Open access • Journal Article • DOI:10.1016/J.DSS.2006.08.006

Reference metadata extraction using a hierarchical knowledge representation framework — Source link

Min-Yuh Day, Richard Tzong-Han Tsai, Cheng-Lung Sung, Chiu-Chen Hsieh ...+5 more authors

Institutions: National Taiwan University, Academia Sinica, Chaoyang University of Technology

Published on: 01 Feb 2007 - Decision Support Systems

Topics: Metadata repository, Metadata, Knowledge extraction, Information integration and Knowledge representation and reasoning

Reference Metadata Extraction Using a Hierarchical Knowledge Representation Framework

Min-Yuh Day [a,b], Richard Tzong-Han Tsai [a], Cheng-Lung Sung [a], Chiu-Chen Hsieh [a], Cheng-Wei Lee [a], Shih-Hung Wu [c], Kun-Pin Wu [a], Chorng-Shyong Ong [b], Wen-Lian Hsu [a,*]

[a] *Institute of Information Science, Academia Sinica, Nankang, Taipei 115, Taiwan, R.O.C.*
[b] *Department of Information Management , National Taiwan University, Taipei 106, Taiwan, R.O.C.*
[c] *Department of CSIE, Chaoyang University of Technology, Taichung County 413, Taiwan, R.O.C.*

Abstract

The integration of bibliographical information on scholarly publications available on the Internet is an important task in the academic community. Accurate reference metadata extraction from such publications is essential for the integration of metadata from heterogeneous reference sources. In this paper, we propose a hierarchical template-based reference metadata extraction method for scholarly publications. We adopt a hierarchical knowledge representation framework called INFOMAP, which automatically extracts metadata. The experimental results show that, by using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different kinds of reference styles with a high degree of precision. The overall average accuracy is 92.39% for the six major reference styles compared in this study.

Keywords: Reference extraction, Metadata extraction, Knowledge representation framework, INFOMAP

[*] Corresponding author. Tel.: +886-2-27883799x1804; Fax.: +886-2-27824814
E-mail addresses: myday@iis.sinica.edu.tw (M.-Y. Day), thsai@iis.sinica.edu.tw (T.-H. Tsai), clsung@iis.sinica.edu.tw (C.-L. Sung), gladys@iis.sinica.edu.tw (C.-C. Hsieh), aska@iis.sinica.edu.tw (C.-W. Lee), shwu@cyut.edu.tw (S.-H. Wu), kpw@iis.sinica.edu.tw (K.-P. Wu), ongcs@im.ntu.edu.tw (C.-S. Ong), hsu@iis.sinica.edu.tw (W.-L. Hsu)

1. Introduction

The integration of bibliographical information on scholarly publications available on the Internet is an important task in the academic community. Accurate reference metadata extraction from such publications is essential for the integration of metadata from heterogeneous reference sources, where metadata is defined as structured data about data [5, 18, 22]. In this paper, reference metadata refers to the sub-fields of references or citations.

Automatic citation extraction is an extremely challenging task due to variations in the use of 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 [3]. Meanwhile, within fields, other separator issues arise because of punctuation and spacing differences. To further compound the problem, there are many dramatically different citation styles (e.g., different field orders).

Some systems attempt to extract citation information from digital document references [6, 9, 10, 11, 12, 13, 15, 17, 19, 20]. CiteSeer [11, 12, 15] 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 citation of the same paper. Chowdhury [6] and Ding et al. [9], on the other hand, use a template mining approach to extract citations from digital documents.

In this paper, we propose a hierarchical template-based Reference Metadata Extraction (RME) method for scholarly publications. The approach we adopt, called INFOMAP, is a hierarchical knowledge representation framework that extracts important concepts from natural language texts [23, 24]. As a general representation of domain knowledge, INFOMAP consists of domain concepts and their related sub-concepts, such as categories, attributes and actions. The relationships of a concept to its associated sub-concepts form a tree-like taxonomy in which INFOMAP classifies both concepts and related concepts [24]. A powerful feature of the framework is its ability to represent and match complicated template structures. Using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different reference styles.

The remainder of this paper is organized as follows. Section 2 describes the background to citation extraction and previous works. Section 3 describes the phases in system development. Section 4 discusses the experiment and the test bed, and Section 5 contains the experimental results. In Section 6, we compare our proposed approach with related works. Finally, in Section 7, we present our conclusions and suggest some directions for future research.

2. Previous Work

Numerous works on extracting citation information from digital document references are reported in literature [6, 9, 10, 11, 12, 13, 15, 17, 19, 20]. References and citations refer to publication lists in a citation format [14]. Citation extraction is a subtask of Information Extraction (IE) that automatically segments unstructured text strings into structured records. This procedure is necessary in order to import the information contained in legacy sources and text collections into a data warehouse for subsequent querying, analysis, mining, and integration [1, 2].

The various methods of citation extraction can be classified as either machine learning or rule-based approaches.

2.1 Machine learning approaches

The basic concept of the machine learning approach is to learn the relationship between the input and output of samples and then predict new data. Although this approach has good adaptability, it must be trained from samples.

