is $MF_{ab,ti}$ and Ratio_SumAuthor is $MF_{ab,au}$ and Ratio_Affiliation is $MF_{ab,af}$ and Ratio_Abstract is $MF_{ab,ab}$ and Number_Word is $MF_{ab,wo}$} and Order_Zone is $MF_{ab,or}$, the zone belongs to Abstract label zone.

5.3 Defuzzification

There are several aggregation operators for "and": Max, Multiply, etc. Weighted sum and multiply operators are used in this experiment. Since $MF_{ti,ti}$ is more important than other membership functions for Rule Title, and $MF_{au,au}$ is more important than other membership functions for Rule Author, different weights are given to each membership function. I.e., the following weights are used. When we assume that feature values of a zone t are expressed as i=Ratio_Title(t), j=Ratio_SumAuthor(t), k=Ratio_Affiliation(t), l=Ratio_Abstract(t), m=Number_Word(t), and n=Zone_Order(t), the probabilities of the zone t to the four important labels are estimated as follows.

Rule Title: $\mathbf{P_{Title}(t)} = \{w_1 \times MF_{ti,ti}(i) + w_2 \times MF_{ti,au}(j) + w_2 \times MF_{ti,af}(k) + w_2 \times MF_{ti,ab}(l) + w_2 \times MF_{ti,wo}(m)\} \times MF_{ti,or}(n)$.

Rule Author: $\mathbf{P_{Author}(t)} = \{w_2 \times MF_{au,ti}(i) + w_1 \times MF_{au,au}(j) + w_2 \times MF_{au,af}(k) + w_2 \times MF_{au,ab}(l) + w_2 \times MF_{au,wo}(m)\} \times MF_{au,or}(n)$.

Rule Affiliation: $\mathbf{P_{Affiliation}(t)} = \{w_2 \times MF_{af,ti}(i) + w_2 \times MF_{af,au}(j) + w_1 \times MF_{af,af}(k) + w_2 \times MF_{af,ab}(l) + w_2 \times MF_{af,wo}(m)\} \times MF_{af,or}(n)$.

Rule Abstract: $\mathbf{P_{Abstract}(t)} = \{w_2 \times MF_{ab,ti}(i) + w_2 \times MF_{ab,au}(j) + w_2 \times MF_{ab,af}(k) + w_1 \times MF_{ab,ab}(l) + w_2 \times MF_{ab,wo}(m)\} \times MF_{ab,or}(n)$.

w_1=4/12 and w_2=2/12 are used in this experiment.

Four probabilities are estimated for each zone in an article, and zones with the highest probability for each label are selected for the label zone. I.e., a zone with the highest $P_{Title}(t)$ in an article is labeled as title, and a zone with the highest $P_{Author}(t)$ in an article is labeled as author. The same method is applied to label affiliation and abstract zones.

5.4 Other rules

Besides the four principal labels for bibliographic data (title, author, affiliation, and abstract), other fields are of interest. Different crisp rules are developed to label the rubric of the article, grant number, databank accession number, e-mail, zip code, pagination, and corporate authors. The rules for these labels also use the word matching functions to detect them. However, these rules are simpler than those for title, author, affiliation, and abstract.

In the case of rubric, when a zone contains more than one rubric word, is in a common location where rubrics are normally found (Order_Zone), and satisfies other conditions for rubric label, the rule labels the zone as rubric. In the case of zip code, when a zone has five-digit numbers (the first two-digit number should be relevant to the state name) followed by relevant state name (two-digit state name or full state name), and satisfies other conditions for zip code label, the rule zip code labels the zone as zip code. Similar rules are applied for other labels. Table 4 shows abbreviated rules for other important labels.

