2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies

Reference Metadata Extraction from Scientific Papers

Zhixin Guo, Hai Jin
Cluster and Grid Computing Lab
Services Computing Technology and System Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
guozhixin@mail.hut.edu.cn

***Abstract*—Bibliographical information of scientific papers is of great value since the Science Citation Index is introduced to measure research impact. Most scientific documents available on the web are unstructured or semi-structured, and the automatic reference metadata extraction process becomes an important task. This paper describes a framework for automatic reference metadata extraction from scientific papers. Our system can extract title, author, journal, volume, year, and page from scientific papers in PDF. We utilize a document metadata knowledge base to guide the reference metadata extraction process. The experiment results show that our system achieves a high accuracy.**

Keywords-metadata extraction; rule-based approach; reference

I. INTRODUCTION

Since the introduction of Science Citation Index, scientists use citations to measure the impact of scientific papers. Citations are also used to facilitate information search and retrieval in digital libraries. Unfortunately, citations are not structured data in scientific papers. A metadata extraction process is needed to get citations. So reference metadata extraction from scientific papers is essential for the integration of metadata from heterogeneous reference sources, where metadata is defined as structured data about data [1]. Here reference metadata refers to the sub-fields of references or citations.

Some large citation indices such as ISI use a manual information extraction process, which require human effort to tag information in citations. The process is time-consuming and expensive. In recent years, automatic reference metadata extraction becomes popular in large digital libraries such as CiteSeer [2, 3] and Google Scholar. The automatic extraction process is a challenging work due to diverse styles and variations in the use of field separators. For example, the author and title fields can be separated by spaces and periods, while the volume and issue fields can be separated by braces or parentheses [4]. Sometimes the full name and abbreviation of conference are both acceptable such as *Proceedings of IEEE International Conference on Advanced Learning Technologies* and *Proc. ICALT*.

In our early works, we proposed a metadata extraction framework for scientific papers [5, 6]. Both document header information such as title, author and abstract, and reference information such as title, author, journal, and volume can be extracted from scientific papers. In this paper an improved method for reference metadata extraction is introduced according to our early work. We use the document header information to assist the extraction of reference metadata.The framework is a part of SemreX system [7], which is P2P based semantic literature sharing system offering the literature sharing services among computer science researchers.

The rest of the paper is organized as follows. Section II overviews the related work about reference metadata extraction. The software architecture is presented in section III, where the design of the document metadata extraction layer of SemreX is presented in detail. The reference metadata extraction algorithm is described in section IV. The results of the experiment are presented in section V. Finally, we conclude the paper in section VI.

II. RELATED WORK

In general, the metadata of a scientific paper refers to the metadata from the paper header (the text before the *introduction* or the end of the first page) and the bibliographic fields [8]. Bibliographical information of scientific papers has great value to help scientists to find previous relevant works and measure their impact. It usually appears at the end of a paper and is stated as unstructured text strings. An automatic metadata extraction process is needed to get bibliographical information. It is a subtask of information extraction that automatically segments unstructured text strings into structured records. This procedure is necessary in order to import the information contained in legacy and text collections into a data warehouse for subsequent querying, analysis, mining and integration [9].

There are two kinds of methods for citation extraction, which are machine learning approach and rule-base approach [4]. In machine learning approaches, the extraction process uses a dataset of the input and output samples to be trained to discover their relationship, and then predicts new data. Approaches like CiteSeer [3] take advantage of probabilistic estimation, based on training sets of tagged bibliographic data. Seymore et al. used *hidden Markov models* (HMM) for the document header metadata extraction [10]. HMM learns

978-0-7695-4564-6/11 $26.00 © 2011 IEEE
DOI 10.1109/PDCAT.2011.72

a generative model over input sequence and labeled sequence pairs. While enjoying wide historical success, standard HMM models have difficulty modeling multiple non-independent features of the observation sequence.

Based on discriminatively trained SVM classifiers, Han et al. presented another method to extract the header metadata [8]. In the paper [11], Peng et al. employed Conditional Random Fields for the task of extracting various common fields from the header and citation of research papers. Those methods are all based on machine learning technique. In general machine learning techniques are robust and adaptable and can be used on any document set. However, generating the labeled training data is time-consuming and costly.

