Oriental Journal of Computer Science & Technology Vol. **2**(2), 101-111 (2009)

Rule-based metadata extraction for heterogeneous references

BOLANLE OJOKOH

Department of Computer Science and Technology,
School of Electronics Engineering and Computer Science, Peking University, Beijing, 100 871 (China).

(Received: November 06, 2009; Accepted: December 19, 2009)

ABSTRACT

References form an essential part of electronic scholarly publications. Accurate and automatic reference metadata generation provides scalability, interoperability and usability for digital libraries and their collections. This paper deals with automatic metadata extraction from the references of general digital documents using rule-based approach. It encompasses automatic extraction of metadata from book and journal references. The system consists of four major components: a means of providing reference input (by uploading the file or providing the set of references in the window provided by the browser), the text converter for converting documents into standard text format, the parser for automatically extracting metadata such as reference style, author, title, journal, volume, number (issue), year, and page information and author, title, publisher, place of publication, year and pages information from book and journal references of the converted documents using pre-defined regular expressions, and the browser for displaying the results.The experimental results show that the proposed framework can be used to extract metadata from different reference styles of book and journal references effectively.

Key words: metadata, implementation, experiment, references, digital libraries, information extraction.

INTRODUCTION

Metadata is, most generally, data that describes other data to enhance their usefulness in content explanation. Metadata extraction is a special case of information extraction. It occurs when an algorithm automatically extracts metadata from a resource's content[9]. Building tools for automatic metadata extraction and representation significantly improve the amount of metadata available, the quality of metadata extracted, and the efficiency and speed of the metadata extraction process[7].

References, are most commonly found in a late section of an article; this section is often labeled “References”, “Bibliography” or “List of References”, and information that is normally contained in this section include the name(s) of author, title, journal, volume, number (issue), year, and page information which have constituted an important kind of metadata; valuable for literature search, analysis, and evaluation[6]. Accurate reference metadata extraction for scholarly publications is essential for the integration of information from the available heterogeneous reference sources[2].

Automatic reference extraction is particularly difficult because of the variations in field separators. For example, in journal references, the author and title fields can be separated by spaces or periods; while the volume and issue fields can be separated by braces or parentheses. Within fields, further separator issues are caused by punctuation and spacing differences. To further compound the problem; there are many dramatically different citation styles (that is, different field orders). Some systems have attempted to extract citation information from digital document references. For instance, Powley and Dale[12] tried extracting only author name for references. Besagni *et al.* [1] used part-of-speech tagging of words in references (from a corpus of pharmacology journal papers) to segment them, and report 90.2% accuracy in extracting author names. Wellner *et al.*[15] used conditional random fields for reference segmentation and coreference resolution on a collection of references from CiteSeer, reporting segmentation accuracy across all fields of 94.9%. Takasu[14] employed hidden Markov models and support vector machines for reference segmentation, reporting high accuracy results, but pointing out that their test corpus, comprising papers from a single journal, had extremely consistent formatting. Ding *et al.*[5] used a rule-based ('template mining') system for reference segmentation, reporting 95% accuracy in extracting only author names. Both Chowdhury[3] and Ding *et al.*[5] worked with tagged references in one style only. Day *et al.*[4] proposed a template-based approach for reference metadata extraction from journal references only and confirmed reliable accuracy for some styles of referencing. They, however, made use of some existing tools like Journal Spider to retrieve citations and Compass for knowledge editing.

This paper proposes a unified approach to reference metadata extraction. It presents an all-encompassing system that extracts metadata from books, journals and other general citations such as conference proceedings and theses and dissertations. The proposed approach solves the problem of small-scale metadata extraction for digital documents. It goes beyond just extracting metadata from references written in a particular style or from just text documents only. It accepts documents that can be in PDF, DOC or text format, and has the capability to represent and match template structures of regular expressions formed for different reference styles from the accepted natural language text after which the set of metadata are extracted from different kinds of reference styles.

The remainder of this paper is organized as follows: Section 2 describes the proposed system's architecture. Section 3 presents the implementation and experimental results while Section 4 gives the conclusion and directions for future research.