5.5 Modification of probability values

Since every probability function is based on word features, it needs modification of probabilities using other important features. In the case of the title zone, font sizes of most title zones are larger than in other zones, and furthermore, title zones usually have tags such as "<H2>" or "<H3>". Therefore, probability of title ($\mathbf{P_{Title}(t)}$) should be increased when a zone has the tags. In the case of the affiliation zone, probability of a zone being an affiliation ($\mathbf{P_{Affiliation}(t)}$) is also increased when the zone has many KeyAffiliation words such city and country names. Therefore, ratio of KeyAffilation and Affiliation words is used as weights for the adjustment of the probability. It is really difficult to distinguish between abstract and other zones when there are no "Abstract" and "Keyword" in the article. Therefore, in the case of the abstract, adjustment is focusing on relation between abstract and other important label zones. I.e., Probability of abstract is decreased when the zone has tags "<H2>" or "<H3>". The probability of abstract is increased when the zone has words in the AbstractSpecial table.

6. EXPERIMENTAL RESULTS

Figure 3 shows the visual version of the labeling module. The zoning and labeling results are checked visually using this module. Pagination, title, author, affiliation, zip code, and abstract are labeled correctly as shown in the figure. Figure 4 shows an example of the labeling procedure. The figures are obtained from the visual labeling module. Figure 4(a) is an input journal article with HTML-format, and Figure 4(b) is the zoning result. Different colors are assigned to different zones. Figure 4(c) is the labeling result. Rubric, title, author, affiliation, and abstract zones are labeled correctly. Figure 5 shows another example of labeling result. Figure 5(a) is the end part of an input journal article, and Figure 5(b) is the zoning result. Figure 5(c) shows the labeling result. Grant Number, Email, and zip code are labeled correctly.

1,149 articles are selected from 47 journal issues picked from forty different journals, and Table 3 shows the experimental results. There are seven errors in title, nine errors in author, four errors in affiliation, one error in abstract,

and three errors in pagination zones. Figure 6 shows some examples of labeling errors. In Figure 6(a), the title zone "Introduction" has the biggest font in the article. However, since there is only one word in the zone, and the word is usually used as a section name in journal article, another zone was labeled as title. In Figure 6(b), there is an author zone between title and affiliation zones, and the zone has author names with affiliation information. The number of words related to author names is less than number of words related to affiliation. Therefore, a zone with more author-related words is labeled as author. In Figure 6(c), the affiliation zone is split into three zones, and the first affiliation zone "The Jackson Laboratory" has fewer affiliation-related words than the second and third affiliation zones. Therefore, the labeling module labels the second and third zones as affiliation.

Twenty-four errors are found in the 1,149 articles, giving an error rate of 2.09 %. We conclude that the accuracy of our labeling module is 97.91%.

7. SUMMARY

This paper describes a fuzzy rule-based algorithm to label the zones (bibliographic information) in HTML-format articles, which are downloaded from websites of medical journals. The proposed labeling module shows promise for analyzing the increasing number of journals that will be published electronically as HTML files. The labeling module employs both geometric and non-geometric features as the basis for the set of rules. Experiments conducted for 1,149 journal articles show 97.91% labeling accuracy. Though the test data set is not large, it demonstrates the potential for large-scale implementation of labeling bibliographic text in online journals.

Since the statistics and memberships obtained from 214 articles are not enough to process several journals, the statistics should be updated from more articles. Individual journal statistics such as location (Order_Zone) and font size of each label will be collected and used to solve some exceptional labeling problems as shown in Figure 5, and to improve the computation time taken by the labeling module. The weights for defuzzification are selected heuristically, and work properly in this experiment. However, neural network algorithms such as Back Propagation will be adapted to estimate more robust weights for the labeling module.

