Automated Labeling Of Biomedical Online Journal Articles

Jongwoo Kim, Daniel X. Le, George R. Thoma
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA

ABSTRACT

An automated labeling (AL) module has been developed to automate the extraction of bibliographic data (e.g., article title, authors, affiliation, abstract, and others) from online biomedical journals for the National Library of Medicine's MEDLINE® database. The AL module employs string matching, statistics, and fuzzy rule-based algorithms to identify segmented zones in an article's HTML pages as specific bibliographic data. Experiments conducted with 1,267 medical articles from 64 journal issues show about 97.71% accuracy.

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

1. INTRODUCTION

Document labeling is an active research area, and several techniques are based on the layout (geometric) structure and/or the logical structure of a document. Most of this research is based on processing the hard copy version of documents. [1-6]

Since growing numbers of journals are published electronically in HTML, PDF, and XML formats, document analysis of electronic journals has become important.

To meet this challenge, an R & D division of the National Library of Medicine (NLM) has developed an automated system named WebMARS for *Web-based Medical Article Records System* [7] to produce bibliographic records from online biomedical journals for the NLM's MEDLINE® database.

Automated labeling (AL) module is one of the subsystems in WebMARS, and it uses statistics and fuzzy rule-based technology [8-12] to identify segmented regions (zones) of an article as useful bibliographic data such as article title, author, affiliation, abstract, rubric, e-mail, zip code, pagination, databank accession number, grant number, corporate author, and three types of support. We focus on the AL module in the WebMARS in this paper.
Section 2 provides an overview of WebMARS, Section 3 describes features used in the AL module, Section 4 describes a fuzzy rule-based algorithm, and Section 5 describes other rules. Experimental results and a Summary are in Sections 6 and 7.

2. WEBMARS

The WebMARS is developed to process biomedical journal articles published in HTML and PDF formats. The system consists of five modules: WebPageCollection, WebLabeling, LabelCleanUp, WebReconcile, and Upload. The WebPageCollection module downloads journal articles from publisher websites, and saves the articles in the WebMARS database. It also converts PDF-format articles to HTML using BCL Magellan [13]. The WebLabeling module consists of automated zoning and labeling (AL) modules. The automated zoning module segments HTML-format articles into zones using HTML tag information, and the AL module labels each zone as one of the important bibliographic fields such as title, author, affiliation, abstract, etc. The LabelCleanUp module removes irrelevant HTML tags, and reformats the contents of each labeled zone in line with MEDLINE conventions. The WebReconcile module allows an operator to validate the results, and finally, the Upload module sends the results to the WebMARS database.

3. FEATURES USED IN THE AL MODULE

The AL module contains fuzzy rule-based algorithms for which features are needed. Most features and rules derived for labeling are based on an analysis of HTML-format articles. Both geometric and contextual features are used. Geometric features are based on the zone order in an article, while contextual ones are derived from zone contents and HTML tag information. Since the label of a zone is often characterized by the words in the zone, word matching is an important function in the module. For example, words indicating affiliation include university, school, and department, and are collected from the affiliation zones in previous articles. A list of these words is used to estimate features for this type of zone. Title, author, abstract word lists are also assembled by selecting the most frequent words in those zones from historic MEDLINE® records.

Other word lists are also made for rubric, e-mail, zip code, pagination, grant number, databank accession number, and corporate author zones.
The Rubric list contains words describing the type of document, and these words frequently appear just above the title zone. The list includes "Editorial", "Interview", "Letter", "News", "Classical Article", "Clinical Conference", and "Journal Article".

E-mail addresses typically end with ".com", ".net", ".org", etc. Also, two characters are often assigned to indicate country names of the e-mail addresses, such as ".us", ".kr", ".ca", and ".cn" standing for USA, Korea, Canada, and China, respectively. Of the 192 words in the word list of E-mail, 180 words are related to country names.

In the case of Zip Code, only U.S. zip codes are considered. This ZipCode list has 328 pairs of state names with the first two digits of U.S. zip codes. Examples are (MD, 20), (Maryland, 20), (MD, 21), (Maryland, 21), and (MD, 26), (Maryland, 26), meaning that zip codes in Maryland start with 20, 21, or 26.

When the research reported in an article is funded by one of the institutes at the National Institutes of Health (NIH) or the U.S. Public Health Service (PHS), there is a grant number and/or its

corresponding granting organization (institute) in the article. The grant number is composed of two alphabetic characters followed by a five-digit number. Each institution identifies its grants by a unique pair of alphabetic characters. Therefore, pairs of (an institution name, two alphabetical characters) are collected and saved in the GrantNumber list.