Other researchers use the rule-based method for automatic metadata extraction. Kawtrakul and Yingsaeree presented a framework for automatic metadata extraction from electronic documents that can be both text documents and images of paper documents to ease metadata creation process [12]. Han et al. introduced a domain rule-based word clustering method for cluster feature representation to improve the classification performance of document header lines and bibliographic fields [13]. They use a template mining approach based on pattern recognition and pattern matching in natural language texts to extract different kinds of information from digital documents. In contrast to machine learning approaches, rule-based approaches use a set of rules that define how to extract metadata based on human observation [4]. They do not need any training process and can be implemented in a straightforward manner. The main disadvantage of rule-based approaches is their lack of adaptability, and they need domain experts to design relevant rules.

In this paper, we propose a rule-base approach to automatic extract reference metadata. Unlike the system list above, we utilize a document metadata knowledge base to assist the extraction of reference metadata. The extracted reference metadata will be searched in the knowledge base. If the metadata are founded in the knowledge base and there are some errors in the metadata, the extracted metadata will be corrected using the information from the knowledge base. In this way, we can achieve a better performance for reference metadata extraction.

III. SOFTWARE ARCHITECTURE

The approach proposed in this paper is a part of the system named SemreX. SemreX is a software system, including document metadata extraction, semantic classification of papers, semantic representation, syntax and semantic query, and P2P communications. Fig. 1 shows the system architecture of SemreX.

Fig. 2 shows the logical flow of document metadata extraction in SemreX. Since a large amount of electronic documents especially scientific papers are stored in PDF files, we start our study from those files. The crawlers search scientific papers on the web and when a crawler find a paper seems to be a scientific paper in PDF, it records the URL information of the paper and downloads the paper immediately. The File Conversion Module converts the file from PDF to other format. The File Identification Module uses the converted text file to estimate whether this paper is a scientific paper. If it is not, the system ignores it; otherwise the system starts the Metadata Extraction Module to extract the document metadata include header metadata and reference metadata of this paper, saves the extracted metadata in the format of OWL, and stores the information in the ontology database to assist further semantic search.

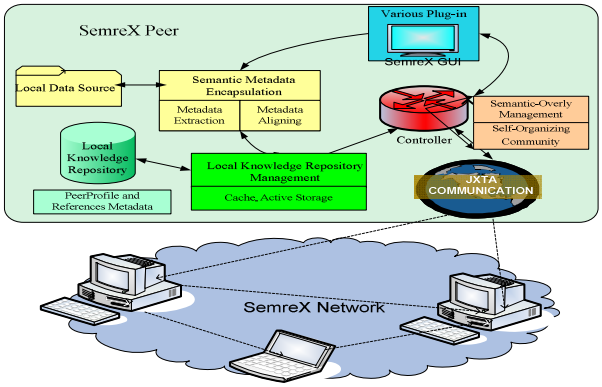

Figure 1. Architecture of SemreX

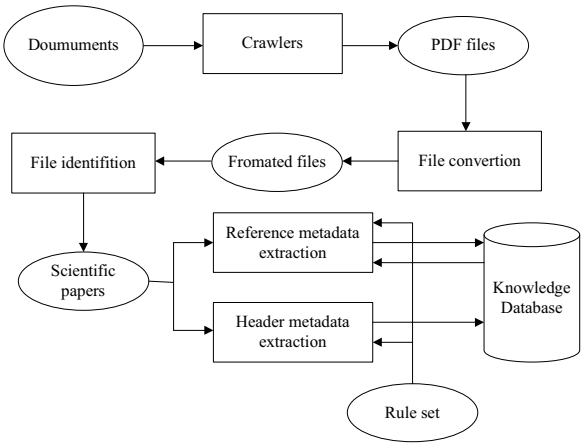

Figure 2. Logical flow of document metadata extraction

A. File Conversion

In order to process the documents and find the elements our system needs to index them and continue scanning the web. Given the almost ubiquitous nature of Adobe's Portable Document Format on the web we start our study from those scientific papers in this format.

Before extraction, we need convert those PDF files to other files in different format which can be easier to deal with. PDF files make use of inner objects to describe information such as text, image, and table. The Adobe Acrobat SDK is released for the developers to create software and plug-ins to interact and customize Acrobat. Those software libraries and components can help researchers to read and covert PDF files. There are a lot of converting applications which have been released to handle the PDF files. According to our system, we try to convert the

source PDF files to two kinds of files which are text files and XML files. The text files contain all the text extracted from the source files and they do not carry any format information. On the other hand, the XML files have the format information about the characters and paragraphs in the papers, such as the font size and the layout. We hope with the help of format information, the extracted metadata could be more accurate. Fig. 3 shows an example of a converted XML file.

```
<Page number="1" height="792.8394"
width="613.4392">
<Options> granularity=word </Options>
<Structure>
<Font name="TimesNewRomanPSMT" size="23.1" />
<Word box="[99.6 666.8 156.9863 689.9]"
textrendering="3">
<Text> Policy </Text>
</Word>
<Word box="[162.7994 666.8 262.0251 689.9]"
textrendering="3">
<Text> Evaluation </Text>
</Word>
<Word box="[266.6 666.8 293.4159 689.8]"
textrendering="3">
<Text> for </Text>
</Word>
```

