A Knowledge-based Approach to Citation Extraction

Min-Yuh Day[1,2], Tzong-Han Tsai[1,3], Cheng-Lung Sung[1],
Cheng-Wei Lee[1], Shih-Hung Wu[4], Chorng-Shyong Ong[2], Wen-Lian Hsu[1]
[1] Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan
[2] Department of Information Management, National Taiwan University, Taipei, Taiwan
[3] Department of Computer Science and Engineering, National Taiwan University, Taipei, Taiwan
[4] Dept. of Computer Science and Information Engineering, Chaoyang Univ. of Technology, Taiwan

{myday, thsai, clsung, aska, hsu}@iis.sinica.edu.tw, shwu@cyut.edu.tw, ongcs@im.ntu.edu.tw

Abstract

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. In this paper, we propose a knowledge-based approach to literature mining and focus on reference metadata extraction methods for scholarly publications. We adopt an ontological knowledge representation framework called INFOMAP to automatically extract the reference metadata. The experimental results show that, by using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different reference styles with a high degree of accuracy. The overall average field accuracy of citation extraction for a Bioinformatics dataset is 97.87% for six reference styles.

.

Keywords: Citation extraction, INFOMAP, Knowledge-based, Ontology

1. Introduction

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 the available heterogeneous reference sources, where metadata is defined as structured data about data [2]. In this paper, reference metadata refers to the sub-fields of references or citations.

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 [1]. Within fields, further separator issues are caused by punctuation and spacing differences. To further compound the problem, there are many dramatically different citation styles (i.e., different field orders).

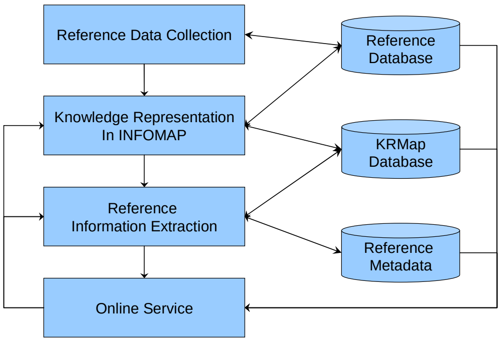
Some systems attempt to extract citation information from digital document references [3, 5, 8, 9, 12, 15, 16]. CiteSeer [8, 9, 12] is an example of an automatic citation indexing system that indexes academic literature in electronic formats. It uses machine learning techniques to identify various forms of citations of the same paper. Chowdhury [3] and Ding et al. [5], on the other hand, use a template mining approach for the extraction of citations from digital documents.

In this paper, we propose a knowledge-based Reference Metadata Extraction (RME) method for scholarly publications. The ontology we adopt, INFOMAP, is a knowledge representation framework that extracts important citation concepts from a natural language text [17, 18]. A powerful feature of INFOMAP is its capability to represent and match complicated template structures, such as hierarchical matching, regular expressions, semantic template matching, frame (non-linear relations) matching, and graph matching. Using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different kinds of reference formats or styles.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach. Section 3 discusses the experimental results. In Section 4, we compare our approach with related works. Finally, in Section 5, we present our conclusions and directions for future research.

2. Proposed Approach

There are four phases in the system framework of our knowledge-based RME for scholarly publications: (1) Reference Data Collection, (2) Knowledge Representation in INFOMAP, (3) Reference Metadata Extraction, and (4) Knowledge-based Reference Metadata Extraction - online service. Figure 1 shows the system framework of knowledge-based RME for scholarly publications.



Reference Data Collection
Knowledge Representation In INFOMAP
Reference Information Extraction
Online Service
Reference Database
KRMap Database
Reference Metadata

Figure 1. The system framework of knowledge-based RME.

We describe the four phases of the system framework in the following sub-sections.