There are eleven databanks to register molecular sequences. The DatabankName list has contains words such as "GENBANK", "GenBank", "EMBL", "Embl", "DDBJ", "Ddbj", "SWISSPROT", "Swiss-Prot", etc.

The Support field has two word lists. One is a list of institutions belonging to the U.S. PHS and the other is a list of institutions belonging to the U.S. Federal Government (excluding U.S. PHS).

Sixteen word lists of the sort mentioned above are assembled and a Ternary Search Tree algorithm [14] is used as the search engine for word matching. Table 1 shows some of these word lists, and about 60 features are extracted from each zone in HTML-format articles using these word lists (as shown in Table 2).

4. FUZZY RULE-BASED ALGORITHM

4.1 Membership function generation

Statistics are used to make membership functions for a fuzzy rule-based algorithm for the AL module. 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 are estimated for Frequency_Title (frequency of title word), Frequency_Author (frequency of author word), Frequency_Affiliation (frequency of affiliation word), Frequency_Abstract (frequency of abstract word), Number_Words (number of words in the zone), and Order_Zone (order of the zone) for title. 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, an average 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 Frequency_Affiliation word for the affiliation label. Figure 1(a) shows the histogram of data collected from the 214 articles. The horizontal axis indicates Frequency_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 Frequency_Affiliation for affiliation label. As shown in Figure 1(c), when a zone has Frequency_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 Frequency_Title, Frequency_Author, Frequency_ Affiliation, Frequency_Abstract, Number_Words, and Order_Zone for affiliation label, respectively.

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 Frequency_Title for affiliation label is described as $MF_{af,ti}$, and the membership function of Frequency_Author 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 (Frequency_Title, Frequency_Author, Frequency_Affiliation, Frequency_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.

4.2 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 (Frequency_Title is $MF_{ti,ti}$ and
Frequency_Author is $MF_{ti,au}$ and
Frequency_Affiliation is $MF_{ti,af}$ and
Frequency_Abstract is $MF_{ti,ab}$ and
Number_Words is $MF_{ti,wo}$) and
Order_Zone is $MF_{ti,or}$,
the zone is **Title zone**.

Author Rule:
If (Frequency_Title is $MF_{au,ti}$ and
Frequency_Author is $MF_{au,au}$ and
Frequency_Affiliation is $MF_{au,af}$ and
Frequency_Abstract is $MF_{au,ab}$ and
Number_Words is $MF_{au,wo}$) and
Order_Zone is $MF_{au,or}$,
the zone is **Author zone**.

Affiliation Rule:
If (Frequency_Title is $MF_{af,ti}$ and
Frequency_Author is $MF_{af,au}$ and
Frequency_Affiliation is $MF_{af,af}$ and
Frequency_Abstract is $MF_{af,ab}$ and
Number_Words is $MF_{af,wo}$) and
Order_Zone is $MF_{af,or}$,
the zone is **Affiliation zone**.

Abstract Rule:
If (Frequency_Title is $MF_{ab,ti}$ and
Frequency_Author is $MF_{ab,au}$ and
Frequency_Affiliation is $MF_{ab,af}$ and
Frequency_Abstract is $MF_{ab,ab}$ and
Number_Words is $MF_{ab,wo}$) and
Order_Zone is $MF_{ab,or}$,
the zone is **Abstract zone**.

The subscripts "ti", "au", "af", "ab", "wo", and "or" indicate title, author, affiliation, abstract, number of words, and zone order, respectively. $MF_{ti,ti,}$, $MF_{ti,au}$, $MF_{ti,af}$, $MF_{ti,ab}$, $MF_{ti,wo}$, and $MF_{ti,or}$ are the memberships of title word, author word, affiliation word, abstract word, number of words in the zone, and order of the zone in title zone. Similar rules are made for other labels such as author, affiliation, and abstract.

As aggregation operators for the fuzzy rules, weighted sum and multiply operators are used in this experiment. Since $MF_{ti,ti}$ is more important than other membership functions for the Title Rule, and $MF_{au,au}$ is more important than other membership functions for the Author Rule, different weights are given to

each membership function. When we assume that feature values of a zone t are expressed as i=frequency of title word (t), j=frequency of author word (t), k=frequency of affiliation word (t), l=frequency of abstract word (t), m=number of words in the zone (t), and n=order of the zoen(t), the probabilities of the zone t to the four important labels are estimated as follows.

