Semi-automatic Metadata Extraction from Scientific Journal Article for Full-text XML Conversion

Sukyoung Kim, Yoonsung Cho, and Kihong Ahn

***Abstract*—By the increasing continuous academic researches, the volume of scientific articles has dramatically reached unpredictable level. To facilitate archive and publication, many scientific journals in Korea are actively adapting Open Access (OA) policy. In addition, it has more attractable than commercial printing of companies that freely provide the full text of article published in scholarly journal through web to user. Because of difficulty to convert automatically unstructured format such as pdf document into full-text, which is structured with accuracy, the most full text conversion works in scholarly journal publisher have been conducted with human interaction. To deal with the problem of low reliability in the automatic metadata extraction and help with minimum human interaction, we propose semi-automated metadata extraction method based on rule-based method and machine learning method. In this experiment, we verified the performance under 26 different journals in Open Access Korea Central (OAK Central). We only cover two part (elements of front and back) as part of an effort to convert full-text xml based on JATS v1.0. As a result, our proposed method reached F1 = 94.1% in front and F1 = 92.5% in back.**

I. Introduction

World Wide Web played a huge role in change from the center of traditional paper media into the center of electronic publishing on circulating structure of scholarly journal article. Electronic publishing has promoted research activity in various research fields internationally by facilitating the expeditious publication. Journals and conference proceeding that exist today publish about 2.5 million articles per years have peer-reviewed, propelled by this advantage [1]. But electronic journals are forged mainly with commercial printing of companies, half of whole scholarly journal article are produced by the top-5 commercial companies [3]. However, these publishers typically provide their contents with the client only who subscribe to their publishers. In order to have competitive power with commercial printing of companies and have stature, non-commercial scholarly publisher have to offer differentiated interoperability service. The kinds of differentiated interoperability service are Open Access (OA), JATS XML, Pubreader, DOI, CrossCheck, CrossMarkm FundRef, Open Researcher and Contributor ID (ORCID), QR code, mobile application and multimedia [3]. From among these, it is critical for differentiated core strategy to provide OA policy, which provides unrestricted online access to articles published in scholarly journals and JATS XML, which describe the content and metadata of journal articles on Web. Especially, providing full text of journal article on Web can maximize not only visibility and accessibility of the scientific journal article but also provides high quality service in view of Digital Library when the well-structured data such as XML is more tractable in managing collections of scholarly articles than unstructured data such as pdf. As Fig. 1, many of the internal and external scholarly journals have been in accordance with OA and full text XML based on JATS. However, because of difficulty to convert automatically unstructured format such as pdf document into full-text, which is structured with accuracy, the most full text conversion works in scholarly journal publisher still have been conducted with human interaction. To reduce time and coast and help with minimum human interaction, most of scholarly journals in Korea are faced with the necessity of automatic metadata extraction and conversion method with accuracy.

Automating this conversion work requires programmatic access to the typographical layout of elements page as well as to their logical/rhetorical function within article [6]. Therefore, we view this subject from Natural Language Processing (NLP) and Information Extraction techniques. Previous research in automatic metadata extraction from pdf format of scholarly journal articles is divided into two main categories. The first is the rule-based method. This method, if rule set is constructed about target domains in advance, gives best performance, but it does not perform well when changed target domain. On the other hand, the second, machine learning method is most popular in information extraction fields. It is not necessary for considering changed target domain when sufficient volume of training set is constructed. With these advantages, most of relevant researches are based on the machine learning methods.

To reduce time and cost by the human-curated XML conversion method, in this paper, we propose a semi-automatic metadata extraction module based on the rule-based method and the machine learning method. Our aim is to deal with the problem of low reliability in the automatic metadata extraction and help with minimum human interaction. From this point of view, we focus on

Sukyoung Kim is with Department of Computer Engineering, Hanbat National University, Daejeon, South Korea (e-mail: kimsk@hanbat.ac.kr).

YoonSung Cho is with Department of Computer Engineering, Hanbat National University, Daejeon, South Korea (e-mail: dibbul5456@gmail.com).