Machine learning approaches solve this problem by automatically learning segmentation models from training data comprised of input strings and their associated segmented records [1, 2]. Such approaches facilitate robust and adaptable automatic metadata extraction [13]. In addition, approaches like Citeseer [11, 12, 15] take advantage of probabilistic estimation, which is based on training sets of tagged bibliographic data, to boost performance. Seymore et al. [19] use the Hidden Markov Model (HMM) to extract important fields from the headers of computer science research papers. Peng et al. [17] employ Conditional Random Fields (CRF) to extract various common fields from the headers and citations of research papers, while Takasu [20] uses a statistical model to extract bibliographic attributes from erroneous references.

2.2 Rule-based Approaches

Rule-based models, such as those developed by Chowdhury [6] and Ding et al. [9], 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. Chowdhury argues that template mining enables citation databases to be built automatically from digital documents. However, rule-based approaches require a domain expert to design a number of rules and maintain them over time. In addition, this approach does not scale well, as deployment of each new domain requires a new set of rules to be designed, crafted, deployed, and maintained.

In contrast to machine learning approaches, rule-based methods of metadata extraction use a set of rules that define how to extract metadata based on human observation. The advantages of such approaches are that they can be implemented in a straightforward manner without training. However, the disadvantages of typical rule-based approaches are their lack of adaptability, difficulty in working with a large number of features, and difficulty in tuning the system because the rules are rigid.

Unlike typical rule-based approaches, which are flat, our template-based approach is hierarchical. As a general representation of domain knowledge, our INFOMAP [23, 24] comprises domain concepts and their related sub-concepts, such as categories, attributes and actions. The relationships between a concept and its associated sub-concepts form a tree-like taxonomy. INFOMAP, which classifies concepts as well as related concepts [24], has been successfully applied in a number of areas, for example, Question Answering Systems [23], Intelligent Tutoring Systems [8, 16], and various other applications [24]. A powerful feature of INFOMAP is its ability to represent and match complicated template structures, such as hierarchical matching, regular expressions, semantic template matching, frame (non-linear relations) matching, and graph matching, which can be used to extract important citation concepts from a natural language text.

INFOMAP is more powerful than typical rule-based approaches, since it provides an integrated hierarchical template editing environment and a flexible template matching engine. The editing environment comprises taxonomy editing, template authoring, template syntax checking, and testing. The matching engine can accept context sensitive rules or templates, such as restricting the existence of an issue field within specific contexts. In addition, INFOMAP achieves better accuracy by using a smaller number of labeled datasets compared to typical machine learning approaches. This reduction in the required amount of annotated training data is crucial in real-world scenarios, because it facilitates more rapid deployment of reference metadata extraction. Using INFOMAP, we can extract author, title, journal, volume, number (issue), year, and page information from different kinds of reference styles more efficiently.

3. Phases in System Development

Our template-based reference metadata extraction system for scholarly publications is comprised of four phases: (1) Reference Data Collection, (2) Knowledge Representation in INFOMAP, (3) Reference Metadata Extraction, and (4) Template-based Reference Metadata Extraction – an online service. Fig. 1 shows the system framework of our approach.

We describe the four phases in the following sub-sections. In addition, in Section 3.3, we present a detailed example to better explain how the system works.

3.1 Reference Data Collection

In the data collection stage, we use Journal Spider to retrieve citations from publicly available indexing and abstracting databases (ISI Web of Science, DBLP, PubMed) in HTML format. We then cache the data in the reference database as the knowledge representation data source.

3.2 Knowledge Representation in INFOMAP

In the knowledge representation stage, we use Compass as the knowledge editing tool for RME in INFOMAP, which is a hierarchical 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 the knowledge of reference concepts in a hierarchical fashion. Fig. 2 shows an example of knowledge representation for template-based RME in INFOMAP.

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. The details of each field are described below.

(1) Author

In the author name field of a citation, each author's name may have a different format. For example, the names of the first author and the other authors may have different sequence formats (e.g., Jane Smith; Smith, Jane; or Smith Jane). Also, authors' names can be expressed with different capitalization (e.g., As Is; Normal; All Uppercase; or Small Caps) and Initials (e.g., full name, A. B.; A.B.; A B; AB; or last name only). For subsequent works by the same authors, the citation may list the authors' names as normal (e.g., Jane Smith), omit them completely, or substitute them with specific punctuation such as "---".

There are also separator and abbreviation issues in author lists. For example, two authors' names may have a separator between them with a comma (,) or a separator before the second author with the conjunction " and ". In addition, when there are three or more authors, the separator before the last author's name may be ",and ". With regard to the abbreviation issue, a citation may list all authors' abbreviated names. If there are 100 (or more) authors, the citation may list all 100, or only list the first author and abbreviate the remainder as ", et al." The abbreviation may also be formatted in italics.

(2) Title

The title is free text that may comprise an article title, book title, journal title, conference title, or report title [9]. In this study, we define title as the "article title" in journal references. The title field in a citation may leave the title as entered, present it in headline-style capitalization, or format it in sentence-style capitalization.