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

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

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

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

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. The same method is applied to label affiliation and abstract zones.

4.3 Modification of probability values

We need to modify of probabilities using other important features since every probability function is based on word 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 ($P_{Title}(t)$) should be increased when a zone has these tags. It is 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, an adjustment is made based on the relation between abstract and other important label zones. I.e., Probability of abstract is decreased when the zone has tags "<H2>" or "<H3>".

5. RULES FOR OTHER LABELS

5.1 Rule for Rubric

Six kinds of rubrics are considered in WebMARS: Editorial, Interview, Letter, News, Classical Article, Clinical Conference, and Journal Article. When a zone contains only one of the six rubric words and is in a location where rubrics are normally found (usually just above title zone), the zone is labeled as "rubric".

5.2 Rule for E-mail

The e-mail list includes words such as ".com", ".net", ".org", ".tv", ".gov", ".ac", and ".edu", ".kr", ".us", ".uk", ".ca", ".cn", etc. When a word has at least one ".", a "@", and a word in the E-mail list, the zone containing the word is labeled as "e-mail".

5.3 Rule for Zip Code

The US zip code list contains pairs of (a two-character state name, the first two digits of the zip code) and (a full state name, the first two digits of the zip code) such as ("MD", "20") and (Maryland", "20"). When a zone has the name of a state with a five-digit number, and the pair of the state name and the first two digits of the number are in the list, the zone is labeled as "zip code".

5.4 Rule for Grant Number

When the research reported in the article is funded by one of the institutes in the NIH or the U.S. PHS, there is a grant number and/or its corresponding grant organization (institute) in the article. When a zone has two alphabetic characters followed by a five-digit number, and the institution name corresponds to the two alphabetic characters in the GrantNumber list, the zone is labeled as "grant number".

5.5 Rule for Databank Name and Accession Numbers

The DatabankName list has words such as "CSD", "DDBJ", "EMBL", "GENBANK", etc. The formats of the accession number are either one alphabetic letter followed by a five-digit number, two alphabetic letters with a six digit number, or three alphabetic letters with a five digit number. When a zone has a sequence of such letters and numbers with one of names in the DatabankName list, it is labeled as "databank accession number".

5.6 Rule for Support

"Support" means the research is funded by the U.S. government, state or local governments, foreign governments, or private organizations. There are three kinds of support: (1) Support, U.S. Government, PHS, (2) Support, U.S. Government, Non-PHS, and (3) Support, Non-U.S. Government. All of the NIH and U.S. PHS grants belong to (1), and grants from other US government agencies belong to (2). Other cases belong to (3). We made two institutional lists for (1) and (2). We also made a lis for words such as "support", "grant", "fund", etc.
When a zone has one of the words in this support word list with institution names which belong to (1), the zone is labeled as "Support, U.S. Government, Public Health Service", and so on for other types of support.

6. EXPERIMENTAL RESULTS

Figure 3 shows examples of the labeling results for HTML-format articles. Figure 3(a) is an input journal article and Figure 3(b) is the labeling result. Rubric, title, author, affiliation, and abstract zones are labeled correctly. Figure 3(c) is the end part of an input journal article and Figure 3(d) shows the labeling result. Grant Number, E-mail, and zip code are labeled correctly.

Figure 4 shows examples of the labeling results for PDF-converted HTML-format articles. Figure 4(a) is an input journal article and Figure 4(b) is the labeling result. Title, author, affiliation, abstract, e-mail, and zip code zones are labeled correctly. Figure 4(c) is another input journal article and Figure 4(d) shows the labeling result. Title, author, affiliation, abstract, e-mail, zip code, and support zones are labeled correctly.

1,267 articles are selected from 64 journal issues picked from 57 different journals (40 journals for HTML-format articles and 17 journals for PDF-converted HTML-format articles), and Table 4 shows the experimental results. There are nine errors in title, ten errors in author, four errors in affiliation, one error in abstract, one error in grant, three errors in pagination, and one error in corporate author zones.

Twenty-nine errors are found in the 1,267 articles (8,464 label zones), giving an error rate of 2.29 %. We conclude that the accuracy of our labeling module is 97.71%.

7. SUMMARY

This paper describes a fuzzy rule-based algorithm to label the zones containing bibliographic information in HTML and PDF-converted HTML-format articles downloaded from medical journal Websites. Experiments conducted for 1,267 journal articles show 99.66% (based on label) and 97.71% (based on article) labeling accuracy. The test results show the potential for large-scale implementation of labeling bibliographic text in online journals.

