IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 1 (Jul. - Aug. 2013), PP 12-16
www.iosrjournals.org

An effective citation metadata extraction process based on BibPro parser

G. Guru Brahmam[1], A. Bhanu Prasad[2]
[1](M.Tech, Software Engineering, Vardhaman College of Engineering/ JNTU-Hyderabad, India)
[2](Associate Professor, Department of IT, Vardhaman College of Engineering/ JNTU-Hyderabad, India)

Abstract: *There is a dramatic increase in academic publications and these publications are integrated to the digital libraries by making use of citation string. Any author can publish the journals conferences with the help of his own citation style. There is no specific format to the conferences and journals citation styles that are publishing on digital libraries. As there is no specific format to the citations it is difficult to any author or researchers when he want to perform field based searching on digital libraries. So it is an interesting problem to extract the components of citations string which is formatted in one of thousand different citation styles. The proposed citation parser named BibPro extracts components of citations strings more accurately with that of existing systems and achieves reasonable performance.*
Keywords: *Data integration, digital libraries, information extraction, sequence alignment.*

I. Introduction

Citations and its formats play vital role in scientific publication digital libraries (DLs), such as CiteSeer, arXiv e-print, DBLP, and Google Scholar. Often researchers use citations to find information related to their articles. The citations used as auxiliary support in information extraction tasks, e.g., automatic document classification [2],[3], indexing and ranking[10], and quality assessment[4]. Most of the Citations contain common fields like authors' name, title, publication venue, date and the number of pages. Also most of the citation management techniques used these common fields as key assumption during citation field extraction.

There are many reasons which make citation metadata extraction processes difficult. The citation used in the extraction is collected from internet and there is no specific style to the citations. Data collected from internet is difficult to process because the data may contain inconsistent data and error.

We propose a sequence-alignment-based citation parser. The citation parser named as BibPro to extract the components of citation which is formatted in one of thousand styles. The technique in this BibPro parser is first capture the structural properties from semi-structured format then transform these into sequenced template.

Order of punctuation marks and local structure in each field are included in the structural properties. During the parsing encoding tables and reserved words concept are used to rep1resent each semantic unit as unique symbol and blocking pattern concept is used to capture the local structure in each field. After implementing the sequenced templates the citation parser BibPro applies alignment techniques to match the query citation string with sequenced templates.

Using the sequenced templates the citation parser Bibpro get the advantages like reduced complexity of citation string, easy transforming of structural properties into sequence. We can use sequence techniques such as search, matching and alignment to compare structural properties.

II. Related Work

In recent years, several techniques are proposed to deal with the citation metadata extraction problem effectively. A brief method survey is given in [6]. They include HTML structural analysis, natural language processing, machine learning techniques and data modeling techniques. The survey in [7],[8,][9],[10],[15],[18],[21], is similar to citation metadata extraction. The approaches in the survey are classified into two categories: learning based approach and knowledge approach.

The learning based approach transforms the citation metadata extraction problem into classification problem and applies machine learning technique to solve it. It is found that currently there are three major machine learning techniques, the Hidden Markov model (HMM) [17],[18],[19],[20],[21], support vector machines (SVM) [16], and conditional Random Fields (CRF) [15],[29], are used during the metadata extraction. Knowledge based approaches are use domain knowledge to describe the interested data, where knowledge include relationships, lexical appearances and context keywords. Citeseer is well known search engine and digital library that uses heuristics to extract components of citation string.

III. Problem Definition

We refer to a citation string as textual string which is used to present metadata of an journal or conference publication in specific formatting style. The typical fields of metadata include author, title and publication information. These fields are separated by punctuation marks. The citation metadata extraction problem defined as follows: given a semi-structured citation string S, S = {Field1 delimiter1 field2 delimiter2 field3 delimiter3 field4 delimiter4 ….}, here fieldi are fields in metadata and delimitei are symbols to separate fields. Most of the citation metadata extraction is interested in extraction of most common fields like title, author name, venue, volume, issue, page and date.