REFERENCES

1. D. Le, J. Kim, G. Pearson, and G.R. Thoma, "Automated Labeling of Zones from Scanned Documents," *Proceedings 1999 Symposium on Document Image Understanding Technology*, pp.219-226, 1999.
2. J. Kim, D. X. Le, and G.R. Thoma, "Automated Labeling in Document Images," *IS&T/SPIE's 13th Annual Symposium on Electronic Imaging 2001*, pp. 111-122, San Jose, California, January 2001.
3. D.X. Le, L.Q. Tran, et. al., "Automated Medical Citation Records Creation for Web-Based On-Line Journals," *14th IEEE Symposium on Computer-Based Medical Systems*, pp. 315-320, Bethesda, Maryland, July 2001.
4. L.A. Zadeh, "Outline of a new approach to the analysis of complex systems," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 1., pp. 28-44, 1973.
5. T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transaction on Systems, Man, and Cybernetics*, Vol. 15, No. 1, pp. 116-132, 1985.
6. T. Yasukawa and M. Sugeno, "A fuzzy-logic-based approach to qualitative modeling," *IEEE Transactions on Fuzzy Systems*, Vol 1., No. 1, pp. 7-31, February, 1993.
7. J. Keller and R. Krishnapuram, and F. Rhee, "Evidence aggregation networks for fuzzy logic inference," *IEEE Transactions on Neural Networks*, Vol. 3, No. 5, pp. 761-769, September 1992.
8. B. Kosko, *Neural Networks and Fuzzy Systems*, Englewood Cliffs, Prentice Hall, NJ, 1992.
9. Bentley and B. Sedgewick, "Ternary Search Trees," *Dr. Dobb's Journal*, pp. 20-25, April 1998.

Table 1. Word List Tables used in the Labeling Module.

Table Name	Words in the Table
Rubric	Review, Orginal Article, etc.
Title	Words frequently used in the title zones.
TitleSpecial	Erratum, etc.
Author	Smith, John, Kim, etc.
Academic Degree	Ph.D., MD, RN, etc.
Affiliation	Words frequently used in the affiliation zones.
KeyAffiliation	University, Department, Institute, etc.
Abstract	Words frequently used in the abstract zones.
AbstractSpecial	Is, are, was, were, have, etc.
WordAbstract	Abstract, Summary, Background, etc.
WordStructuredAbstract	Aim, Result, Conclusion, etc.
Keyword	Keyword, Index word, etc.
Introduction	Introduction, Introduzione, etc.
Corporate	Society, Group, etc.
ZipCode	State names with their first two-digit zipcodes.
GrantNumber	Lists of institutes of US government (NIH, USPHS, etc) with their grant codes (AA, DK, etc.).
Databank	GenBank, EMBL, Embl, DDBJ, Ddbj

Table 2. Features used in the Labeling Module.

Zone Features	Variable Names
Geometric Feature:	
Zone order in sequence from the top	Order_Zone (A number)
Non-Geometric Features:	
Number of Characters and Words	Number_Character, Number_Word
Number of “Editorial”, “Article”, etc	Number_Rubric
Number of Title words such as “of”, “in”, “in”, “the”, “with”, etc.	Number_Title
Number of “M.D.”, “Ph.D.”, “RN”, etc.	Number_Degree
Number of Middle Name, “Jr”, “Sr”, “II”, etc.	Number_Middlename
Number of Author Names such as “van”, “de”, “lee”, “kim”, “wang”, etc	Number_Author
Number_Degree+Number_MiddleName+Number_Author	Number_SumAuthor
Number of Affiliation words such as “department”, “university”, “medicine”, etc.	Number_Affiliation
Number of City, State, Country, School, etc.	Number_KeyAffiliation
Number of Abstract words such as “the”, “of”, “in”, “and”, “to”, etc.	Number_Abstract
Number of “is”, “was”, “are”, “were”, “be”, etc.	Number_KeyAbstract
Number of “Abstract”, “Summary”, etc.	Number_WordAbstract
Number of “Keywords”, “Index Words”, etc.	Number_Keyword
Number of Databank word; “GenBank”,”EMBL”, “DDBJ”, and “Ddbj”	Number_Databank
Ratio of Number_Title in a zone	Ratio_Title
Ratio of Number_SumAuthor in a zone	Ratio_SumAuthor
Ratio of Number_Affiliation in a zone	Ratio_Affiliation
Ratio of Number_KeyAffiliation in a zone	Ratio_KeyAffiliation
Ratio of Number_Abstract in a zone	Ratio_Abstract

Table 3. Membership functions used in the Labeling Module.