Figure 3. An example of converted XML file

We can acquire the following important information from the source PDF file:

1) Position: the position of each word described as box="[X1 Y1 X2 Y2]" where X means x-coordinate and Y means y-coordinate.
2) Content: all extracted text of the PDF file.
3) Font type: the font type of each word.
4) Font size: the font size of each word.
5) Page: the page number that the word locates at.

Those characters can help us to identify whether this paper is a scientific paper and extract the header metadata and bibliographies more accurately.

B. File Identification

Our system tries to extract metadata from scientific papers in PDF. Since those PDF files are downloaded from the web, it is essential for us to identify whether those documents are scientific papers. Scientific papers are usually regular papers published in journals or proceedings. Although papers in different journals or proceedings have different format requirements, they may have some common characters in the format. We define several rules to identify the scientific papers. Those rules based on the analysis of the converted text files from sources PDF files are summarized as follows:

1) If the number of lines in the text file is less than 10, the origin PDF file is not a scientific paper.
2) If there are words like *IEEE*, *ACM* or *PROCEEDINGS* in the first page, it should be a scientific paper.
3) If there are words like abstract or introduction in the first two pages, or references in the last two pages, it should be a scientific paper.

The first rule is to prevent the cases that some papers in PDF are made up of pictures. When those papers are put to the file conversion module, the text files we get only contain few words although they may really be the scientific papers. The other rules use some keywords to identify scientific papers. Although those rules are very simple, we find they are useful to identify the scientific papers in the experiments.

C. Metadata Extraction

When a paper is identified to be a scientific paper, it will be sent to the metadata extraction module. The document metadata including the header information and bibliographies will be extracted. The header information includes title, authors, abstract and pages. They are stored in the knowledge base. Reference metadata are extracted using a rule-base approach and are checked if there are any matched document in the knowledge base. If matched, the extraction process uses the metadata stored in the knowledge base to make sure the extracted reference metadata are right.

IV. Reference Metadata Extraction Algorithms

There are many reference styles in scientific papers. Each style has its own format to represent the citation. For example, the reference formatted in the IEEE style looks like:

S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," IEEE Computer, Vol.32, No.6, 1999, pp.67-71.

The reference formatted in the ACM style is as:

Lawrence, S., Giles, C. L., and Bollacker, K. 1999. Digital libraries and autonomous citation indexing, IEEE Computer, 32(6), 67-71.

Or it could be formatted in the APA style as:

Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing, IEEE Computer, 32(6), 67-71.

Even in different journals or proceedings sponsored by IEEE, the reference style has some differences. For example, in some IEEE styles, the year appears after the volume and number, but in other IEEE styles, the year appears just after the name of the journal.

It is a challenging work to automatic extract reference metadata as author, title, page, date, et al from so many different formatted references. For example, the punctuation of dot is usually used to separate the author name and title, but it also appears in the name of author to separate the family name and others. Some reference styles use dot to separate the author and title, others use comma to separate them. Although there are a lot of reference styles which make the extraction more difficult, there are also some rules in the styles. Some of the rules are listed as follows:

1) In the field of the author name of a citation, each author's name may have a different format. Some use the full name, and some only use the last name. For the same person whose name is Steve Lawrence, the appearance of the author name of a citation could be *Steve Lawrence*, *Lawrence, S.*, *S. Lawrence*, and *Lawrence*. If there are two or more than two authors, the separator before the last author's name may be "*, and*". If there are three or more than three authors, the citation

may list all authors, separate them with a comma, or only list the first several authors and abbreviate the remainder as "*, et al.*"

2) The field of journal or conference name of a citation usually appears after the title of a paper. If the paper is published in a conference, the separator before the conference name should be *In the proceedings of* and *Proc.*, or the abbreviated name is used such as *SIGMOD'09* or *WWW'09*.
3) The field of date of a citation may include year, season, or month and date. But usually it only presents the year which format may use 4-digits like 2011. The separator before and after the year can be comma, or parenthesis like *(2011)*.
4) The title of a citation usually appears after the author name. The format of title could be in headline-style capitalization or in sentence-style capitalization. Sometimes the separator before and after the title can be double quotation marks.
5) The volume and issue are parts of citation published in a journal. They are usually listed together. The volume uses the word *vol.* as prefix, and the issue uses the word *no.* as prefix. Sometimes there are no prefixes and the volume and issue are formatted using commas like *32(2)*, where the volume is 32, and the issue is 2.
6) The page field of a citation usually appears at the end of a citation. It may have a prefix like *pp.*, or it may just be formatted as numbers. There are several major styles for formatting page numbers: only the first page; the range of pages; or abbreviate the last page.