IV. Bibpro Parser Workflow

The BibPro parser is mainly divided into two modules; one is template database construction and query processing module shown in fig.1. In template data base construction module set of citations and associated metadata is given as input to canonicalization algorithm. During this process the structural features like INDEX FORM and STYLE FORM, of citation is stored. A STYLE FORM is a symbolic representation of a citation string in which each field of metadata as well as each punctuation mark is represented by a single symbol.

INDEX FORM is canonical symbolic representation of a citation string. This canonical symbolic representation will be used later to align with a given query citation. Next the query INDEX FORM is used to search the template data bases for similar INDEXES to that of the given query citation string. Then we make use these detail alignment information to give the final output.

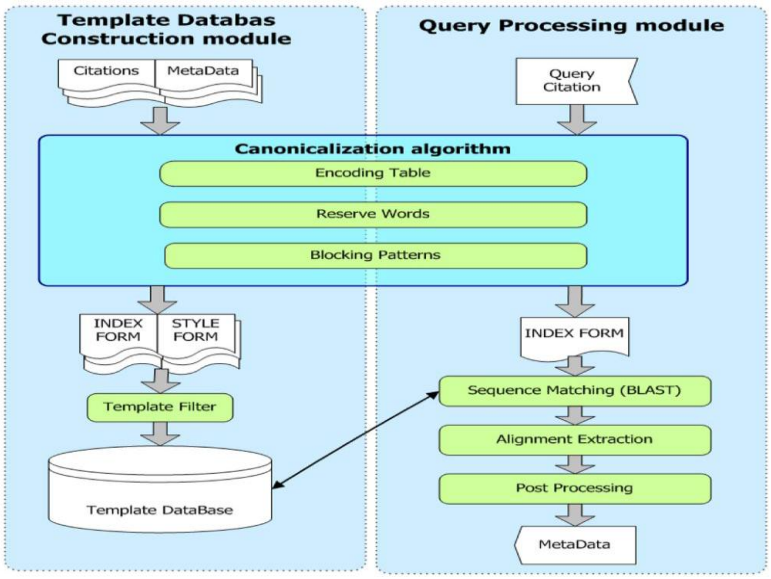

Fig. 1

Canonicalization Algorithm: Normal regular rules are performing the parsing of the data**.** Using rules recognize the features of data. Common rules are not extract efficient results. All levels are contains noisy results. Removing the noisy results apply the canonical strings format under extraction of data. Canonical string extracts the results with structural properties, without redundant and noisy results specification process. All kinds of components results are display as a protein sequence result at a final stage of result. It is display the result as a systematic content.

Encoding Table: It can perform the operation in token encoding and give the mapping for generation of protein sequence. It can perform the operations based on Meta data environment process. In mapping environment perform the operation like transformation technique. It can perform the operation like symbol alignment procedure. In each and every citation string alignment process works on corresponding symbol. In the encoding process 23 symbols are used to replace fields with appropriate symbols. With these symbols fields are replaced like Author with ‘A’, Title with ‘T’, venue with ‘L’, Volume with ‘V’ and etc. The encoding table in the canonicalization algorithm is shown in Table 1.

Table 1. Encoding Table

Category	Symbol	Representation Field
Field Token	A	Author
	T	Title
	L	Venue (Journal, BookTitle, Technical Report)
	V	Volume
	W	Issue
	P	Page
	Y	Date (Year Month)
	F	Editor
	S	Institution
	M	Publisher
	X	Unknown Single Token
	B	Unknown Continuous Token
	N	Numeral Token
Delimiter Token	R	,
	D	.
	G	"
	E	'
	C	:
	Z	;
	H	-
	I	([< {
	K	)] > }
	Q	/ _ ! @ # $ % ^ & * + = \ | ? ◦ ~

Reserved Words: One term related that field automatically applies directly as reserved words contents. We use the reserved word to recognize the token belongs to which field. It can perform the operation like automatically and internally. Any words are occurring frequently in particular environment also apply the reserved words technique. We are applying the few reserved words in particular field. The reserved word used the canonicalization algorithm is shown in Table 2.