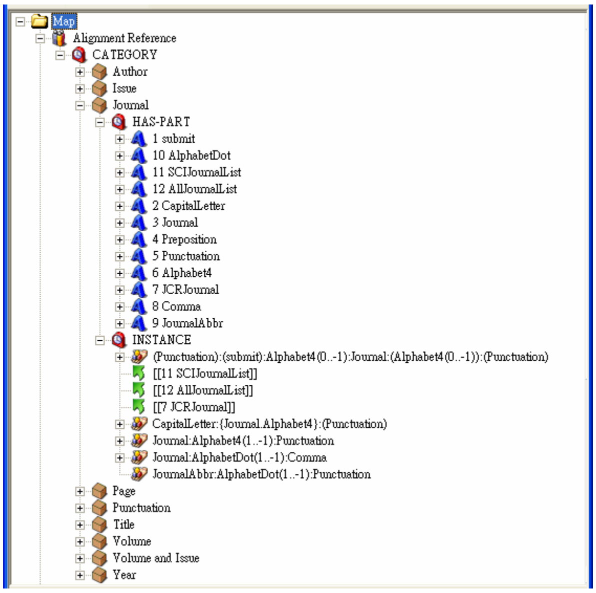
2.1. Reference Data Collection

In the data collection stage, we use Journal Spider (a journal agent) to collect journal data from the Journal Citation Reports (JCR) indexed by the ISI and digital libraries on the Web. The major part of the citation data is taken from the ISI web of science, DBLP, Citeseer, and PubMed. We then cache the data (around 160,000 records) in the reference database as the knowledge representation data source.

2.2. Knowledge Representation in INFOMAP

In the knowledge representation stage, we use Compass as the knowledge editing tool for RME in INFOMAP [17, 18]. INFOMAP is an ontological knowledge representation framework that provides an integrated environment for extracting important citation concepts from a reference. The format of INFOMAP is a tree-like knowledge representation scheme that organizes knowledge of reference concepts in a hierarchical fashion. An example of knowledge representation for knowledge-based RME in INFOMAP is shown in Figure 2.

We represent the basic knowledge of reference concepts in INFOMAP to extract the author, title, journal, volume, issue, year, and page information from different types of reference styles.



```
Map
  Alignment Reference
    CATEGORY
      Author
      Issue
      Journal
        HAS-PART
          1 submit
          10 AlphabetDot
          11 SCIJournalList
          12 AllJournalList
          2 CapitalLetter
          3 Journal
          4 Preposition
          5 Punctuation
          6 Alphabet4
          7 JCRJournal
          8 Comma
          9 JournalAbbr
        INSTANCE
          (Punctuation):(submit):Alphabet4(0..-1):Journal:(Alphabet4(0..-1)):(Punctuation)
          [[11 SCIJournalList]]
          [[12 AllJournalList]]
          [[7 JCRJournal]]
          CapitalLetter:{Journal.Alphabet4}:(Punctuation)
          Journal:Alphabet4(1..-1):Punctuation
          Journal:AlphabetDot(1..-1):Comma
          JournalAbbr:AlphabetDot(1..-1):Punctuation
      Page
      Punctuation
      Title
      Volume
      Volume and Issue
      Year
```

Figure 2. An example of knowledge representation for knowledge-based RME in INFOMAP

2.3. Reference Metadata Extraction

As there are many reference styles in scholarly publications, citations of an article can be given in dramatically different formats. Table 1 presents examples of six different reference styles of citations listed for the paper "Successful Knowledge Management Projects" by Davenport, DeLong, and Beers's [4]. For example, the reference formatted in the APA style looks like:

> Davenport, T., DeLong, D., & Beers, M. (1998). Successful knowledge management projects. Sloan Management Review, 39(2), 43-57.

Or it could be formatted in the IEEE style as:

> [1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan Management Review, vol. 39, no. 2, pp. 43-57, 1998.

There are several ways to separate fields that have different kinds of reference styles. For example, the author field and title field can be separated by periods or commas. Within fields, further separator issues occur because of different uses of punctuation, such as periods, commas, colons, semi-colons, question marks, and spacing.

Nevertheless, we can still extract tagged field information from different formats of reference data. In the reference information extraction stage, we use INFOMAP and the Alignment Reference Citation Agent to extract author, title, journal, volume, number (issue), year, and page information from different kinds of reference formats.