Kihong Ahn is with Department of Computer Engineering, Hanbat National University, Daejeon, South Korea (e-mail: khahn@hanbat.ac.kr).

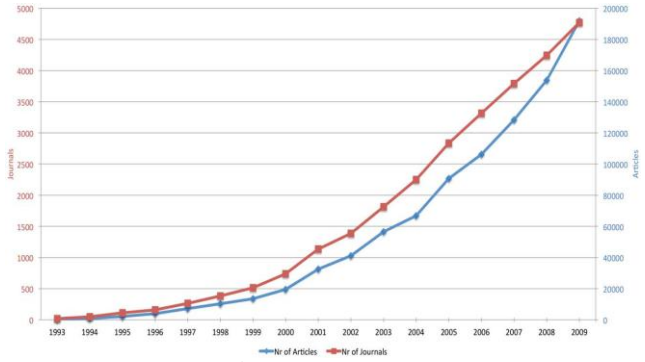

Fig. 1. The development of open access publishing
("The Development of Open Access Journal Publishing from 1993 to 2009," Mikeal Laakso et al, 2011, PLoS ONE, 6, p. 6.)

precision to have high reliability in the automatic methods.

This paper is organized as follows: section 2 introduces related work and section 3 presents our semi-automatic conversion method, and section 4 shows error analysis and evaluation. Finally, we give our conclusion and discussion for future directions.

II. Related Work

A. Journal Article Tag Suit

The National Center for Biotechnology Information (NCBI) originally created the Journal Archiving and Interchange Tag Suite with the intent of providing a common format in which publishers and archives could exchange journal content [22]. The JATS v1.0 is a revision of the NLM Journal Archiving and Interchange Tag Suite version 3.0. JATS 1.0 provides a common XML format in which publishers and archives can exchange journal content. This XML is mainly consists of front, body, and back. The element of front describes article header information, which is usually content on first page. The second, body is describes the main body, which includes contents from Introduction to the front of Reference. The back describes whole reference.

B. Rule-based Approach

The rule-based methods generate extraction rules by using knowledge about the target domain and use the rules to extract metadata from pdf documents. This approach is difficult to guarantee the performance when the target domain is changed. But the best performing systems are often handcrafted [19]. The CiteSeer system as the first search engine for scientific literature to incorporate Autonomous Citation Indexing use rule-based metadata extraction system [19]. Also, to overcome weakness in new application domain, after the rule templates about diverse document set are defined and these template are used to classify new document by assigning it to a group of documents of similar layout [23]. The Layout-Aware PDF Text Extraction (LA-PDFText) provides an open source system that extract text blocks from pdf document and classifier them into logical units based on rule [8].

C. Machine Learning Approach

The machine learning methods are popular methods in many text processing fields. This approach is flexible in new document domain but requires that a sufficient training data is available and the task manually labels a set of training data [19]. In Previous research, with enhanced state transition probability, full second order Hidden Markov Models (HMM) proposed [21]. However, HMMs are based on the assumption that features of the model they represent are not independent from each other. To solve this label bias problem and include a wide variety of arbitrary, non-independent feature of the input, Conditional Random Fields (CRFs) method is proposed [22]. With this advantage, The SectLabel extracted logical structure by using richer representation of the document that includes features from Optical Character Recognition (OCR) [7]. To relax segmentation uncertainty and improve extraction performance, [10] proposed co-reference information extraction method. Another method, Support Vector Machine (SVM) was used. To improve the line classification, an iterative convergence procedure was proposed [12]. Also, there are hybrid methods, which mixed one or more machine learning methods. [15] proposed metadata extraction method based on measurement fusion rule. In this experiments, the three learning method such as HMM, SVM and CRF are used.