Proposed System Architecture
Mathematical Model of the System

R is the Rule-Based Reference Metadata Extractor. It consists of six components. The functions of these components and the relationship among them are expressed in equations 1-26.

$$\text{R is a 6-tuple (Input, Converter, Parser, Browser, Output, Database)} \quad \text{...(1)}$$

$$Input = \begin{cases} u & \text{if p = null} \\ p & \text{otherwise} \end{cases} \quad \text{...(2)}$$

where,

u refers to uploaded document (e.g. in PDF, HTML or Word format), while p is the pasted references showing that the input is either an uploaded document or pasted references.

The converter receives the uploaded document and converts to text (equivalent format of pasted references) as shown in equation 3.

$$Converter : u \rightarrow p \text{ , then} \quad \text{...(3)}$$

The set of references from which metadata are to be extracted could be book, journal or other types of references respectively (Equation 4). Each reference is being mapped to its set of metadata as described in equation 5, and consists of a set of metadata and separators as outlined in equations 6.

$$p \in \left\{ r_b, r_j, r_o \right\} \quad \text{...(4)}$$

and

$$\left(f : r_b \rightarrow m_b \middle| r_j \rightarrow m_j \middle| r_o \rightarrow m_o \right) \quad \text{...(5)}$$

where, r_b, r_j and r_o refer to book, journal and other references respectively and m_b, m_j and m_o are used to represent set of book, journal and other metadata respectively. Every reference consists of a set of metadata and a set of separators (equation 7) as described in equation 6.

$$r_b \rightarrow \{m_b, s\}, r_j \rightarrow \{m_j, s\}, r_o \rightarrow \{m_o, s\} \quad ...(6)$$

where,

$$s \in \{., (,), ", ", ', ', ;, :\} \quad ...(7)$$

The metadata present in the books, journals and other references are outlined in equations 8-10.

$$m_b \in \{author, year, title, edition, volume, place, publisher, pages\} \quad ..(8)$$

$$m_j \in \{author, year, title, journal, volume, issue, pages\} \quad ...(9)$$

$$m_o \in \{author, year, title, website, encyclopedia, conference, affiliation\} \quad ...(10)$$

The Parser receives the converted references as input, maps each of the reference to the pattern that fits from the template of patterns defining the reference styles and subsequently extracts metadata by structuring every record to fields. These processes are outlined by equations 11-20.

$$\text{(iii) Parser (Input, Map, Extract)} \quad ...(11)$$

$$\text{Input} = \text{p} \quad ...(12)$$

$$styles = \left\{\begin{matrix} AMA, APA, ACM, BIOI, Chicago, IEEE, JCB, MISQ, MIA, \\ Turabian, AMA, APA, ACM, BIOI, Chicago, IEEE, JCB, \\ MISQ, MIA, Turabian, common, other \end{matrix}\right\} \quad ...(134)$$

For other references, there are different types as shown in equation 14

$$\text{type} = \{\text{works, website, encyclopedia, conf. proc., thesis, unpublished}\} \quad ...(14)$$

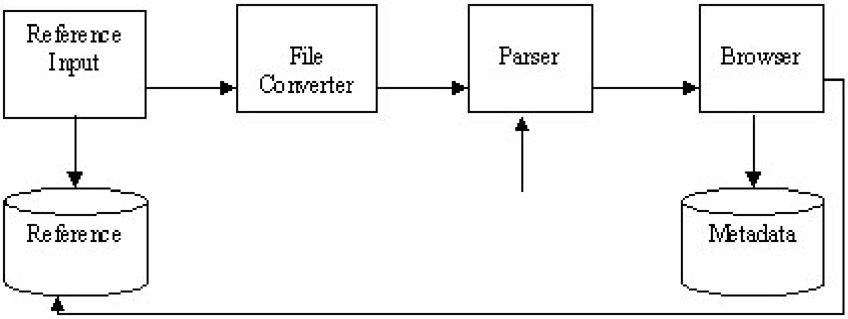


Fig. 1: Reference metadata extractor architecture

```
<MISQbook>:=<Author/editor-name>\.<Title>, < Publisher >,< Place of publication >,
<Year>\.
<Author/editor-name>:= (.*)
<Title>:= (.*)
<Publisher>:=  * ([A-Za-z\s0-9\(\)\-:,\&]*)
 <Place of publication>:= * ([A-Za-z\s0-9\(\)\-:,\&]*)
<Year>:= ([0-9]{4})
```