Table 1. Examples of different journal reference styles

Journal Reference styles	Reference style example
Bioinformatics style (BIOI)	Davenport, T., DeLong, D., & Beers, M. (1998) Successful knowledge management projects. Sloan Management Review, 39(2), 43-57.
ACM style (ACM)	1. Davenport, T., DeLong, D. and Beers, M. 1998. Successful knowledge management projects. Sloan Management Review, 39 (2). 43-57.
IEEE style (IEEE)	[1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan Management Review, vol. 39, no. 2, pp. 43-57, 1998.
APA style (APA)	Davenport, T., DeLong, D., & Beers, M. (1998). Successful knowledge management projects. *Sloan Management Review, 39*(2), 43-57.
JCB style (JCB)	Davenport, T., DeLong, D., & Beers, M. 1998. Successful knowledge management projects. Sloan Management Review 39(2), 43-57.
MISQ style (MISQ)	Davenport, T., DeLong, D., and Beers, M. "Successful knowledge management projects," Sloan Management Review (39:2) 1998, pp 43-57.

2.4. Knowledge-based Reference Metadata Extraction - Online Service

There are three parts to the online web system of knowledge-based RME for scholarly publications: (1) the system input area for journal publication references; (2) the system output of RME; and (3) BibTeX format output for data exchange and integration.

1. The system input of the knowledge-based reference metadata extraction for scholarly Publications is shown in Figure 3.

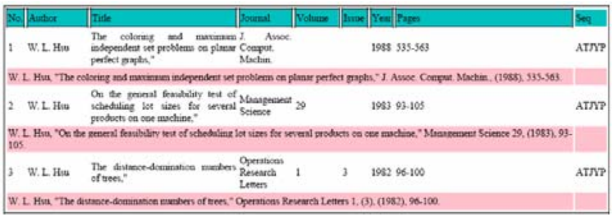
W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.

W. L. Hsu, "On the general feasibility test of scheduling lot sizes for several products on one machine," Management Science 29, (1983), 93-105.

W. L. Hsu, "The distance-domination numbers of trees," Operations Research Letters 1, (3), (1982), 96-100.

Figure 3. The system input of the knowledge-based RME

2. The system output of RME is shown in Figure 4.



No.	Author	Title	Journal	Volume	Issue	Year	Pages	Seq
1	W. L. Hsu	The coloring and maximum independent set problems on planar perfect graphs,"	J. Assoc. Comput. Machin.			1988	535-563	ATJYP
W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.								
2	W. L. Hsu	On the general feasibility test of scheduling lot sizes for several products on one machine,"	Management Science	29		1983	93-105	ATJYP
W. L. Hsu, "On the general feasibility test of scheduling lot sizes for several products on one machine," Management Science 29, (1983), 93-105.								
3	W. L. Hsu	The distance-domination numbers of trees,"	Operations Research Letters	1	3	1982	96-100	ATJYP
W. L. Hsu, "The distance-domination numbers of trees," Operations Research Letters 1, (3), (1982), 96-100.								

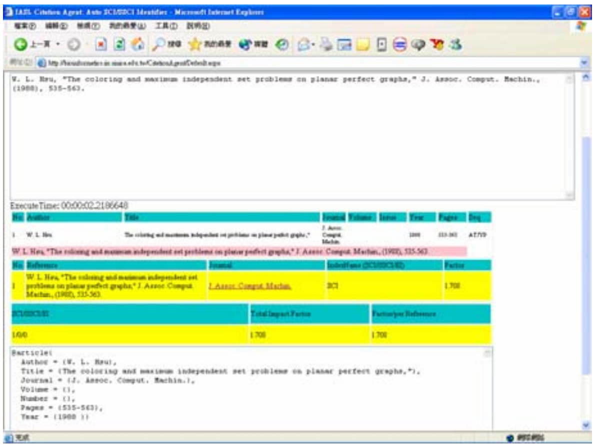
Figure 4. The system output of the knowledge-based RME