As future work, journal specific information such as location (Order_Zone) and font size of each label should be collected to improve labeling accuracy and the computation speed of the labeling module.

8. REFERENCES

[1] F. Hones and J. Lichter, "Layout Extraction of Mixed Mode Documents," **Machine Vision and Applications** 7, 1994, pp. 237-246.

[2] Y. Tateisi and N. Itoh, "Using Stochastic Syntactic Analysis for Extracting a Logical Structure from a Document Image," **Proc. IEEE Int. Conf. Neural Networks**, Vol. 2,1994, pp. 391-394.

[3] T. Kanungo, S. Mao, "Stochastic Language Model for Style-Directed Physical Layout Analysis of Documents," **IEEE Transactions on Image Processing**, vol. 12, no. 5, May, 2003, pp. 583-596.

[4] J. Kim, D. Le, and G. Thoma, "Automated Labeling Document Images," **Proc. SPIE Electronic Imaging**, Vol. 4307, Document Recognition and Retrieval VIII, San Jose, CA January, 2001, pp.111-122.

[5] J. Kim, D. Le, and G. Thoma, "Automated labeling algorithms for biomedical document images**," Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics**, Vol. V, Orlando FL, July, 2003, pp.352-57.

[6] J. Kim, D. Le, and G. Thoma, "Automated labeling from Biomedical Journals Published in Foreign Languages**" Proc. 8th World Multiconference on Systemics, Cybernetics and Informatics**, Vol. V, Orlando FL, July, 2004, pp.444-449.

[7] 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**, Bethesda, Maryland, July, 2001, pp. 315-320.

[8] J. Kim, D. Le, and G. Thoma, **"**Automated labeling of bibliographic data extracted from biomedical online journals," **Proc. SPIE Electronic Imaging**, Vol. 5010, January, 2003, pp. 47-56.

[9] L.A. Zadeh, "Outline of a new approach to the analysis of complex systems," **IEEE Transactions on Systems, Man, and Cybernetics**, Vol. 1, 1973, pp. 28-44.

[10] T. Yasukawa and M. Sugeno, "A fuzzy-logic-based approach to qualitative modeling," **IEEE Transactions on Fuzzy Systems**, Vol 1., No. 1, February, 1993, pp. 7-31.

[11] J. Keller and R. Krishnapuram, and F. Rhee, "Evidence aggregation networks for fuzzy logic inference," **IEEE Transactions on Neural Networks**, Vol. 3, No. 5, September 1992, pp. 761-769.

[12] B. Kosko, **Neural Networks and Fuzzy Systems**, Englewood Cliffs, Prentice Hall, NJ, 1992.

[13]BCL Technologies, Inc., **BCL Magellan,** version 6.5, 2002.

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

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

Table Name	Words in the Table
Rubric	"Editorial", "Interview", "Letter", "News", "Classical Article", "Clinical Conference", and "Journal Article".
Title	Words frequently used in the title zones.
Author	Frequently used last and first names ("Smith", "John", "Kim", etc.)
Academic Degree	"Ph.D. ", "MS", "MD", "RN", etc.
Affiliation	Words frequently used in the affiliation zones. ("University", "Department", "Institute", etc.)
Abstract	Words frequently used in the abstract zones. ("the", "of", "in", "and", "to", etc.)
WordAbstract	"Abstract", "Summary", "Background", etc.
WordStructuredAbstract	"Aim", "Result", "Conclusion", etc.
Keyword	"Keyword", "Index word", etc.
Introduction	"Introduction", "Introduzione", etc.
Corporate	"Society", "Group", etc.
E-mail	".com", ".net", ".org", ".biz", ".info", ".tv", ".ws",".gov", ".ac", ".edu" ".kr", ".us", ".uk", etc.
ZipCode	State names with their first two-digit zipcodes. ("MD", "20"), ("Maryland", "20"), ("MD", "21"), ("Maryland", "21"), ("MD", "26"), ("Maryland", "26"), ("MI", "48"), ("Michigan", "48"), ("MI", "49"), ("Michigan", "49")
GrantNumber	Lists of institutes of US government (NIH, USPHS, etc) with their grant codes (AA, DK, etc.).
DatabankNamw	"GenBank", "Embl", "Ddbj", "Swiss-Prot", "CSD", "GDB", "HGML", "OMIN", "PDB", "PIR", "PRFSEQDB".