III. Semi-Automatic Extraction Method

To convert pdf format into well-structured XML automatically, extraction model have to recovery text structure in pdf document by analyzing spatial and layout feature of raw texts. Than The metadata set defined by JATS 1.0 are (article-title, author, affiliation, pub-date, license, abstract, and volume, etc.). The problem to cover range of publication format style (geometric information, layout feature, font feature) from heterogeneous journals still cause that the results tagged as specific class are misleading. Although many of the previous works takes these features into account, the extraction result still is poor in noisy input data (domain is out of space or input data is out of sequence). We consider that it can greatly influence accuracy to refine the input data before applying the machine learning method instead of using full text on pdf document directly into the input data of the machine learning method. Our proposed method applied the rule-base method to refine input data and pre-empt the problem, which is misleading by labeling text chunk block as basic type. Also we adopted machine learning method to label each individual text line to a target class such as element defined in this paper according to JATS 1.0. A brief summary of our proposed work is as follow:

A) An open-source tool, LA-PDFText based on rule, extracts text blocks on pdf documents and determine each type of text block.
B) Block segmentation and line feature are constructed in order to split text block that might be belong to multiple class into text lines.
C) In phase of machine learning, each individual text line and token is labeled as target class with the Condition Random Fields (CRFs) model.

A. Rule-based Text Block Classification

There are various open sources to detect text line or block. Apahce Tika and Apache PDFBox return simple string in pdf document without additional features. CrossRef pdf2xml provides xml format of text with line spatial feature and font feature. The most recent open source of these is the LA-PDFText. It is an open-source tool for accurately extracting text from full-text scientific articles [8]. This open source developed to help with Natural Language Processing (NLP) developer who has to extract text from pdf documents. LA-PDFText extracts not only text block from pdf document but also provides the interface that developer can generate custom rules to classify the type of text block. Fig. 2 is the result of text block classification with LA-PDFText. In this experiments, we designed the rules, which are customized for each journal by analyzing whole format of journals. The used rules for block type classification are in Table 2.

B. Block Segmentation and Line Feature Generation

Text block that separated by the rule spilt cause to consider just layout feature and font feature of text, so it lacks amount of information to parsing by the end result(JATS elements). A text line in Text block can be labeled to target element, or one or more elements. Therefore, First, text block separate a text line unit and once again, to do classification by each line target element. The best case is labeled to one of target element to clarify by rule which decided about text block. In Article-title's case, layout and font feature is clearly different from other text so, in including text for text block case, is almost single class. But Most of text block is labeled to one or more multi class, so it needs segmentation work.

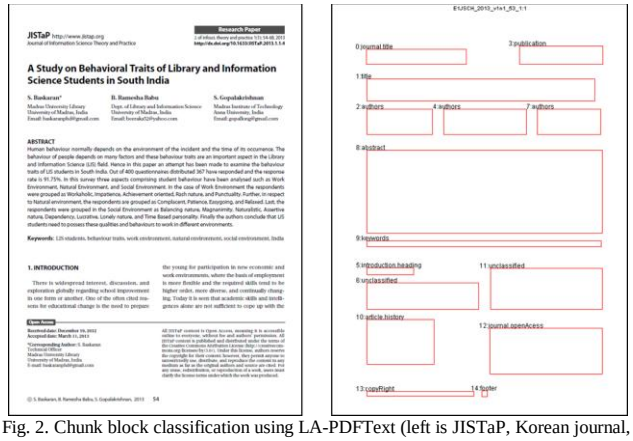

Fig. 2. Chunk block classification using LA-PDFText (left is JISTaP, Korean journal, right is the result of chunk block classification)

TABLE I
THE LIST OF USED RULE FOR TYPE CLASSIFICATION

Block Type	Condition
title	*pageNumber==1* *inTopHalf==true* *mostPopularFontSize>=15* *readLeftRightMidLine()=="MIDLINE"*
publication	*readNumberofLine()<=5* *pageNumber==1* *inTophalf==true* *regularExpression==true ("[vol|no]")*
author	*pageNumber==1* *inTopGalf==true* *regularExpresesion==true ("^([A-Za0z\\.]*))...* *fontSize=="large_font"*
openAccess	*redNumberofLine()>=5* *pageNumber==1* *inTopHalf==true* *regularExpression==true (e.g. 'This is an open Access')*
keyword	*pageNumber==1* *regularExpression==true ("^(Key|Index)")*
	. . .
abstract	*inTropHalf==true* *readLeftRightMidLine()=="MIDLINE"* *readNumberofLine()>=5* *RegularExpression==true ("^(Abst|ABSTR)")*