3. The system output of BibTex Format is shown in Figure 5.

```
@article{
  Author = {W. L. Hsu},
  Title = {The coloring and maximum independent set problems on planar perfect graphs,"},
  Journal = {J. Assoc. Comput. Machin.},
  Volume = {},
  Number = {},
  Pages = {535-563},
  Year = {1988 }}
@article{
  Author = {W. L. Hsu},
  Title = {On the general feasibility test of scheduling lot sizes for several products on one machine,"},
  Journal = {Management Science},
  Volume = {29},
  Number = {},
  Pages = {93-105},
  Year = {1983 }}
@article{
  Author = {W. L. Hsu},
  Title = {The distance-domination numbers of trees,"},
  Journal = {Operations Research Letters},
  Volume = {1},
  Number = {3},
  Pages = {96-100},
  Year = {1982 }}
```

Figure 5. The system output of BibTex Format

Users can input the plain text of a journal publication reference that is in different citation styles into the system. The online system will automatically extract author, title, journal, volume, number (issue), year, and page metadata from the different reference styles and output the BibTex format for data exchange and integration. Figure 6 shows the online service of knowledge-based RME.



IASL Citation Agent: Auto SCI/SSCI Identifier - Microsoft Internet Explorer

http://bioinformatics.iis.sinica.edu.tw/CitationAgent/Default.aspx

```
W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin.,
(1988), 535-563.
```

ExecuteTime: 00:00:02.2186648

No.	Author	Title	Journal	Volume	Issue	Year	Pages	Deq
1	W. L. Hsu	The coloring and maximum independent set problems on planar perfect graphs,"	J. Assoc. Comput. Machin.			1988	535-563	[illegible]

W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563

No.	Reference	Journal	IndexName (SCI/SSCI/EI)	Factor
1	W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563	J. Assoc. Comput. Machin.	SCI	1.708

SCI/SSCI/EI	Total Impact Factor	Factor/per Reference
1/0/0	1.708	1.708

```
@article{
  Author = {W. L. Hsu},
  Title = {The coloring and maximum independent set problems on planar perfect graphs,"},
  Journal = {J. Assoc. Comput. Machin.},
  Volume = {},
  Number = {},
  Pages = {535-563},
  Year = {1988 }}
```

Figure 6. The online service of knowledge-based RME,
(http://bioinformatics.iis.sinica.edu.tw/CitationAgent/)

3. Experimental Results and Discussion

The experimental results of knowledge-based metadata extraction for scholarly publications are presented below.

3.1. Experimental results of citation extraction from six reference styles

We used EndNote to collect Bioinformatics citation data for 2004 from PubMed (http://www.ncbi.nlm.nih.gov/entrez/). A total of 907 bibliography records were collected from PubMed digital libraries on the Web. Reference testing data was generated for each of the six reference styles (BIOI, ACM, IEEE, APA, MISQ, and JCB). We then randomly selected 500 records for testing from each of the six reference styles.

In this experiment, we consider a field to be correctly extracted only when the field values in the reference testing data are correctly extracted. The accuracy of citation extraction is defined as follows:

$$Accuracy = \frac{Number\ of\ correctly\ extracted\ \ fields}{Total\ number\ of\ \ fields} \qquad (1)$$

The performance measure we define here is the field accuracy, which is different from word accuracy and instance accuracy defined in [14].

Figure 7 summarizes the experimental results of citation extraction for the six different reference styles. The overall average accuracy of citation extraction for the six styles is 97.87%. The best overall average accuracy is 98.20% for the MISQ style. In particular, the average accuracy of the journal field is 99.77% for the six styles, and the individual accuracy of the MISQ reference style is 100%. These results indicate that our method is quite reliable.