Example: Page field in PP, Volume field place the vol etc.

As we do not know the metadata in the encoding process of citation string use reserved words to recognize the tokens. Also not all fields have explicit reserved. Hence we use to term frequency concept to replace the field with appropriate reserved word. The term frequency concept is the words which have higher term frequency in a specific field than in other fields should be treated as reserved words for that field and the words that occurs that occurs frequently in all fields are avoided. Next we compute weight for the each term in citation string. Based on computed weight common fields are filtered out.

Blocking Patterns: Reserved words followed punctuation marks are present, and then total content treats as a local structure. It can follow the dependency procedure in implementation. We have some specific patterns in implementation. Each and every pattern expresses the sum of common rules. After utilization all patterns it's possible to generate final structure in output environment. Final output displays as a global structure environment. It is one good filtering procedure. Blocking Patterns uses regular expressions to identify the dependency between the reserved words and punctuation marks. The regular expressions to author, venue, volume, page, issue are shown Table 3.

Template database Construction and query processing: All citations components display as a template. In template total components are displayed here that is called as an index form. Total components of citations data display metadata template. It can provide the result as a one style form in specification. Here we are uses the transform technique under template content display here. Whenever any user forward any kind of query directly display the result as a result form.

Table 2. Reserved words in BibPro parser

Field	Encoded Knowledge or Keyword
Author	Name abbreviation
Journal	"Transactions", "Trans", "Journal"
Booktitle (Conference)	"Proceedings", "Proc", "Workshop", "Conf", "Conference", "Symposium", "Sympos", "Symp", "International", "Intern", "Annual", "Annu"
Technical Report	"Tech", "rep", "Rpt", "TR", "Master", "Masters", "Ph", "PhD", "Thesis", "thesis", "Dissertation", "dissertation"
Volume	"Volume", "volume", "Vol", "vol", "Vo", "vo"
Issue (Number)	"Number", "number", "Nr", "nr", "No", "no", "NO", "Nos"
Page	"pp", "page", "pages", "PP", "Page", "Pages", "pg", "PG"
Month	"January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December", "Jan", "Feb", "Mar", "Apr", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec", "Sept"
Year	1900-2010
Editor	"eds", "Eds", "editors", "Editors", "editor", "Eds", "ED", "Ed", "ed", "edited"
Institution	"University", "Univ", "Department", "Dept", "Corporation"
Publisher	"Press", "Pub", "Publishers", "Inc", "Publications"

Table 3. Regular Expressions

Field	Regular Expression
Author	A[^LVWPYSF]+A A[^LVWPYSF]*[BX] [BX][^LVWPYSF]*A
Venue	L[^AVWPYSF]+L L[^AVWPYSF]*[BX] [BX][^AVWPYSF]*L
Volume	V[^ALWPYSF]*N
Page	P[^ALVWYSF]*N
Issue	W[^ALVPYSF]*N

V. Conclusion

Still parsing citations is a challenging problem due to the diverse nature of citation formats. We proposed a template based citation parsing system called BibPro, which extends our previous work by using the order of punctuation marks in a citation string to represent its format. When parsing citation string through online, bibpro transforms the citation string into a protein sequence and applies sequence alignment techniques to find out the most similar template for extraction of metadata from the citation. According to our experiments, bibpro performs well and is scalable.

References

[1] D. Lee, J. Kang, P. Mitra, C.L. Giles, and B.-W. On, "Are Your Citations Clean?," Comm. ACM, vol. 50, pp. 33-38, 2007.