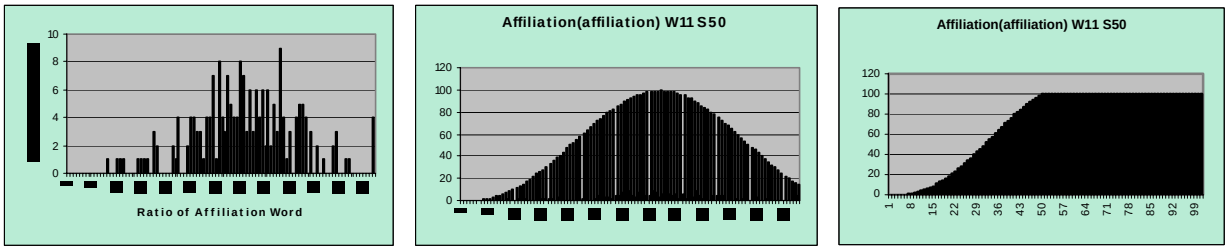
Label/Feature	Ratio_ Title	Ratio_ SumAuthor	Ratio_ Affiliation	Ratio_ Abstract	Number_ Word	Order_ Zone
Title	$MF_{ti,ti}$	$MF_{ti,au}$	$MF_{ti,af}$	$MF_{ti,ab}$	$MF_{ti,wo}$	$MF_{ti,or}$
Author	$MF_{au,ti}$	$MF_{au,au}$	$MF_{au,af}$	$MF_{au,ab}$	$MF_{au,wo}$	$MF_{au,or}$
Affiliation	$MF_{af,ti}$	$MF_{af,au}$	$MF_{af,af}$	$MF_{af,ab}$	$MF_{af,wo}$	$MF_{af,or}$
Abstract	$MF_{ab,ti}$	$MF_{ab,au}$	$MF_{ab,af}$	$MF_{ab,ab}$	$MF_{ab,wo}$	$MF_{ab,or}$

Table 4. Rules for the Labeling Module.

Label	Rules
Rubric	Number_Rubric > 0 and Number_Word <= 4
Corporate	Number_Corporate > 0 and Order_Zone is similar to Author
Grant Number	Number_Grant > 0 and there is a word such as “supported”
Databank	Number_Databank > 0 and there is a word such as “submitted” or “deposited”
E-mail	Number_Email > 0 and there is a word such as “E-mail”, “Email”, etc.
Zip Code	Number_ZipCode > 0 and Number_Affiliation > 0
Pagination	Number_JournalName > 0 and Number_Year or Month > 0 and Number_Volume or _Issue > 0

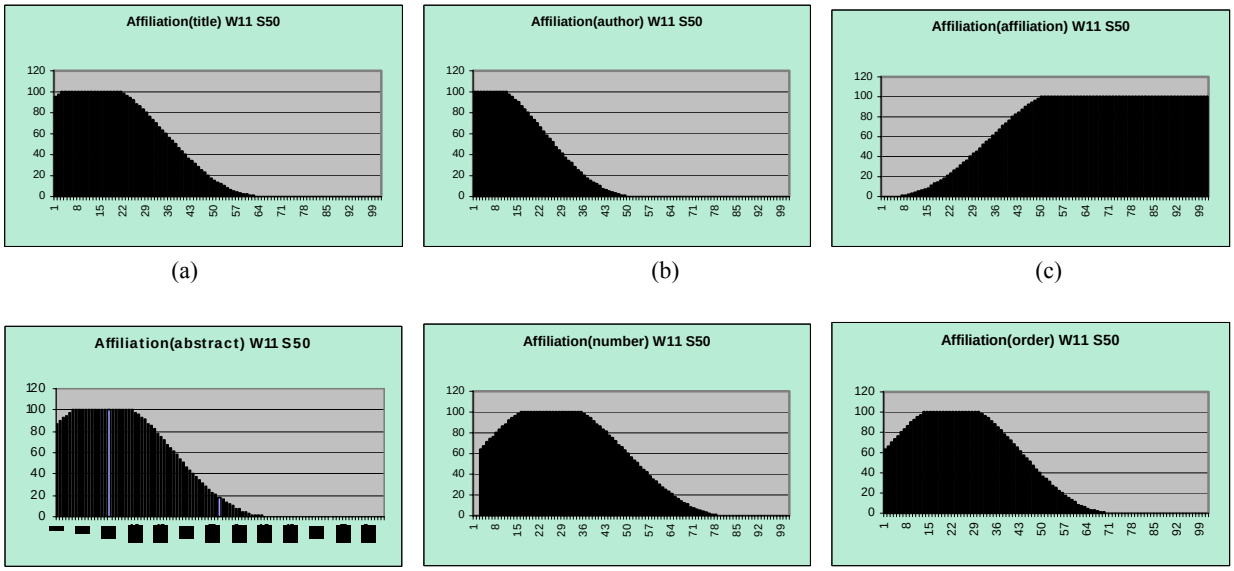
Table 5. Test Results of the Labeling Module.