This is a rule sample, which are used to detect text chunk block (sample journal : GISTaP)

1) Block Segmentation

Text block that separated by the rule spilt cause to consider just layout feature and font feature of text, so it lacks amount of information to parsing by the end result(JATS elements). A text line in Text block can be labeled to target element, or one or more elements. Therefore, First, text block separate a text line unit and once again, to do classification by each line target element. The best case is labeled to one of target element to clarify by rule which decided about text block.

In Article-title's case, layout and font feature is clearly different from other text so, in including text for text block of case, is almost single class. But Most of text block is labeled to one or more multi class, so it needs segmentation work.

TABLE II
THE LIST OF USED FEATURE

Feature	Description	Scope
INITNUM	First letter starts with digit	Front, Back
INITCAP	First letter starts with capital	Front, Back
ALLCAP	All letter is capital	Front, Back
LARFONT	Font size is large	Front, Back
NORFONT	Front size is normal	Front, Back
EMAIL	Character is with e-mail pattern	Front, Back
DATE	Character is with time expression	Front, Back
PAGE	Character is with page pattern	Front, Back
DOI	Character is with doi pttern	Front
NUMBER	Character including digit	Front
FONT-SIZE	Font size	Front, Back
FONT-STYLE	Font style	Front, Back
ISSN	Character is with issn pattern	Front
PAGE	The page including current character	Front
LINE_LONG	Text line is long	Back
LINE_MDLLE	Text line is middle	Back
LINE_SHORT	Text line is short	Back
PERSON	Character is with person name	Front, Back
ORGANIZATION	Character is with organization	Front, Back
LOCATION	Character is with location	Front, Back
LAST-DOT	Last letter end with dot	Back
LAST-COMMA	Last letter end with comma	Back
BLOCK-TYPE	Text block type including current text	Front, Back

2) Line Feature Generation

It is point of performance of Machine-learning that wise

choice of feature set [10]. Combination to key feature which can classify the class is more helpful for performance quality than collecting various feature set. In this experiment, 23 kinds of feature are used. Table 2 is list of feature used in this experiment. Lexical feature, font feature, word entity feature, chunk type feature and so on of text line is used. The type of text block which is already determined by established rule clarify indirect target element of the text. Such a text block classification through rule-based method is very useful in extracting key feature.

C. Machine Learning-based Classification

The module of semi-automatic metadata extraction combined with rule-based and machine-learning is illustrated in Fig. 3. In this step, text line set divided in previous step is need to classification as each target element again. In this experiment, we use Conditional Random Field (CRF) in order to classify appropriate target element by considering a lot of context of text line and feature of each text. CRF is applied on field to solve sequence labeling problem because it can solve the label bias problem of Hidden Markov Model (HMM) and support arbitrary dependent feature and joint inference over entire sequence. The classification model we propose is divided into two models to extract front metadata and back metadata. First, model which tag elements of front area defined by JATS. Second, model which tag elements of back area.

1) Front Classification

Front area that defined by JATS is consist of 18 kinds of sub-element in <journal-meta> and 39 kinds of sub-element in <article-meta>. In First column of Table 6, in this experiment, there are 18 kinds of element that defined target class. Out of text block that determined single class by 18 previous rules, the other text block that has multiple classes is applied front CRF model.

2) Back Classification

Back part is consists of seven elements. Nature of LA-PDFText that finds the continuous text block, all of reference body is cognized one block by text layout and font feature. To extract each individual reference block, we make two kinds of classification step. One is looking up and separating individual reference blocks in all of reference body. Previously, separate the reference text in body by line, tracking down first and last line, separating to individual reference. The other is dividing individual reference as token unit. Because of reference has a multi class on one text line. After divided text token set as token unit generating feature-vector, eventual seven elements (author, title, source, year, volume, number and page) are classified.