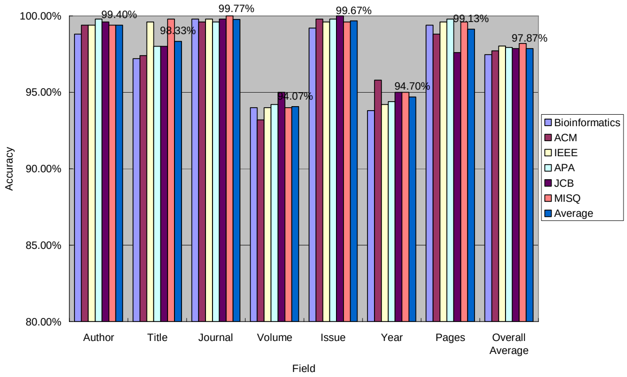
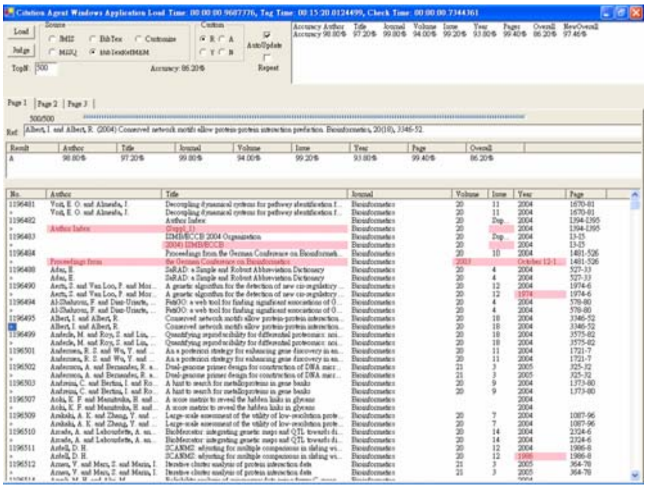


Figure 7. Experimental results of citation extraction from six reference styles

Figure 8 shows the results of the selected reference database experiment. The results show that by using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different reference styles with a high degree of precision.



Citation Agent Windows Application Load Time: 00:00:00.9687776, Tag Time: 00:15:20.0124499, Check Time: 00:00:00.7344361

TopN: 500

Accuracy: 86.20%

Accuracy	Author	Title	Journal	Volume	Issue	Year	Pages	Overall	NewOverall
Accuracy	98.80%	97.20%	99.80%	94.00%	99.20%	93.80%	99.40%	86.20%	97.46%

500/500

Ref: Albert, I. and Albert, R. (2004) Conserved network motifs allow protein-protein interaction prediction. Bioinformatics, 20(18), 3346-52.

Result	Author	Title	Journal	Volume	Issue	Year	Page	Overall
A	98.80%	97.20%	99.80%	94.00%	99.20%	93.80%	99.40%	86.20%

