An Overview on Implementation of Citation Parser using Knowledge Base Techniques

Anand V. Saurkar[1], **A. R. Itkikar**[2]

[1]DMIETR, Sawangi (M), Wardha (MH), India

[2]Sipna COET, Amravati (MH), India

Abstract: ***Use of the bibliographical information of publications available on the Internet is an important task in academic research. Accurate reference metadata extraction for publications is essential for the integration of information from heterogeneous reference sources. It is an essential task in the research paper development to point out references for proper document. Giving a proper acknowledgement to a document or part of document is called citation. A "citation" is the way to tell readers the source from which certain material has come. In general citation implies a relation between part or whole of the cited document and a part or whole of the citing document. To cite a particular document there are several methods. We shall develop a citation tool based on knowledge based citation parsing technique. By developing a current system we shall extract the cited document which is useful for scientific publication or document citing.***

Keywords: Citation, Parser, Text search, knowledge Base.

1. Introduction

In the publication or academic communities, citations are a key part of linking distinct pieces of knowledge into a well-structured record of a field. Although citations can come in many forms, a common standard places a reference to each paper, article or book used by the authors, in a separate section of the work [1]. This list, often known as a bibliography, acts as an acknowledgement to these materials and provides exact publication information to identify each source and allow a reader to locate it for further study. Users often use citations to find information of interest in Digital Libraries, while researchers depend on citations to determine the impact of a particular article. Parsing citations is essential for integrating bibliographical information published on the Internet. Parsing citations is essential for integrating bibliographical information published on the Internet. Most citation management techniques are based on the assumption that we can correctly identify the main components of a citation, such as authors' names, title, publication venue, date, and the number of pages. However, for a variety of reasons, it is difficult to design a parser that can automatically parse citations scattered over the Internet [1]. Potential problems include data entry errors, diverse citation formats, the lack of (enforcement of) a standard, imperfect citation gathering software, common author names, abbreviations of publication venues, and large-scale citation data. We propose a knowledge-based citation parser, to extract components of citations in any given formats. The basic idea of this citation parser is to capture the structural properties from semi structured format and transform these properties into a sequence template. The structural properties of a citation string include the order of punctuation marks and local structure in each field of a citation string. We use an encoding table and reserved words, which is automatically trained from the data set, to represent each semantic unit as a unique symbol; and use a blocking process to capture local structure in each citation field. There are various methods like machine learning technique and knowledge based technique to cite a document. We shall propose a system which is based on knowledge based hierarchy technique.

The problem of citation parsing has been the focus of past research initiatives, as documented in the literature [2]. Existing citation parsers can be generally divided into two categories: template matching and machine learning based approaches. A template matching approach takes an input citation and matches its syntactic pattern against known templates. The template with the best fit to the input is then used to label the citation's tokens as fields. The canonical example of a template based approach is Para Tools. Disadvantage of Para Tool is on the basis of available template it generate result [1].

This technique works fairly well for citations which adhere to simple citation patterns, but is susceptible to errors when it tries to extract fields from citations with much punctuation, since there may be multiple templates that fit equally well. If the wrong template is chosen, entire fields will be tagged incorrectly. The limitations of the template-based approach have encouraged researchers to try alternative models for citation parsing.

Hidden Markov Models (HMMs) are a powerful probabilistic tool and has been applied extensively on various language related tasks [2] [4]. HMMs are a finite state automaton with stochastic state transitions and symbol emissions. HMMs may be used for citation parsing by formulating a model in the following way: each state is associated with a citation field name (hereby called "tag") such as title, author or date. A labeled training dataset is first used to train a HMM. This model is then used to recover the most-likely state sequence that produces the sequence of observation symbols. The HMM method is less effective due to its assumptions of independent and non-overlapping features [6].

Conditional Random Fields (CRF) model are used to parse citations in another experiment. CRFs are undirected graphical models trained to maximize a conditional probability and have been applied on tasks such as name entity extraction. Training time is a concern, as CRFs converge slowly. It requires approximately 500 iterations for the model based on the same training set to stabilize [5].

The Maximum Entropy Model provides flexibility given sufficient training datasets and serves as a balance between the two mentioned machine learning models. Since then, maximum entropy techniques have widely used for natural language tasks such as identifying sentence boundaries and text classification. Disadvantage of this system is additional supporting databases such as a journal name database, publisher database, country name database, etc. can also be collected to help the system identify and extract fields [2].

2. Analysis of Problem

The integration of bibliographical information of scholarly publications available on the Internet is an important task in academic research. Accurate reference metadata extraction for scholarly publications is essential for the integration of information from heterogeneous reference sources.

Automatic citation extraction is difficult due to variations between field separators. For example, 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.

3. Proposed Work and Objectives

To solve above problem regarding automatic citation extraction I will proposed a system is based on knowledge based hierarchy technique. A feature of this system is its capability to represent and match complicated structures, such as hierarchical matching, regular expressions, semantic matching. Using this system, we will extract author, title, journal, volume, number (issue), year, and page information from different kinds of reference. A hierarchical approach is use to detect source of information from source data base. Knowledge data base is a combination of all data available from a particular domain set. I will apply a citation technique on the same. Objective of this proposed system collection of reference data, save it and apply parsing on that data to cite a particular inputted string.

4. Conclusion

Parsing citations is challenging due to the diverse nature of citation formats. In this report, we present an implementation of citation parser on basis of knowledge based hierarchy. The basic concept of this parser is to transform semi structured properties of a citation string into a sequence template, and apply parsing technique to further resolve the structured information.

References

[1] BibPro: A Citation Parser Based on Sequence Alignment Chien-Chih Chen, Kai-Hsiang Yang, Chuen-Liang Chen, and Jan-Ming Ho. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 24, No. 2, February 2012.

[2] Citation Parsing Using Maximum Entropy and Repairs ,Ng Yong Kiat , Department of Computer Science ,School of Computing , National University of Singapore .2004/2005.

[3] A Knowledge-based Approach to Citation Extraction, Min-Yuh Day1,2, Tzong-Han Tsai1,3, Cheng-Lung Sung1, Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, shwu@cyut.edu.tw.

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

[5] 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.

[6] P. Yin, M. Zhang, Z. Deng, and D. Yang. Metadata extraction from bibliographies using bigram HMM. In Proc. of the 7th Intl. Conf. on Asian Digital Libraries, LCNS 3334, pages 310-319, 2004.

[7] Citation Matching in Sanskrit Corpora Using Local Alignment Abhinandan S. Prasad and Shrisha Rao International Institute of Information Technology, Bangalore abhinandan.sp@iiitb.net, srao@iiitb.ac.in

[8] A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching, Deepank Gupta, Bob Morris, Terry Catapano, and Guido Sautter Netaji Subhash Institute of Technology, Plazi.

[9] V. Borkar, K. Deshmukh, and S. Sarawagi, “Automatic Segmentation of Text into Structured Records,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2001.