IV. Error Analysis and Evaluation

A. Experimental Data

In this paper, the dataset to test our semi-automatic metadata extraction module is collected from Open Access Korea Central (OAK Central), which is a free archive of scholarly & scientific journal literature in Korea. We downloaded 560 papers from 26 different journals in OAK Central. In whole dataset, 260 papers are used for training set and rest 300 papers are used for testing set.

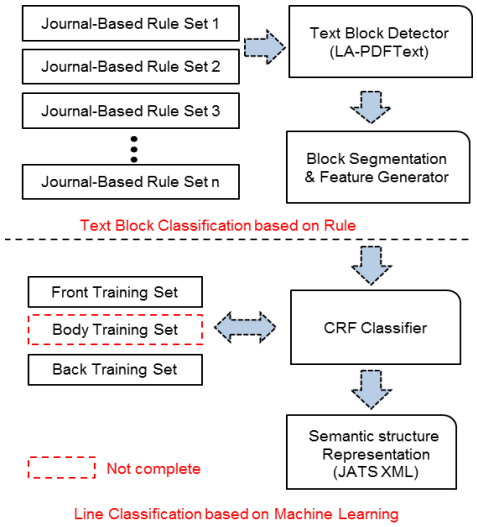

Fig. 3. Overview of Semi-Automatic Metadata Extraction Module

B. Evaluation Metrics

To verify our proposed method, we used basic measure, Precision (P), Recall (R), and F1-measure. This thesis defines the following:.

A: The number of true positive elements (e.g. 'title' token tagged as 'title').

B: The number of false negative elements (e.g. non-title token tagged as 'title')

C: The number of false positive elements (e.g. 'title' token tagged as anything but 'title')

Precision (P) = A / (A + C)
Recall (R) = A / (A + B)
F1 = (2 * P * R) / (P + R)

C. Evaluation of Extraction Result

Table 4 is precision, recall and F1-measure of front elements. Before applying machine learning method, using rule-based methods gets high precision on abstracting text block and Classifying basic type about text block. Text block which belong single class is tagged as JATS element, not applied CRF. Typical single class includes title, doi, abstract, accepted-date and received date. If we analysis single and multiple class about text block previously and use refined data in machine learning method, can expect reliable output. Table 5 is a result of back part. In the case of back part, we extract metadata only using machine learning method.

TABLE VII
CONFUSION MATRIX OF FRONT ELEMENTS

	j-id	*j-title*	*a-title*	*a-doi*	*c-group*	*aff.*	*a-date*	*r-date*	*c-holde*	*license*	*abstra.*	*volume*	*issue*	*other*
j-id	18	0	0	1	0	0	0	0	0	0	0	0	0	0
j-title	0	76	0	0	0	0	0	0	0	0	0	0	0	22
a-title	0	0	98	0	0	0	0	0	0	0	0	0	0	0
a-doi	1	0	0	80	0	0	0	0	0	0	0	0	0	13
c-group	0	0	0	0	226	14	0	0	0	0	3	0	0	34
aff.	0	0	0	0	0	160	0	0	0	0	3	0	0	31
acc-date	0	0	0	0	0	0	89	0	0	0	0	0	0	0
rec-date	0	0	0	0	0	0	0	89	0	0	0	0	0	0
co-holder	0	0	0	0	0	0	0	0	60	0	0	0	0	18
license	0	0	0	0	0	0	0	0	0	54	0	0	0	4
abstract	0	0	0	0	0	0	0	0	0	0	95	0	0	3
volume	0	0	0	0	0	0	0	0	0	0	0	98	0	1
issue	0	0	0	0	0	0	0	0	0	0	0	0	88	1

D.Error Analysis

Table 7 is classification confusion matrix about front area of JATS XML. Table 6 is classification confusion matrix about back area. We need separating as each individual reference and classifying as JATS back elements from individual reference because extracted reference body by rule-based method includes all of reference area. Eventually, classifying as JATS back elements take a lot of effect on output of separated from reference body. Most errors of reference are originated in separating as individual reference procedure. It is case that two references are classified one reference and one reference is classified two references.