Table 2. Features used in the Labeling Module.

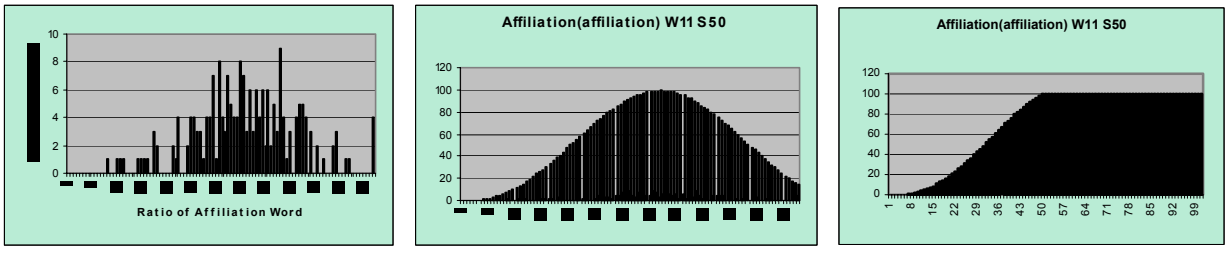
Variable Names	Features
	Geometric Feature:
Order_Zone	Zone order in sequence from the top
	Non-Geometric Features:
Number_Character	Number of Characters
Number_Word	Number of Words
Number_Rubric	Number of Rubric such as "Editorial", "Article", etc
Number_Title	Number of Title words such as "of", "in", "in", "the", "with", etc.
Number_Degree	Number of "M.D.", "Ph.D.", "RN", etc.
Number_Middlename	Number of Middle Name, "Jr", "Sr", "II", etc.
Number_Author	Number of Author Names such as "van", "de", "lee", "kim", "wang", etc
Number_SumAuthor	Number_Degree+Number_MiddleName+Number_Author
Number_Affiliation	Number of Affiliation words related to City, State, Country, School, etc.
Number_Abstract	Number of Abstract words such as "the", "of", "in", "and", "to", etc.
Number_Databank	Number of Databank word; "GenBank","EMBL", "DDBJ", and "Ddbj"
Frequency _Title	Frequency of Number_Title in a zone
Frequency _Author	Frequency of Number_SumAuthor in a zone
Frequency _Affiliation	Frequency of Number_Affiliation in a zone
Frequency _Abstract	Frequency of Number_Abstract in a zone

Table 3. Membership functions used in the Labeling Module.

Label/Feature	Frequency_ Title	Frequency_ Author	Frequency_ Affiliation	Frequency_ 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. Test Results of the Labeling Module.

Label	Number	Error	Error (%) of each label
Rubric	456	0	0
Title	1267	9	0.7
Author	1242	10	0.8
Affiliation	1186	4	0.4
Abstract	1102	1	0.1
Grant	360	1	0.3
Databank	14	0	0
E-mail	997	0	0
Zip Code	663	0	0
Pagination	1142	3	0.3
Corporate Author	9	1	11,1
Support	26	0	0
Total Label Zones	**8464**	**29**	**0.34**
Total Articles	**1267**	**29**	**2.29**



(a) (b) (c)

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

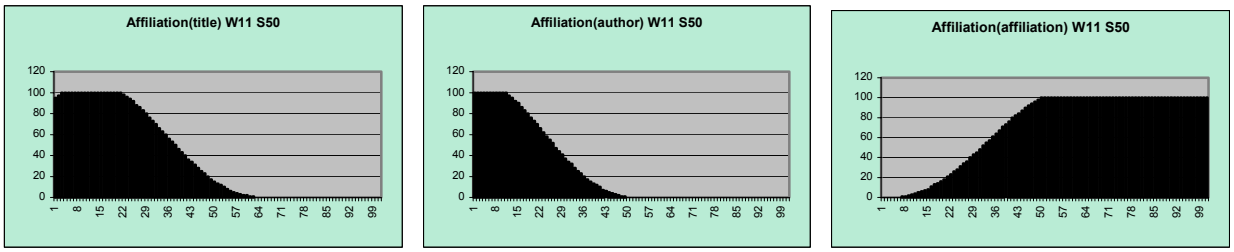


Figure 2. Membership functions of affiliation label. Horizontal axis indicates feature values, and vertical axis indicates membership values. Membership functions of (a) Frequency_Title, (b) Frequency_SumAuthor, (c) Frequency_Affiliation, (d) Frequency_Abstract, (e) Number_Words, and (f) Order_Zone.