No.	Author	Title	Journal	Volume	Issue	Year	Page
1196481	Voit, E. O. and Almeida, J.	Decoupling dynamical systems for pathway identification f...	Bioinformatics	20	11	2004	1670-81
*	Voit, E. O. and Almeida, J.	Decoupling dynamical systems for pathway identification f...	Bioinformatics	20	11	2004	1670-81
1196482		Author Index	Bioinformatics	20	Sup...	2004	1394-1395
*	Author Index	(Suppl_1)	Bioinformatics	20		2004	1394-1395
1196483		ISMB/ECCB 2004 Organization	Bioinformatics	20	Sup...	2004	13-15
*		2004) ISMB/ECCB	Bioinformatics	20		2004	13-15
1196484		Proceedings from the German Conference on Bioinformati...	Bioinformatics	20	10	2004	1481-526
*	Proceedings from	the German Conference on Bioinformatics	Bioinformatics	2003		October 12-1	1481-526
1196488	Adar, E.	SaRAD: a Simple and Robust Abbreviation Dictionary	Bioinformatics	20	4	2004	527-33
*	Adar, E.	SaRAD: a Simple and Robust Abbreviation Dictionary	Bioinformatics	20	4	2004	527-33
1196490	Aerts, S. and Van Loo, P. and Mor...	A genetic algorithm for the detection of new cis-regulatory ...	Bioinformatics	20	12	2004	1974-6
*	Aerts, S. and Van Loo, P. and Mor...	A genetic algorithm for the detection of new cis-regulatory ...	Bioinformatics	20	12	1974	1974-6
1196494	Al-Shahrour, F. and Diaz-Uriarte, ...	FatiGO: a web tool for finding significant associations of G...	Bioinformatics	20	4	2004	578-80
*	Al-Shahrour, F. and Diaz-Uriarte, ...	FatiGO: a web tool for finding significant associations of G...	Bioinformatics	20	4	2004	578-80
1196495	Albert, I. and Albert, R.	Conserved network motifs allow protein-protein interaction...	Bioinformatics	20	18	2004	3346-52
*	Albert, I. and Albert, R.	Conserved network motifs allow protein-protein interaction...	Bioinformatics	20	18	2004	3346-52
1196499	Anderle, M. and Roy, S. and Lin, ...	Quantifying reproducibility for differential proteomics: noi...	Bioinformatics	20	18	2004	3575-82
*	Anderle, M. and Roy, S. and Lin, ...	Quantifying reproducibility for differential proteomics: noi...	Bioinformatics	20	18	2004	3575-82
1196501	Andersson, R. S. and Wu, Y. and ...	An a posteriori strategy for enhancing gene discovery in an...	Bioinformatics	20	11	2004	1721-7
*	Andersson, R. S. and Wu, Y. and ...	An a posteriori strategy for enhancing gene discovery in an...	Bioinformatics	20	11	2004	1721-7
1196502	Andersson, A. and Bernander, R. a...	Dual-genome primer design for construction of DNA micr...	Bioinformatics	21	3	2005	325-32
*	Andersson, A. and Bernander, R. a...	Dual-genome primer design for construction of DNA micr...	Bioinformatics	21	3	2005	325-32
1196503	Andreini, C. and Bertini, I. and Ro...	A hunt to search for metalloproteins in gene banks	Bioinformatics	20	9	2004	1373-80
*	Andreini, C. and Bertini, I. and Ro...	A hunt to search for metalloproteins in gene banks	Bioinformatics	20	9	2004	1373-80
1196507	Aoki, K. F. and Mamitsuka, H. and ...	A score matrix to reveal the hidden links in glycans	Bioinformatics			2004	
*	Aoki, K. F. and Mamitsuka, H. and ...	A score matrix to reveal the hidden links in glycans	Bioinformatics			2004	
1196509	Arakaki, A. K. and Zhang, Y. and ...	Large-scale assessment of the utility of low-resolution prote...	Bioinformatics	20	7	2004	1087-96
*	Arakaki, A. K. and Zhang, Y. and ...	Large-scale assessment of the utility of low-resolution prote...	Bioinformatics	20	7	2004	1087-96
1196510	Arcade, A. and Labourdette, A. an...	BioMercator: integrating genetic maps and QTL towards di...	Bioinformatics	20	14	2004	2324-6
*	Arcade, A. and Labourdette, A. an...	BioMercator: integrating genetic maps and QTL towards di...	Bioinformatics	20	14	2004	2324-6
1196511	Ardell, D. H.	SCANMS: adjusting for multiple comparisons in sliding wi...	Bioinformatics	20	12	2004	1986-8
*	Ardell, D. H.	SCANMS: adjusting for multiple comparisons in sliding wi...	Bioinformatics	20	12	1986	1986-8
1196512	Arnau, V. and Mars, S. and Marin, I.	Iterative cluster analysis of protein interaction data	Bioinformatics	21	3	2005	364-78
*	Arnau, V. and Mars, S. and Marin, I.	Iterative cluster analysis of protein interaction data	Bioinformatics	21	3	2005	364-78

Figure 8. Results of the selected reference database experiment

3.2. Analysis of the structure of reference styles

There are punctuation and spacing differences between field separators in the structure of reference styles. For example, the structures used between the volume and issue fields are difficult in the APA style and the IEEE style. The volume and issue fields are separated by parentheses in the APA style, where as a comma is used in the IEEE style.

To estimate the generality of our approach, we randomly selected more than 30 styles to test our method. Table 2 summarizes the analysis of field relation structure. For example, there are two possible field relation structures for the author field. One is "<Author><Year>", which accounts for 54.29%; and the other is "<Author><Title>", which accounts for 42.8%. However, the most common sequence structure of reference styles is "<Author> <Year> <Title> <Journal> <Volume> <Issue> <Pages>".