Fig. 2: Regular expressions for analyzing the MISQ book reference style

$$Map : Parser(s, m_b) \rightarrow stored\ patterns \rightarrow \begin{cases} AMA_b \\ APA_b \\ ACM_b \\ BIOI_b \\ Chicago_b \\ IEEE_b \\ JCB_b \\ MISQ_b \\ MLA_b \\ Turabian_b \\ error & \text{otherwise} \end{cases} \quad ...(15)$$

$$Map : Parser(s, m_j) \rightarrow stored\ patterns \rightarrow \begin{cases} AMA_j \\ APA_j \\ ACM_j \\ BIOI_j \\ Chicago_j \\ IEEE_j \\ JCB_j \\ MISQ_j \\ MLA_j \\ Turabian_j \\ common_j \\ other_j \\ error & \text{otherwise} \end{cases} \quad ...(16)$$

$$Map : Parser(s, m_o) \rightarrow stored\ patterns \rightarrow \begin{cases} works \\ website \\ encyclopedia \\ conference\ proceeding \\ thesis \\ unpublished \\ error & \text{otherwise} \end{cases} \quad ...(17)$$

For every reference mapped without error,

$$Extract : (s * m_b) \rightarrow author * year * title * edition * volume * place * publisher * pages \quad ...(18)$$

or

$$Extract : (s * m_j) \rightarrow author * year * title * journal * volume * issue * pages \quad ...(19)$$

or

$$Extract : (s * m_o) \rightarrow author * yr * title * website * encyclopedia * conference * affiliation \quad ...(20)$$

The output is being displayed by the browser, while the metadata is being stored in the database as outlined in equations 21-26.

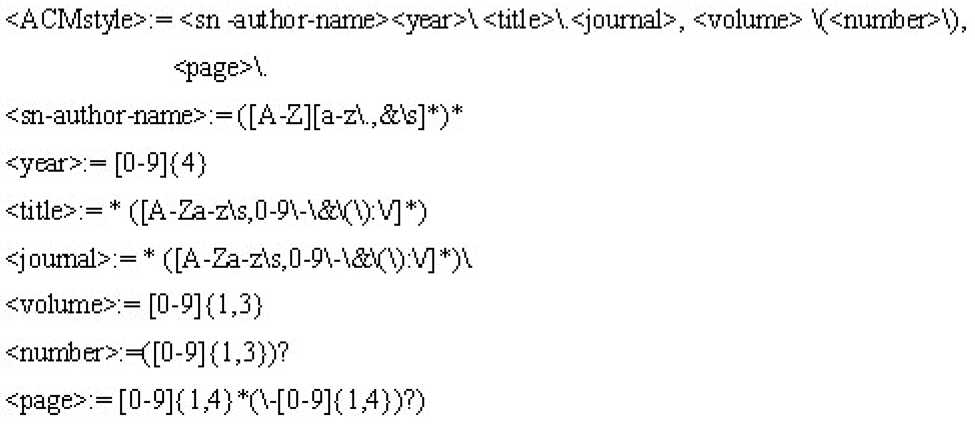
$$\forall m_b \text{ extracted, Browser displays Output} \in \{\text{style}, m_b\} \quad ...(21)$$

$$\forall m_j \text{ extracted, Browser displays Output} \in \{\text{style}, m_j\} \quad ...(22)$$

$$\forall m_o \text{ extracted, Browser displays Output} \in \{\text{type}, m_o\} \quad ...(23)$$

$$Store : (Database : r_b, m_b) \quad ...(24)$$

$$Store : (Database : r_j, m_j) \quad ...(25)$$



```
<ACMstyle>:= <sn -author-name><year>\ <title>\.<journal>, <volume> \(<number>\),
                    <page>\.
<sn-author-name>:=([A-Z][a-z\.,&\s]*)*
<year>:= [0-9]{4}
<title>:= * ([A-Za-z\s,0-9\-\&\(\):V]*)
<journal>:= * ([A-Za-z\s,0-9\-\&\(\):V]*)\
<volume>:= [0-9]{1,3}
<number>:=([0-9]{1,3})?
<page>:= [0-9]{1,4}*(\-[0-9]{1,4})?)
```