After we study the rules of a citation, we find that a template-based approach should be proper for reference metadata extraction. For example, the ACM style of the paper *Digital libraries and autonomous citation indexing* shown in the beginning of section IV, can be presented using the following template:

AUTHORS._DATE_._TITLE_._JOURNAL_,_VOLUME_(_ISSUE_),_PAGES_.

If the text of a citation match the template, and follow some rules listed above, we can extract properly the reference metadata such as title, journal or conference, publisher, date and page. We use a template base to describe the different reference styles. Table I shows the template elements definition and their meanings.

TABLE I. TEMPLATE ELEMENTS DEFINITION

Name	Meanings
AUTHORS	Authors' name
TITLE	Title
JOURNAL	Name of journal
PROCEEDINGS	Name of proceeding
SCHOOL	Dissertation school
PAGES	Page
DATE	Date
VOLUME	Volume
ISSUE	Issue
URL	Web page
PUBLISHER	Publisher
PUBLISHERLOC	Address
EDITOR	Editor
ANY	Arbitrary string

There are two problems to be solved before we extract reference metadata from a citation using a template-based approach. One is where we can get those templates, and the other is how to match the template. We analyze the reference styles in digital libraries, define and store the templates in a template base. Totally we have defined about 576 templates for different reference styles. Furthermore, we build the knowledge base about the person name, full or abbreviate name of journals and conferences. When the strings of a citation are sent to the metadata extraction process, the system first checks the words using the knowledge base and finds what they could belong to, the author name, title, journal or conference name, or date. Then it searches the template base, finds the most similar template to match the strings, and uses the template to extract the reference metadata. The template matching algorithm can be found [5].

When the reference metadata are extracted from a citation, the system searches the knowledge base about document metadata which is built before the reference metadata extract process and tries to match a document to the citation. If any document is matched, it replaces the extracted reference metadata using the metadata in the knowledge base because the metadata in the knowledge base are more accurate. The whole algorithm can be described as follows:

Step 1. Get the PDF file from the web and convert it to other formats to get text and layout information.

Step 2. Identify whether the file is a scientific paper. If not, finish the process.

Step 3. Find the location of the references in the files using the keyword *references*, *citation* or *work cited.*

Step 4. Tokenize each citation string and separate each citation from others.

Step 5. For each citation, extract citation components such as title, authors, date, and pages, using the template based algorithm.

Step 6. Apply direct match between the extracted title, authors in step5 and the title, authors stored in the knowledge base. If extract match is found, use the information in the knowledge base to correct the extracted metadata.

Step 7. Store the reference metadata of each citation in the database for further search.

V. EXPERIMENTAL RESULTS

We test our system over a set of 100 scientific papers in computer science domain downloaded from digital libraries of IEEE and ACM. Those papers are all journal or conference papers. The test papers are all in PDF. Three of them are discarded because they could not be converted from PDF files to text files and xml files. In the rest 97 papers there are totally 2157 citations needed to be extracted.

We use our system to extract metadata such as title, authors, journal, proceedings, date, and pages. For each citation in our sample we test whether the automatically extracted metadata are correct. Such test is based on a manual verification of the correctness of the extracted metadata against the original paper. Because this process is tedious and boring, we have to limit our sample size. The knowledge base we use to assist the reference metadata extraction has stored about 1.2 million document metadata.

The metadata are all verified to be correct using the bibliographic records downloaded from DBLP. We compare the results using or not using the knowledge base. Table II shows the accuracy results of our test.