[2] M. Cristo, P. Calado, M.A. Goncalves, E.S. de Moura, B. Ribeiro-Neto, and N. Ziviani, "Link-Based Similarity Measures for the Classification of Web Documents," J. Am. Soc. for Information Science and Technology, vol. 57, pp. 208-221, 2006.

[3] T. Couto, M. Cristo, M.A. Goncalves, P. Calado, N. Ziviani, E. Moura, and B. Ribeiro-Neto, "A Comparative Study of Citations and Links in Document Classification," Proc. Sixth ACM/IEEE-CS Joint Conf. Digital Libraries, 2006.

[4] M.A. Goncalves, B.L. Moreira, E.A. Fox, and L.T. Watson, "'What Is a Good Digital Library?' - A Quality Model for Digital Libraries," Information Processing and Management, vol. 43, pp. 1416-1437, 2007.

[5] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Proc. Seventh Int'l Conf. World Wide Web 7,1998.

[6] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira, "A Brief Survey of Web Data Extraction Tools," SIGMOD Record, vol. 31, pp. 84-93, 2002.

[7] C.L. Giles, K. Bollacker, and S. Lawrence, "CiteSeer: An Automatic Citation Indexing System," DL '98: Proc. Third ACM Conf. Digital Libraries, pp. 89-98, 1998.

[8] K.D. Bollacker, S. Lawrence, and C.L. Giles, "CiteSeer: An Autonous Web Agent for Automatic Retrieval and Identification of Interesting Publications," Proc. Second Int'l Conf. Autonomous Agents, 1998.

[9] S. Lawrence, C.L. Giles, and K.D. Bollacker, "Autonomous Citation Matching," Proc. Third Ann. Conf. Autonomous Agents, 1999.

[10] S. Lawrence, C.L. Giles, and K.D. Bollacker, "Digital Libraries and Autonomous Citation Indexing," Computer, vol. 32, no. 6, pp. 67- 71, June 1999.

[11] 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. 43, pp. 152-167, 2007.

[12] E. Agichtein and V. Ganti, "Mining Reference Tables for Automatic Text Segmentation," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2004.

[13] E. Cortez, A.S. da Silva, M.A. Gon calves, F. Mesquita, and E.S. de Moura, "FLUX-CiM: Flexible Unsupervised Extraction of Citation Metadata," Proc. Seventh ACM/IEEE-CS Joint Conf. Digital Libraries, pp. 215-224, 2007.

[14] Andrew McCallum's Code and Data, http://www.cs.umass.edu/ ~mccallum/code-data.html, 2005.

[15] F. Peng and A. McCallum, "Accurate Information Extraction from Research Papers Using Conditional Random Fields," Proc. Human Language Technology Conf. and North Am. Chapter of the Assoc. for Computational Linguistics (HLT-NAACL), pp. 329-336, 2004.

[16] H. Han, C.L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E.A. Fox, "Automatic Document Metadata Extraction Using Support Vector Machines," Proc. Third ACM/IEEE-CS Joint Conf. Digital Libraries, 2003.

[17] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction," Proc. Workshop Machine Learning for Information Extraction (AAAI '99), pp. 37- 42, 1999.

[18] V. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic Segmentation of Text into Structured Records," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2001.

[19] A. Takasu, "Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model," Proc. Third ACM/IEEECS Joint Conf. Digital Libraries, 2003. [20] P. Yin, M. Zhang, Z. Deng, and D. Yang, "Metadata Extraction from Bibliographies Using Bigram HMM," Proc. Seventh Int'l Conf. Asian Digital Libraries, pp. 310-319, 2004.

[21] E. Hetzner, "A Simple Method for Citation Metadata Extraction Using Hidden Markov Models," Proc. Eighth ACM/IEEE-CS Joint Conf. Digital Libraries, 2008.

[22] I.-A. Huang, J.-M. Ho, H.-Y. Kao, and W.-C. Lin, "Extracting Citation Metadata from Online Publication Lists Using BLAST," Proc. Eighth Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD '04), Jan. 2004.