Fig. 3: Regular expressions for analyzing the ACM Journal Reference Style

$$Store : (Database : r_o, m_o) \quad ...(26)$$

Reference Metadata Extractor Architecture

The architecture of the reference metadata extractor is shown in Fig. 1. It consists of four main components:

Reference Input, Text Converter, Parser and Browser

The input to the system could be uploaded text, pdf, word or html documents from which references are stripped or a set of references pasted in a text area provided by the system. Any input in other formats will be converted to text by the text converter. The parser receives the text document/ file and extracts the references metadata based on a set of patterns (regular expressions) through which it analyses the references parsed. There are some common styles through which journal and book information can be provided in references by the system[11]. Examples of some common and generally used reference styles used to test the efficiency of the system are American Medical Association (AMA) , Association for Computing Machinery (ACM), Bioinformatics (BIOI), Chicago, Institute of Electrical and Electronics Engineers (IEEE), Journal of Cell Biology (JCB), MIS Quarterly (MISQ), MLA and Turabian.

Table 1: Experimental results of precision of reference metadata extraction from ten journal reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue	Pages	Average
APA	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
JCB	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MISQ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
IEEE	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ACM	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0
BIOI	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Chicago	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MLA	1.0	1.0	1.0	1.0	1.0	1.0	0.95	0.99
AMA	0.99	1.0	0.99	0.99	0.99	-	0.96	0.99
Turabian	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Overall average	0.998	1.0	0.999	0.999	0.999	1.0	0.99	

Table 2: Experimental results of recall of reference metadata extraction from ten journal reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue	Pages	Average
APA	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
JCB	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MISQ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
IEEE	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
ACM	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
BIOI	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Chicago	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MLA	1.0	1.0	1.0	1.0	1.0	1.0	0.95	0.99
AMA	0.98	0.98	0.98	0.98	0.98	-	0.94	0.97
Turabian	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Overall Average	0.98	0.98	0.98	0.98	0.98	0.98	0.97	

Examples of such reference styles are: Witten, I.H., & Bainbridge, D. 2003. How to Build a Digital Library. Morgan Kaufmann Publishers, California. 518pp. for the JCB book style and T. Davenport, D. DeLong, and M. Beers, "Successful Knowledge Management Projects," Sloan Management Review, vol. 39, no. 2, pp. 43-57, 1998. for the IEEE Journal style.

The set of regular expressions formed for some selected styles for journal and book reference styles are presented in Figs. 2 and 3.

The metadata extracted from the lists of references are displayed to the users through the browser and can be stored inside the metadata database for future use. There are six tables in the database named references. There is a table each for book, journal and other references. The others contain the extracted reference metadata for the journal references, book references and each of the other types of references respectively.

System Implementation and Experimental Results

The rule-based reference metadata was implemented on a Pentium M 1.6GHz Machine with 1GB RAM and 100GB Hard disk space. The application runs on Windows XP Operating System. Appserv-win 32-2.5 – a suite comprising Apache,

Table 3: Experimental results of accuracy of reference metadata extraction from ten journal reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue	Pages	Average
APA	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
JCB	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MISQ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
IEEE	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.91
ACM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.0
BIOI	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
Chicago	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MLA	1.0	1.0	1.0	1.0	1.0	1.0	0.95	0.99
AMA	0.97	0.98	0.97	0.97	0.98	-	0.94	0.97
Turabian	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Overall Average	0.98	0.98	0.98	0.98	0.98	0.98	0.97	

Table 4: Experimental results of F-measure of reference metadata extraction from ten journal reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue	Pages	Average
APA	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
JCB	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MISQ	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
IEEE	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
ACM	0.99	1.0	1.0	1.0	1.0	1.0	1.0	1.0
BIOI	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Chicago	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MLA	1.0	1.0	1.0	1.0	1.0	1.0	0.95	0.99
AMA	0.98	0.99	0.98	0.98	0.98	-	0.95	0.98
Turabian	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Overall Average	0.99	0.99	0.99	0.99	0.99	0.99	0.98	

MySQL and PHP was used to run the programs, while Internet Explorer helps to display the web pages. Macromedia Dream weaver was used to edit the scripts, while a good user interface was provided for the system using HTML codes. PHP particularly suits the purpose of its use here as the scripting language especially because it supports complex pattern matching with regular expressions. MySQL was used for managing the database.