Table 2. Analysis of Field Relation Structures

Field	Field Relation Structure	Percentage%
Author	<Author><Year>	54.29%
	<Author><Title>	42.86%
	N/A	2.85%
Year	<Author><Year><Title>	48.57%
	<Journal><Year><Volume>	20.00%
	<Issue><Year><Pages>	14.29%
	<Author><Year><Journal>	5.71%
	<Pages><Year>	2.86%
	<Volume><Year><Pages>	2.86%
	N/A	5.71%
Title	<Year><Title><Journal>	48.57%
	<Author><Title><Journal>	42.86%
	N/A	8.57%
Journal	<Title><Journal><Volume>	71.43%
	<Title><Journal><Year>	20.00%
	<Year><Journal><Volume>	5.71%
	N/A	2.86%
Volume	<Journal><Volume><Pages>	40.00%
	<Journal><Volume><Issue>	31.43%
	<Year><Volume><Issue>	14.29%
	<Year><Volume><Pages>	5.71%
	<Journal><Volume><Volume>	2.86%
	<Journal><Volume><Year>	2.86%
	N/A	2.85%
Issue	<Volume><Issue><Pages>	34.29%
	<Volume><Issue><Year>	14.29%
	N/A	51.42%
Pages	<Volume><Pages>	42.86%
	<Issue><Pages>	34.29%
	<Year><Pages>	17.14%
	<Volume><Pages><Year>	2.86%
	N/A	2.85%

We conducted experiments on the other 30 styles without additional knowledge editing and obtained an average accuracy of reference extraction of approximately 87% for those styles. In particular, the overall average accuracy of reference extraction is 88.20% for the MLA Style. The experimental results show that our knowledge-based method is reliable for different kinds of unseen reference styles.

Thirteen types of punctuation that can be used in some fields for different reference styles. For example, a comma can be used in author, volume, issue, and page fields. It can also used as the major field separator for the reference styles of BIOI, ACM, IEEE, APA, MISQ, and JCB; while the major field separator in the JCB style is a period.

We use templates in INFOMAP to represent the punctuation differences between field separators in the structure of reference styles.

4. Related Works

Extensive works on extracting citations information from digital document references have been reported in the literature [3, 5, 8, 9, 12, 14, 15]. In this section, we compare related works that are based on machine learning models and rule-based models.

First, a machine learning approach such as Citeseer [8, 9, 12] takes advantage of probabilistic estimation, which is based on the training sets of tagged bibliographical data, to boost performance. The citation parsing technique of Citeseer can identify titles and authors with approximately 80% accuracy and page numbers with approximately 40% accuracy [8]. Seymore et al. [15] use the Hidden Markov Model (HMM) to extract important fields from the headers of computer science research papers, and achieved an overall word accuracy of 92.9%. Peng et al. [14] employ Conditional Random Fields (CRF) to extract various common fields from the headers and citations of research papers. The overall word accuracy is 98.3% for extracting fields from paper headers. The reference dataset they use is the Cora dataset [14], which contains 500 references categorized into 13 fields: author, title, editor, book_title, date, journal, volume, tech, institution, pages, location, publisher, and note. Peng et al. [14] achieve an overall word accuracy of 85.1%(HMM) compared to 95.37%(CRF) and an overall instance accuracy of 10%(HMM) compared to 77.33%(CRF) for paper references.

Second, rule-based models such as those developed by Chowdhury [3] and Ding et al. [5] use a template mining approach for citation extraction from digital documents. Ding et al. [5] use three templates for extracting information from cited articles (citations) and obtain a quite satisfactory result (more than 90%) for the distribution of information extracted from each unit in cited articles. The advantage of their rule-based model is its efficiency in extracting reference information. However, they treat references in one style only from tagged texts (e.g., references formatted in HTML), whereas our method treats references in more than six reference styles from plain text.

In contrast to the approaches that use a small amount of testing data, our proposed knowledge-based RME method for scholarly publications can extract reference information from 907 records in various reference styles with a high degree of precision (the overall average field accuracy is 97.87% for six major styles, 98.20% for the

MISQ Style and 87% for 30 other randomly selected styles).