TABLE II. METADATA EXTRACTION ACCURACY

Metadata	Accuracy	
	Not Using Knowledge Base (%)	*Using Knowledge Base (%)*
Title	88.1	88.1
Authors	91.5	94.7
Journal	83.5	90.2
Proceedings	84.6	91.8
Volume	76.7	82.4
Issue	73.2	86.1
Date	86.7	91.2
Pages	75.6	83.8
URL	93.3	93.3

From the results we observe that the reference metadata extract accuracy has been improved a lot by using the knowledge base. The accuracy of URL is the highest because it always has the prefix "http://". Its accuracy can not be improved using the knowledge base since the knowledge base does not store the metadata of such kind of papers. The accuracy of authors is the second highest because the author name usually appears in the front of a reference. There are usually two kinds of mistakes to extract the authors' metadata. One is in the case the reference does not contain any author information. The other type of author error is caused by the ambiguity of the author name boundary, for example, if the title is too short, sometimes the system mistakenly extracts the author name together with the title. The accuracy of pages, volume and issue is lower than the others. An error could be caused by missing page information. For example, in the following reference entry:

R. R. Bouckaert, Low level information extraction: a Bayesian network based approach, Workshop on Text Learning (TextML 2002), 2002.

The system pareses the last word "2002" to be the page information, but it should be the date information. The accuracy of volume is the lowest among those reference metadata. The errors in the volume information could be caused by missing volume information in the reference or the ambiguity of boundary. For example, in the following reference entry:

A. Sen, Metadata management: past, present an future, Decision Support Systems 37(1) 2004 151-173

The system mistakenly extracts "Decision Support Systems 37" as the journal name, where the word "37" should be the volume information.

VI. CONCLUSIONS

In this paper, we propose a rule-based reference metadata extraction system for scientific papers. This system, which is a part of the project SemreX, can retrieve reference information on the title, authors, journal, proceedings, volume, issue, date, pages and URL with good accuracy. The metadata extracted by our system can help people find the relationship between authors and the impact of a paper. Unlike machine learning approaches, our approach does not need training. However, it requires a domain expert to design and maintain a number of templates. Another problem about our system is that it often fails to extract metadata from papers that have somewhat complicated structures. We are now trying to revise our system and making it more robust to retrieve desired information successfully. A wide variety of documents types can also be considered in addition to the PDF papers. The system can be extended to handle scientific papers in different formats such as Postscript or HTML. Those are our avenues for ongoing and future research.

REFERENCES

[1] A. Sen, "Metadata management:past, present and future," *Decision Sypport Systems*, 2004, Vol.31, No.1, pp.151-173.

[2] S. Lawrence, C. L. Giles, and K. Bollacker , "Digital libraries and autonomous citation indexing," *IEEE Computer*, Vol.32, No.6, 1999, pp.67-71.

[3] H. Li, I. Councill, W. Lee, and C. L. Giles, "CiteSeerx: an architecture and web service design for an academic document search engine," *Proceedings of the 15th international conference on World Wide Web*, 2006, pp.883-884.

[4] M-Y. Day, R. T-H. Tsai, C-L. Sung, C-C. Hsieh, C-W. Lee, S-H. Wu, K-P. Wu, C-S. Ong, and W-L. Hsu, "Reference metadata extraction using a hierarchical knowledge representation framework," *Decision Support Systems*, Vol.41, No.1, pp.152-167.

[5] Z. Guo, H. Jin, and H. Chen, "Semantic document reference metadata extraction in SemreX," *Journal of Computer Research and Development*, 2006, Vol.43, No.8, pp.1368-1374 (in Chinese).

[6] Z. Huang, H. Jin, P. Yuan, and Z. Han, "Header metadata extraction from semi-structured documents using template matching," *Proceedings of* I*FIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS'06)* - Lecture Notes in Computer Science, Vol.4278, Springer-Verlag, 2006, pp.1776-1785.

[7] H. Jin and H. Chen, "SemreX: efficient search in semantic overlay for literature retrieval," *Future Generation Computer Systems*, 2008, Vol.24, No.6, pp.475-488.

[8] H. Han, C. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. Fox, "Automatic document metadata extraction using support vector machines," *Proceedings of Joint Conference on Digital Libraries*, 2003, pp.37-48.

[9] E. Agichtein and V. Ganti, "Mining reference tables for automatic text segmentation," *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp.20-29.

[10] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden markov model structure for information extraction," *Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction*, 1999, pp.37-42.

[11] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," *Proceedings of Human Language Technology Conference*, 2004, pp.329-336.

[12] A.Kawtrakul and C. Yingsaeree, "A unified framework for automatic metadata extraction from electronic document," *Proceedings of the International Advanced Digital Library Conference*, 2005, pp.71-77.

[13] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles, and X. Zhang, "Rule-based word clustering for document metadata extraction," *Proceedings of ACM Symposium on Applied Computing*, 2005, pp.1049-1052.