The reference metadata extractor can receive an uploaded document in different formats and convert it to a text document. It locates the reference section after which it does the extraction of the metadata. The reference(s) for which metadata are to be extracted can also be entered in a multi-line HTML text area and passed to the system after clicking on the hyperlink "Strip Reference". The result is displayed in tabular format on a web page. Four different groups of references were tested with the system. These include ten different journal reference styles and ten different book reference styles.

Some other general reference types such as encyclopedia articles, conference proceedings, works by associations or corporations and thesis or dissertation were also tested with the system. Patterns were also generated for some other types of journal references that do not fall under any of

Table 5: Experimental results of Precision of reference metadata extraction from ten book reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue
APA	0.99	1.0	0.99	1.0	0.99	0.99
JCB	1.0	1.0	1.0	1.0	1.0	1
MISQ	0.97	1.0	0.87	0.88	0.90	0.92
IEEE	0.92	1.0	0.92	0.98	1.0	0.96
ACM	0.99	1.0	0.99	0.99	1.0	0.99
BIOI	0.96	1.0	0.97	0.98	0.98	0.98
Chicago	0.96	1.0	1.0	1.0	0.97	0.99
MLA	1.0	1.0	1.0	1.0	1.0	1
AMA	1.0	1.0	1.0	1.0	1.0	1
Turabian	0.96	1.0	1.0	1.0	0.97	0.99
Overall Average	0.98	1.0	0.97	0.98	0.98	

Table 6: Experimental results of Recall of reference metadata extraction from ten book reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue
APA	0.95	0.96	0.95	0.94	0.95	0.95
JCB	0.93	0.93	0.93	0.93	0.93	0.93
MISQ	0.94	0.94	0.82	0.93	0.93	0.91
IEEE	0.81	0.83	0.81	0.82	0.83	0.82
ACM	0.95	0.96	0.95	0.95	0.96	0.95
BIOI	0.93	0.94	0.94	0.94	0.94	0.94
Chicago	0.98	0.99	0.99	0.99	0.99	0.99
MLA	0.93	0.93	0.93	0.93	0.93	0.93
AMA	0.98	0.98	0.98	0.98	0.98	0.98
Turabian	0.98	0.99	0.99	0.99	0.99	0.99
Overall Average	0.94	0.95	0.93	0.94	0.94	

the identified styles. They were referred to as "common" styles.

A sample output of the rule-based reference metadata extractor is shown in figure 4. In quantifying the performance of the reference metadata extractor, standard quality measure criteria, namely: recall, precision, accuracy, and F-measure were used. These are defined in Eqs 27 – 30. Han *et al.*[7], Hu *et al.* [8], Taghva *et al.* [13], and Powley and Dale[12] used the same criteria for evaluating their information extraction systems. The system was tested with 980 records of journal references in 10 different styles downloaded from the Web and 1000 records of book references in 10 different styles (some of which were manually generated).

An exact performance comparison may not be possible, because of differences in the documents used for testing the different systems [10]. However, attempts are made to refer to related work and how the results compare with theirs. The results of the experimental evaluation of the reference metadata extractor are presented as follows.

$$\text{Precision} = \frac{A}{A+C} \qquad ...(27)$$

Table 7: Experimental results of Accuracy of reference metadata extraction from ten book reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue
APA	0.94	0.95	0.94	0.94	0.94	0.94
JCB	0.93	0.93	0.93	0.93	0.93	0.93
MISQ	0.91	0.94	0.82	0.83	0.85	0.87
IEEE	0.76	0.83	0.76	0.81	0.83	0.80
ACM	0.94	0.95	0.94	0.94	0.95	0.94
BIOI	0.89	0.94	0.91	0.92	0.92	0.92
Chicago	0.95	0.99	0.99	0.99	0.96	0.98
MLA	0.93	0.93	0.93	0.93	0.93	0.93
AMA	0.98	0.98	0.98	0.98	0.98	0.98
Turabian	0.95	0.99	0.99	0.99	0.96	0.98
Overall Average	0.93	0.94	0.92	0.93	0.93	