V. CONCLUSION AND FUTURE WORK

Appearance of JATS which standard Open Access (OA) policy shows further need for the research that extracts structured data and analysis semantic logical structure. Many scholarly journal publishers in Korea still rely on manual information extraction owing to low reliability in the automatic metadata extraction system. To deal with the problem of low reliability in the automatic metadata extraction and help with minimum human interaction, we propose semi-automated metadata extraction method based on rule-based method and machine learning method.

The model proposed in this experiment mixed rule-based and machine-learning is verified under scholarly publisher that treat small scale journal, not search engine or digital library which highlights scalability. We verified the performance under 26 different journals in Open Access Korea Central (OAK Central). Thus, proposed model prosecute text block classification and tags block which type of clear single class as target class, do not use machine-learning method. If we do not analyze logical structure of journal and only treat all of text in article with machine-learning, we hardly see a result of extraction for service. If we remove ambiguity of text block through rule-based method and apply machine-learning method after refining data, we can expect reliable high quality data.

Several issues remain to be investigated. First, the portion of body in JATS is difficulty to extract metadata such as figure extraction, table recognition, and mathematical expression automatically. Second, our experiments only are conducted in small specific journals (26 journals in OAK Central) to get high reliability and help with minimum human interaction. Finally, open data integration problem are occurred because we use the rule-based methods

TABLE IV
THE ACCURACY OF FRONT ELEMENTS

Article-front	**P**recision	**R**ecall	**F**1
Element	**96.47**	89.35	**92.51**
journal-id	94.74	94.74	94.74
journal-title	100.0	77.55	87.36
article-title	100.0	100.0	100.0
article-id(doi)	98.77	85.11	91.43
contrib-group	100.0	81.59	89.86
affiliation	91.95	82.47	86.96
accepted-date	100.0	100.0	100.0
received-date	100.0	100.0	100.0
copyright-holder	100.0	76.92	86.96
license	100.0	93.10	86.43
abstract	94.06	96.94	95.48
volume	100.0	98.99	99.49
issue	100.0	98.88	99.44

Precision, recall and F1-measure of the front from OAK Central data set (%)

TABLE V
THE ACCURACY OF BACK ELEMENTS

Article-front	**P**recision	**R**ecall	**F**1
Average	**96.80**	92.00	**94.15**
person-group	98.49	97.12	97.80
article-title	85.05	97.12	90.69
page	98.58	93.78	96.12
year	99.22	95.12	97.12
source	97.28	80.74	88.24
volume	98.96	90.60	94.59
number	100.0	89.53	94.48

Precision, recall and F1-measure of the front OAK Central data set (%)

TABLE VI
CONFUSION MATRIX OF FRONT ELEMENTS

	name	*title*	*page*	*year*	*sou.*	*vol.*	*num.*	*other*
name	3135	42	0	0	0	0	0	51
title	18	3241	0	0	33	0	0	42
page	3	9	2712	24	9	6	0	129
year	3	18	3	3039	3	3	0	126
source	24	477	15	0	2679	0	0	123
volume	0	6	21	0	24	1995	0	156
number	0	0	0	0	6	12	462	36

REFERENCES

[1] Y. Gargouri, C. Hajjem, V. Larivière, Y. Gingras, L. Carr, T. Brody and S. Harnad, "Self-selected or mandated, open access increases citation impact for higher quality research," *Plos one*, 5(10), 2010, e13636.

[2] M. Laakso, P. Welling, , H. Bukvova, L. Nyman, B. C. Björk and T. Hedlund, "The development of open access journal publishing from 1993 to 2009," *PloS one*, 6(6), 2011, e20961.

[3] S. Huh, "ScienceCentral: open access full-text archive of scientific journals based on Journal Article Tag Suite regardless of their languages," *Biochemia medica*, 23(3), 2013, p.235-236.

[4] B. C. Björk, P. Welling, M. Laakso, P. Majlender, T. Hedlund and G. Guðnason, "Open access to the scientific journal literature: situation 2009," *PloS one*, 5(6), 2010, e11273.