Label	Number	Error	Error (%) of each label
Rubric	434	0	0
Title	1149	7	0.6
Author	1138	9	0.8
Affiliation	1080	4	0.4
Abstract	996	1	0.1
Grant	351	0	0
Databank	13	0	0
E-mail	915	0	0
Zip Code	586	0	0
Pagination	1142	3	0.3
Corporation	6	0	0
Total Articles	**1149**		
Total Error		**24**	
Total Error (%)			**2.09**



(a) (b) (c)

Figure 1. Procedure for generating a membership function of Ratio_Affiliation for affiliation label. (a) Histogram of Ratio_Affiliation in affiliation label. (b) Normalized smoothing result of the histogram (a). (c) Membership function of Ratio_Affiliation for affiliation label.



(a) (b) (c)

(d) (e) (e)

Figure 2. Membership functions of affiliation label. Horizontal axis indicates feature values, and vertical axis indicates membership values. Membership functions of (a) Ratio_Title, (b) Ratio_SumAuthor, (c) Ratio_Affiliation, (d) Ratio_Abstract, (e) Number_Words, and (f) Order_Zone.

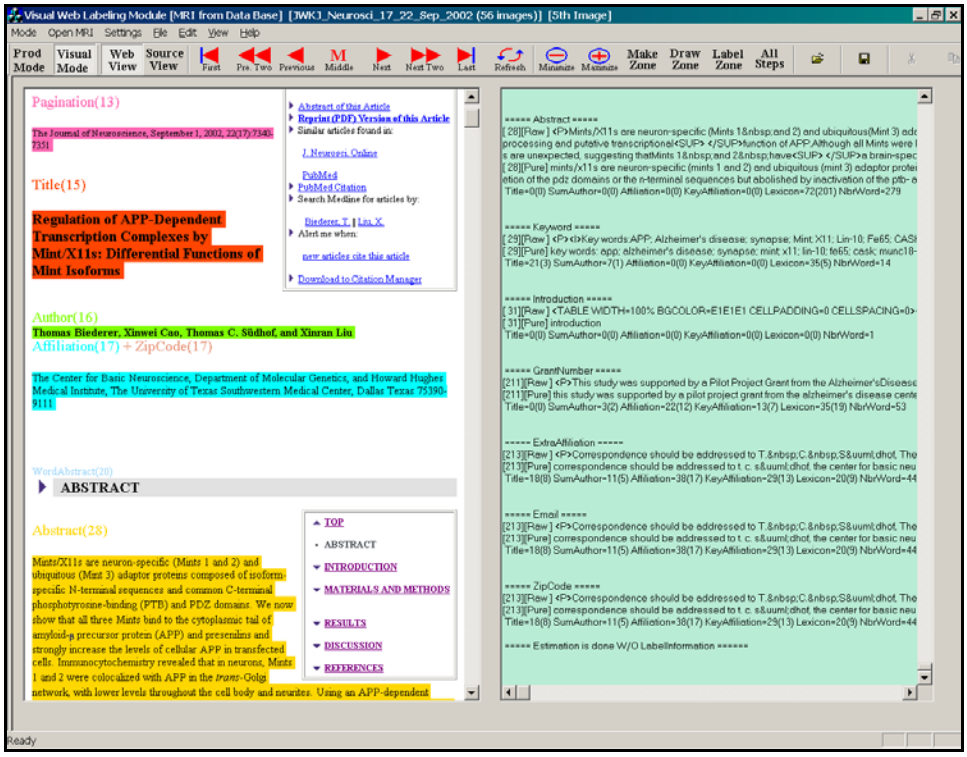
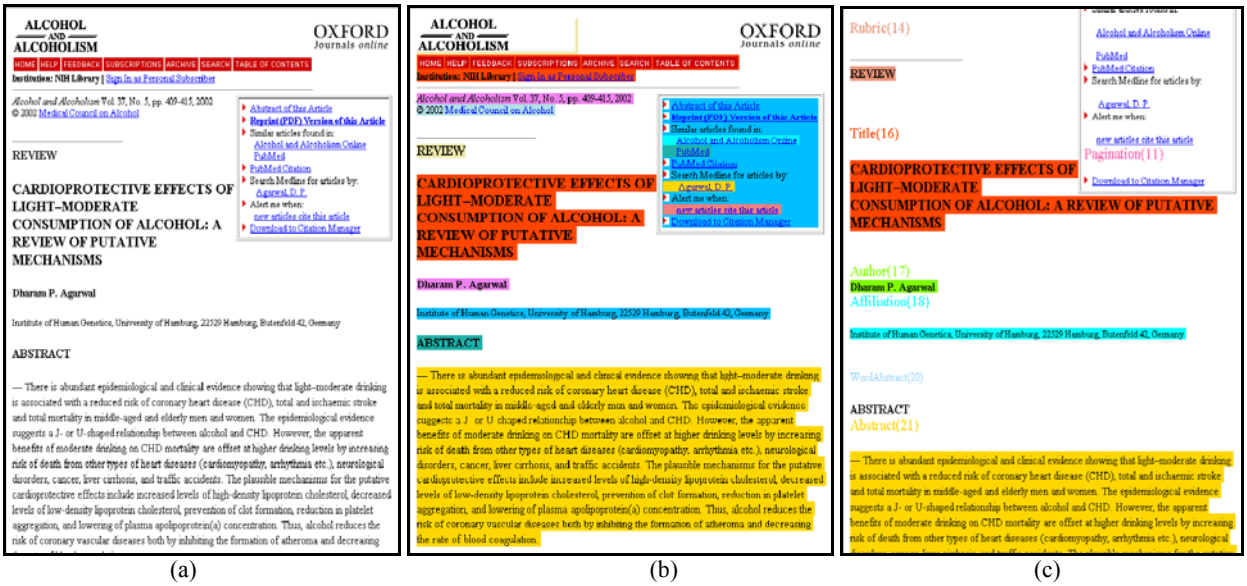
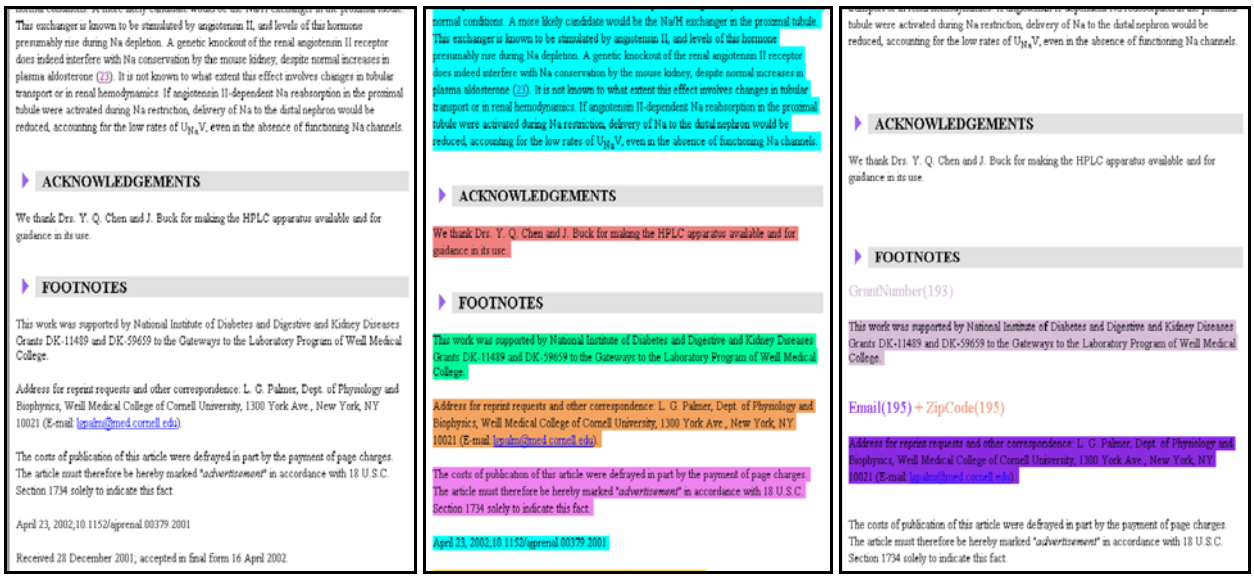