Table 8: Experimental results of F-measure of reference metadata extraction from ten book reference styles

Reference style	Author	Year	Title	Journal	Volume	Issue
APA	0.97	0.98	0.97	0.97	0.97	0.97
JCB	0.96	0.96	0.96	0.96	0.94	0.96
MISQ	0.95	0.97	0.90	0.90	0.91	0.93
IEEE	0.86	0.91	0.86	0.89	0.91	0.87
ACM	0.97	0.98	0.97	0.97	0.98	0.97
BIOI	0.94	0.97	0.95	0.96	0.96	0.96
Chicago	0.97	0.99	0.98	0.98	0.98	0.98
MLA	0.96	0.96	0.96	0.96	0.96	0.96
AMA	0.99	0.99	0.99	0.99	0.99	0.99
Turabian	0.97	0.99	0.98	0.98	0.98	0.98
Overall Average	0.95	0.97	0.95	0.96	0.96	

$$\text{Recall} = \frac{A}{A+B} \quad ...(28)$$

$$\text{Accuracy} = \frac{A+D}{A+B+C+D} \quad ...(29)$$

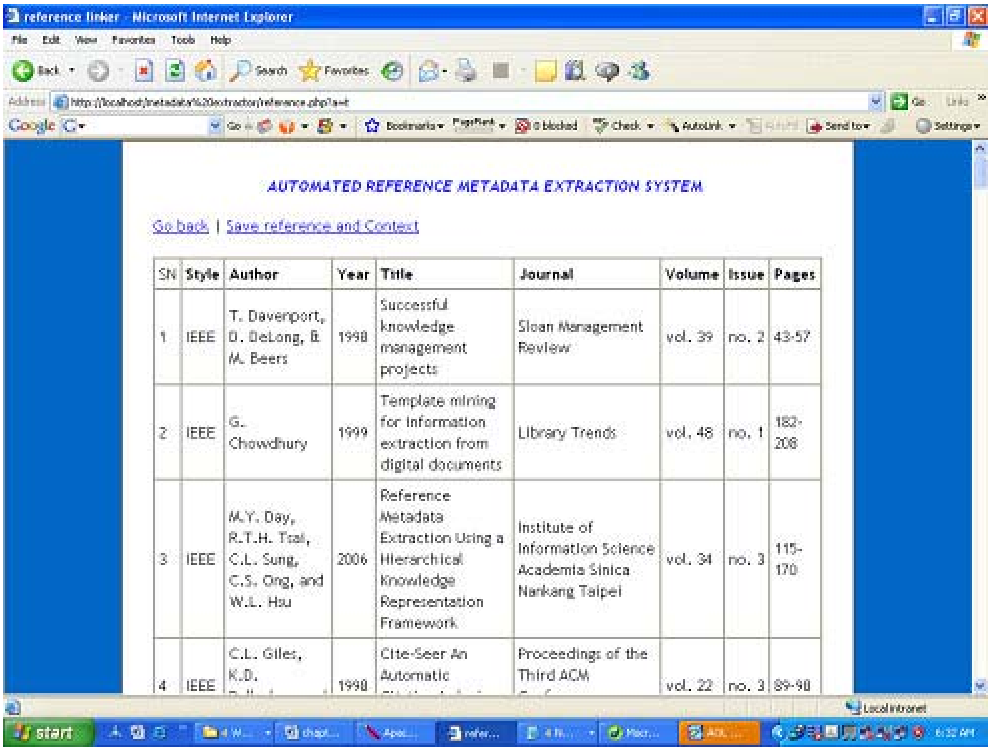
$$\text{F-measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad ...(30)$$

A is the number of correctly extracted fields. B is the number of fields with existing but not extracted data. C is the number of fields with wrongly extracted data, while D is the number of fields with not existing and not extracted data.

Tables 1 – 8 summarize the experimental results of metadata extraction for the ten different journal and book reference styles.