5. Conclusions and Future Research

RME is a challenging problem due to the diverse nature of reference styles. In this paper, we have proposed a knowledge-based RME method for scholarly publications. The experimental results indicate that, by using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different reference styles with a high degree of precision. The overall average field accuracy of citation extraction is 97.87% for six reference styles.

We highlight two major directions for future research, namely, ontology learning and machine learning. The integration of knowledge acquisition with machine-learning techniques has already proved extremely beneficial for knowledge acquisition [6, 10, 11, 13]. For ontology learning, there are two possible ways to acquire the concept relationship on ontology learning: (1) a prior approach: given domain documents, find all the concept-instance pairs [7]; and (2) an online-testing approach: enumerate all possible concept-instance pairs and then evaluate the concept-instance pairs by a search engine.

In the future, we will integrate both the ontological and the machine learning approaches (such as Maximum-Entropy Method (MEM), Hidden Markov Model (HMM), Conditional Random Fields (CRF) and Support Vector Machines (SVM)) to boost the performance of citation information extraction and produce a more robust prototype that can deal with free style references as well as input containing errors.

5. Acknowledgements

This research was supported in part by the National Science Council under GRANT NSC93-2752-E-001-001.

6. References

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

[2] K. Burnett, K. B. Ng, and S. Park, "A comparison of the two traditions of metadata development", *Journal of the American Society for Information Science*, vol. 50, pp. 1209 - 1217, 1999.

[3] G. Chowdhury, "Template mining for information extraction from digital documents", *Library Trends*, vol. 48, pp. 182-208, 1999.

[4] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects", *Sloan Management Review*, vol. 39, pp. 43-57, 1998.

[5] Y. Ding, G. Chowdhury, and S. Foo, "Template mining for the extraction of citation from digital documents", Proceedings of the Second Asian Digital Library Conference, Taiwan, 1999. pp. 47-62.

[6] Y. Ding and S. Foo, "Ontology research and development. Part I - a review of ontology generation", *Journal of Information Science*, vol. 28, pp. 123-136, 2002.

[7] M. Fleischman, E. Hovy, and A. Echihabi, "Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked", Proceedings of ACL-2003 Conference, 2003. pp. 1-7.

[8] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: An Automatic Citation Indexing System", Digital Libraries 98 - The Third ACM Conference on Digital Libraries, 1998. pp. 89-98.

[9] A. Goodrum, K. McCain, S. Lawrence, and C. Giles, "Scholarly publishing in the Internet age: a citation analysis of computer science literature", *Information Processing & Management*, vol. 37, pp. 661-675, 2001.

[10] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines", JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, 2003. pp. 37-48.

[11] P. Jacso, "The future of citation indexing: An interview with Eugene Garfield", *Online*, vol. 28, pp. 38-40, 2004.

[12] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing", *Computer*, vol. 32, pp. 67-71, 1999.

[13] A. Maedche and S. Staab, "Ontology learning from text", *Natural Language Processing and Information Systems*, vol. 1959, pp. 364-364, 2001.

[14] F. Peng and A. McCallum, "Accurate Information Extraction from Research Papers using Conditional Random Fields", Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2004. pp. 329-336.

[15] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden Markov model structure for information extraction", AAAI-99 Workshop on Machine Learning for Information Extraction, 1999. pp. 37-42.

[16] A. Takasu, "Bibliographic attribute extraction from erroneous references based on a statistical model", JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, 2003. pp. 49-60.

[17] S.-H. Wu, M.-Y. Day, and W.-L. Hsu, "FAQ-centered Organizational Memory", Proceeding of the Knowledge Management and Organizational Memory workshop on the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001. pp. 112-120.

[18] S.-H. Wu, T.-H. Tsai, and W.-L. Hsu, "Domain Event Extraction and Representation with Domain Ontology", Proceedings of the IJCAI-03 Workshop on Information Integration on the Web, Acapulco, Mexico, 2003. pp. 33-38.