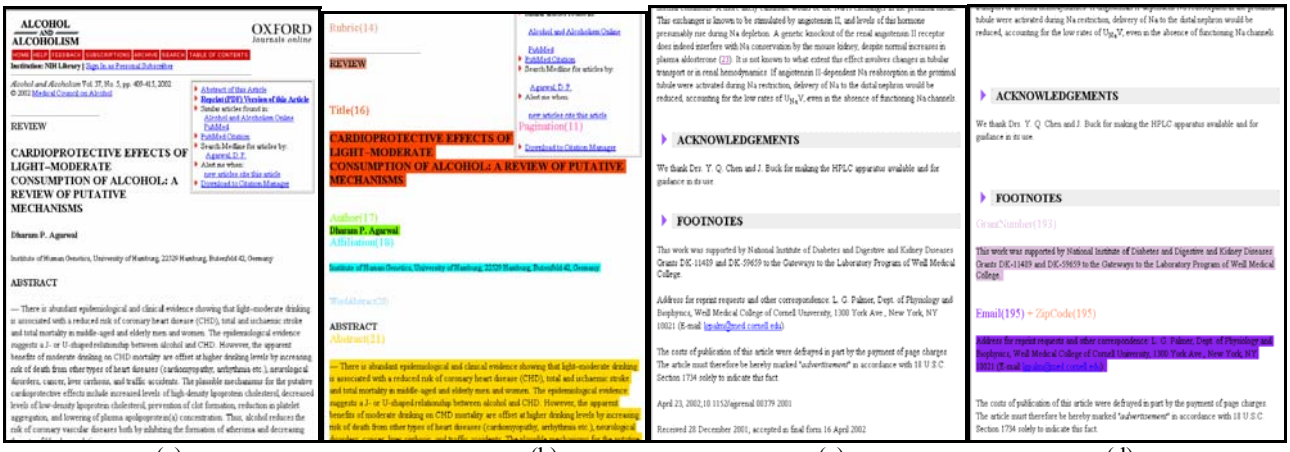


Figure 3. Example of the Labeling Result. (a) A HTML-format article, (b) Labeling Result, (c) A HTML-format article, (d) Labeling Result

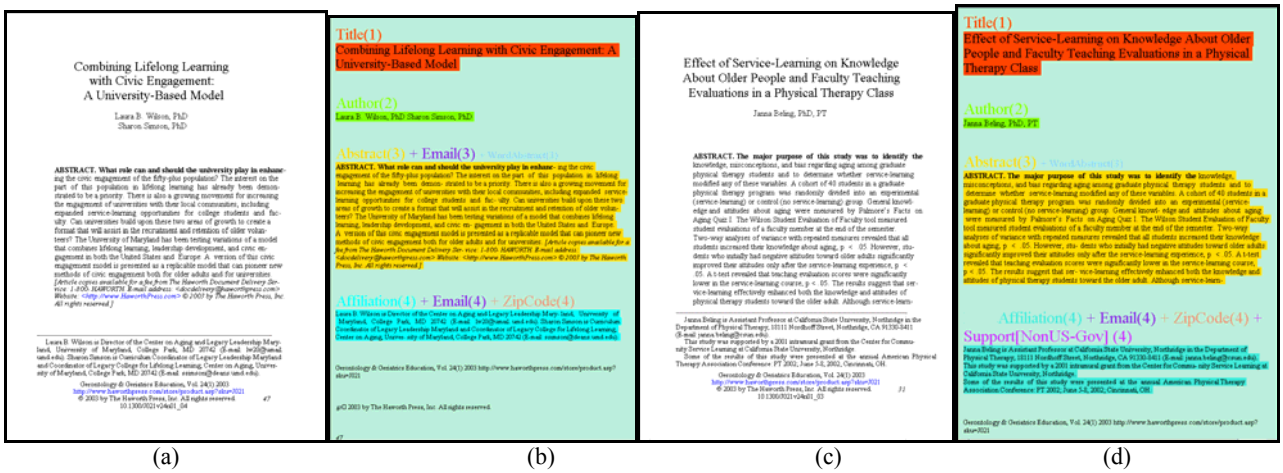


Figure 4. Example of the Labeling Result. (a) A PDF-converted HTML-format article, (b) Labeling Result, (c) A PDF-converted HTML-format article, (d) Labeling Result