The average precision for the journal styles are very high (1.0 for all the styles) except for MLA and AMA Journal Styles with precision of 0.99 each. This is due to the peculiary close patterns of the author, title and journal fields following one another consecutively in the two styles.

The average recall and accuracy is 1.0 each for the MISQ, JCB, Turabian and Chicago styles. MLA style has recall and accuracy of 0.99; APA, BIOI and AMA each has recall and accuracy of 0.97, while IEEE has the least (0.91). This shows that for most of the styles, reference metadata are extracted accurately for most references. IEEE extracts with the least accuracy and precision because of the closeness in the patterns of some consecutive fields, hence the inability to match some the metadata of some references with some special characters such as question marks in their titles. Averagely, the year and issue fields of the journal references are extracted with the highest degree of precision. For the recall, accuracy and F-measure, all the fields are extracted with almost the same degree, except for the page field which has a bit lower recall, accuracy and F-measure. This is mostly due to the fact that in the extraction of the


reference linker - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://localhost/metadata%20extractor/reference.php?a=t
Go
Links
Google
Bookmarks
PageRank
0 blocked
Check
AutoLink
Send to
Settings
Local intranet
start
6:32 AM

AUTOMATED REFERENCE METADATA EXTRACTION SYSTEM

Go back | Save reference and Context

SN	Style	Author	Year	Title	Journal	Volume	Issue	Pages
1	IEEE	T. Davenport, D. DeLong, & M. Beers	1998	Successful knowledge management projects	Sloan Management Review	vol. 39	no. 2	43-57
2	IEEE	G. Chowdhury	1999	Template mining for information extraction from digital documents	Library Trends	vol. 48	no. 1	182-208
3	IEEE	M.Y. Day, R.T.H. Tsai, C.L. Sung, C.S. Ong, and W.L. Hsu	2006	Reference Metadata Extraction Using a Hierarchical Knowledge Representation Framework	Institute of Information Science Academia Sinica Nankang Taipei	vol. 34	no. 3	115-170
4	IEEE	C.L. Giles, K.D.	1998	Cite-Seer An Automatic	Proceedings of the Third ACM	vol. 22	no. 3	89-98

Fig. 4: A sample output of the reference metadata extractor

page information in some styles, some part of the metadata are left not extracted, for example extracting “24” as page instead of “24-39”.

For the book references, there was a greater challenge in the metadata extraction process for the system because almost all the styles have consecutively related patterns for the fields. Moreover, many of the styles have consecutive related separators. Nevertheless, the JCB, MLA and AMA styles have 1.0 precision, some four others have 0.99 precision, while MISQ has the least precision of 0.92. This occurs mostly because the title and place fields have the nearest fields to them having very close patterns.

Recall, Accuracy and F-measure for the book styles range from 0.91 to 0.99 for eight out of the ten styles. There was none with 1.0 as obtained for some journal styles because of the challenge highlighted earlier on. However, MISQ had 0.87 accuracy while IEEE had 0.80 because of the prominence of the separator issue and closeness of patterns of consecutive fields in the two styles.

On the average, the year information was extracted with the highest degree of precision, recall, accuracy and F-measure, while the other fields have relatively close precision, recall, accuracy and F-measure except for the title field with slightly lower values.

Some rule-based models such as those developed by Chowdhury[3] and Ding *et al.*[5] used a template mining approach to extract citations from digital documents. Ding *et al.*[5] used three templates to extract information from cited articles (citations) and obtained a satisfactory result (more than 90%) for the distribution of information extracted from each unit in the cited articles. The advantage of their rule-based model is its efficiency in extracting reference information. However, it only processed references from tagged text in one style. (e.g., references formatted in HTML). None of them considered references of different styles.

This research work may fairly be compared with the work of Day et al.[4]. They proposed a template-based Reference Metadata Extraction method for only journal references in different styles while this research includes book and other references. They implemented their system with reasonably high degree of accuracy in six out of the ten journal reference styles tested with the system developed in this research work. This reference metadata extractor, unlike in[4] indicates the style of reference, which may be an added advantage for researchers who want to validate the style of their manuscript references before sending them out for publication. The overall average field accuracy was 92.39% for the six major styles tested with their system while that of this research is 0.98 (98%). They reported 99.31% average accuracy for the MISQ style, while that obtained by this system is 100%. For the author field, the accuracy of their approach was 90.18% for six reference styles, while it is 98% for the ten styles tested by this system. The strength of the proposed approach is that it is very certain to extract accurately metadata for references in the tested reference styles and other related styles in different formats. Unlike machine learning approaches, there is no need for training. However, it requires a domain expert to design and maintain a number of templates, in case templates for more reference styles are to be built in the future.