[5] A. Constantin, S. Pettifer, and A. Voronkov, "PDFX: fully-automated PDF-to-XML conversion of scientific literature," *Proceedings of the 2013 ACM symposium on Document engineering*, ACM, 2013, p.177-180.

[6] R. Kern, K. Jack, M. Hristakeva and M. Granitzer, "TeamBeam Meta-Data Extraction from Scientific Literature," *D-Lib Magazine*, 18(7), 2012, 1.

[7] M. T. Luong, T. D. Nguyen and M. Y. Kan, "Logical structure recovery in scholarly articles with rich document features," *International Journal of Digital Library Systems (IJDLS)*, 1(4), 2010, p.1-23.

[8] C. Ramakrishnan, A. Patnia, E. H. Hovy and G. A. Burns, "Layout-aware text extraction from full-text PDF of scientific articles," *Source code for biology and medicine*, 7(1), 2012, 7.

[9] M. Granitzer, M. Hristakeva, R. Knight, K. Jack, and R. Kern., "A comparison of layout based bibliographic metadata extraction techniques," in *Proc. the 2nd International Conference on Web Intelligence, Mining and Semantics*. ACM, June, 2012, p. 19.

[10] F. Peng, and A. McCallum, "Information extraction from research papers using conditional random fields," *Information Processing & Management*, 42(4), 2006, p.963-979.

[11] S. Klampfl and R. Kern, "An Unsupervised Machine Learning Approach to Body Text and Table of Contents Extraction from Digital Scientific Articles," *Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg, 2013, p. 144-155.

[12] H. Han, C. L . Giles, E. Manavoglu, H. Zha, Z. Zhang and E. A. Fox, "Automatic document metadata extraction using support vector machines," *Digital Libraries, Joint Conference on*. IEEE, 2003, p. 37-48.

[13] J. Chen and H Chen. "A Structured Information Extraction Algorithm for Scientific Papers based on Feature Rules Learning," *Journal of Software (1796217X)*, 8(1), 2013.

[14] D. Tkaczyk, L. Bolikowski, A. Czeczko and K. Rusek, "A modular metadata extraction system for born-digital articles.," *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*, IEEE, 2012, p. 11-16.

[15] J. Zhao and H. Liu, "Metadata Extraction Approach of PDF Documents Based on Measurement Fusion," *Journal of Multimedia*, 8(6), 2013

[16] P. A. Praczyk and J. Nogueras-Iso, "Automatic extraction of figures from scientific publications in high-energy physics," Information Technology and Libraries, 32(4), 2013, p. 25-52.

[17] G. Eysenbach, "Citation advantage of open access articles," PLoS biology, 4(5), 2006, e157.

[18] K. Antelman, "Do open-access articles have a greater research impact," College & research libraries, 65(5), 2004, p.372-382.

[19] I. G. Councill, C. L. Giles, E. Di Iorio, M. Gori, M. Maggini and A. Pucci, "Towards next generation citeseer: A flexible architecture for digital library deployment," *Research and Advanced Technology for Digital Libraries*, Springer Berlin Heidelberg, 2006, p.111-122.

[20] Z. Fuzhi, and Z. Zhao, "A Metadata Extraction Approach from Papers Based on Meta-learning," 2013.

[21] B. Ojokoh, M. Zhang and J. Tang, "A trigram hidden Markov model for metadata extraction from heterogeneous references," *Information Sciences* 181(9), p.1538-1551.

[22] J. Lafferty, A. McCallum and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[23] J. Beck, "NISO Z39. 96The Journal Article Tag Suite (JATS): What Happened to the NLM DTDs?," *The journal of electronic publishing: JEP*, 14(1), 2011.

[24] P. Flynn, L. Zhou, K. Maly, S. Zeil and M. Zubair, "Automated template-based metadata extraction architecture," *Asian Digital Libraries, Looking Back 10 Years and Forging New Frontiers*, Springer Berlin Heidelberg, 2007, p.327-336.