Figure 3. The visual version of the labeling module.



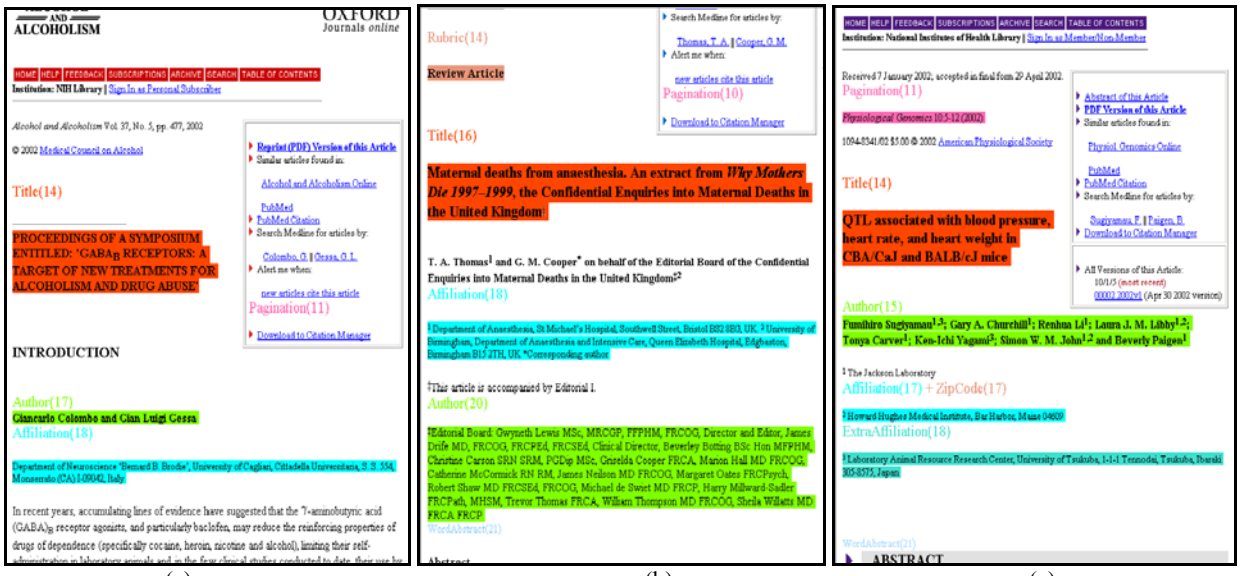
(a) (b) (c)

Figure 4. Example of the Labeling Result. (a) An HTML-format article, (b) Zoning Result, and (c) Labeling Result.



(a) (b) (c)

Figure 5. Example of the Labeling Result. (a) An HTML-format article, (b) Zoning Result, and (c) Labeling Result.



(a) (b) (c)

Figure 6. Examples of labeling errors in (a) title, (b) author and (c) affiliation zones.