CONCLUSION

Future research

This paper has presented a framework for automatic extraction of reference metadata from electronic documents of various formats and styles. The system consists of four major components: a means of providing reference input (by uploading the file or providing the set of references in the window provided by the browser), the text converter for converting documents into standard text format, the parser for automatically extracting metadata from the converted text using a set of predefined regular expressions, and the browser for displaying the results. The experimental results show that the system is very effective in carrying out the automated extraction process generally. There is higher degree of precision and accuracy in extracting metadata from journal references than book references because of the kind of separators found in book reference styles. The system extracts metadata from some styles of referencing not worked on by previous researchers. This facility can be used by authors to validate their references while preparing their manuscripts for publication. Its adoption for a previously unknown reference style only requires the addition of a new template for such. In future, the possibility of combining machine learning techniques with the adopted method will be looked into in order to produce a more generalized system.

REFERENCES

1. Besagni, D., Belaid, A., Benet, N. A segmentation method for bibliographic references by contextual tagging of fields, In Document Analysis and Recognition Proceedings **1**(3-6): 84-88 (2003).
2. K. Burnett, K.B. Ng, S. Park, A comparison of the two traditions of metadata development, *Journal of the American Society for Information Science* **50**(13): 1209-1217 (1999).
3. G. Chowdhury, Template Mining for Information Extraction from Digital Documents, *Library Trends* **48**(1): 182-208 (1999).
4. M.Y. Day, T.H. Tsai, C.L. Sung, C.W. Lee, S.H. Wu, C.S. Ong, W.L. Hsu, Reference Metadata Extraction Using a Hierarchical Knowledge Representation Framework, Institute of Information Science Academia Sinica Nankang Taipei 115 (2006).
5. Y. Ding, G. Chowdhury, S. Foo, Template Mining for the Extraction of Citation from Digital Documents, Proceedings of the Second Asian Digital Library Conference Taiwan 47-62 (1999).
6. C.L. Giles, K.D Bollacker, S. Lawrence, Digital Libraries and Autonomous Citation Indexing, Computer **32**(6): 67-71 (1999).
7. H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, E.A. Fox, Automatic Document Metadata Extraction Using Support Vector Machine, Proceedings of the Joint Conference on Digital Libraries (JCDL'03), Houston, Texas USA (2003).
8. Y. Hu, H. Li, Y. Cao, L Teng, D. Meyerzon, Q. Zheng,. Automatic extraction of titles from general documents using machine learning, *Information Processing and Management,* **42**: 1276-1293 (2006).
9. A. Kawtrakul, C. Yingsaeree, A Unified Framework for Automatic Metadata Extraction from Electronic Document (2005).
10. M. Krämer, H. Kaprykowsky, D. Keysers, T. Breuel, Bibliographic Metadata Extraction using Probabilistic finite state transducers (2007).
11. B.A. Ojokoh, S.O. Falaki, O.S. Adewale, Automated Information Extraction System for Heterogeneous Digital Library Documents. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007) Doctoral Consortium, Vancouver, British Columbia, Canada, 18-23 (2007).
12. B. Powley, R. Dale, Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification, Conference RIAO2007, Pittsburgh PA, U.S.A (2007).
13. K. Taghva, J. Coombs, R. Pereda, T. Nartker, Address Extraction using Hidden Markov Models, Information Science Research Institute, USA (2005).
14. A. Takasu, Bibliographic attribute extraction from erroneous references based on a statistical model, Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), 49-60 (2003).

[15. B. Wellner, A. McCallum, F. Peng, M. Hay, An integrated, conditional model of information extraction and coreference with application to citation matching. In UAI 2004:Proceedings of the 20th conference on uncertainty in artificial intelligence, 593-601 (2004).