(3) Journal

According to Bradford's Law [4], the well-known bibliometric law of publications, the majority of journal references are taken from a minority of journal titles. Vinkler [21] observed that 55% of journal references are derived from only 10% of journal titles, while Ding et al. [9] suggested that a database containing the most frequently cited journals can be used to extract a title. Journal names may be formatted as full journal names, or be replaced by different kinds of abbreviation; for example, journal names may be abbreviated without periods.

(4) Year

The year format may use 4-digits or 2-digits (e.g., '99 for 1999). It may also include the Season (e.g., Spring, Summer, Fall, or Winter), or the Month and Date (e.g., Jan. 1, or Feb. 15).

(5) Volume

A general volume format may include "Volume", or the abbreviation "Vol.", or (); for example:

Barki, Henri; Hartwick, Jon "Interpersonal Conflict and Its Management in Information

System Development," MIS Quarterly (25:2), June 2001, pp. 195-228.

We can represent journal volume information “(25:2)” in the partial structure of a reference “<Journal> <Space> (<Volume> : <Issue>), <Space> <Year>” or “<Journal> : (<Volume> : “:”<No>) , <Year>”.

(6) Issue

A general issue format may include “Issue”, or the abbreviation “No.”, or ().

(7) Page

In addition to normal page numbers, there are five other major styles for formatting page numbers: show only the first page (e.g., 123); abbreviate the last page (e.g., 123-5); abbreviate the last page with 2 digits (e.g., 123-25); show the range of pages (e.g., 123-125); or show only the first page for journals, and the full range of pages for other publications.

3.3 Reference Metadata Extraction

Since 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 that could be used for citations in the paper “Successful Knowledge Management Projects” by Davenport, DeLong and Beers’s [7]. 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 the fields in different 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 punctuation, such as periods, commas, colons, semi-colons, question marks, and spacing.

Nevertheless, we can still extract tagged field information from different reference data formats, as shown in Fig. 5. 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 styles.

The alignment reference citation agent is an application program that preprocesses the reference data, and sends a reference string to INFOMAP. After receiving the attribute and value candidates of the reference metadata, the alignment reference citation agent aligns the attribute and value candidates with a field string sequence, e.g., "ATJYVIP" (Author, Title, Journal, Year, Volume, Issue, Page) for use by the post-processing module.

To better explain how the system works with INFOMAP and the alignment reference citation agent, we provide an example of extracting the "author" field. We also describe how the algorithm resolves conflicts when a given text string maps to more than one possible interpretation.

The INFOMAP representations are mapped to the observation of the author field, and the author's name variations are incorporated into the INFOMAP structure, as shown in Fig. 2. For example, the knowledge representation of the author names "Davenport, T., DeLong, D., & Beers, M." in INFOMAP can be formulated as a template like "AuthorBeginCue:OneAuthor(2..-1)", which is created under the INSTANCE function node of the "MultipleAuthors" node, a CATEGORY of Author, in INFOMAP. There are two elements in the template: "AuthorBeginCue" and "OneAuthor(2..-1)". The template "AuthorBeginCue:OneAuthor(2..-1)" indicates that an identified author begin cue reference, followed by at least two authors can be interpreted as an instance of "MultipleAuthors". The element "AuthorBeginCue" represents a regular expression at the beginning of a string "^" in the REG-EXP function node; while "OneAuthor(2..-1)" represents that "OneAuthor" can be repeated at least twice with no upper limit. Note that we use "-1" to represent infinity. In addition, "OneAuthor" can be represented as a template, such as “EnglishName:(Separator):(Conjunction)”, to indicate that an identified English name followed by an option separator and an option conjunction can be interpreted as an instance of “OneAuthor”. Meanwhile, “EnglishName” can be represented by the template “FamilyName:(Punctuation):FirstName(1..2):(Punctuation)”, where “FirstName(1..2)” represents that “FirstName” is repeated once or twice.

As INFOMAP is domain dependent, the number of templates varies according to the application. In this example, we employ 41 templates (4,099 nodes) to capture the author substructure. We use INFOMAP’s template matching feature, which is designed for identifying syntax patterns of a reference string, to extract the information (e.g., Author) in the above example. The syntax templates in the INFOMAP framework form the basis for matching syntax structures. Although regular expressions have been widely used for such matching, they are difficult to organize and inappropriate for complex structures. Using INFOMAP, we can specify the syntax of a concept by regular expressions, which we then connect by our frame expressions to represent more complex structures. This allows us to identify a pattern within a certain context for a context-sensitive language.

On receipt of an input reference string, such as “Davenport, T., DeLong, D., & Beers, M. (1998). Successful knowledge management projects. Sloan management review, 39(2), 43-57.”, INFOMAP’s template matching engine uses dynamic programming to match it with the syntax templates, as in the above example “AuthorBeginCue:OneAuthor(2..-1)”. The alignment reference citation agent uses two approaches: all possible matches, and longest match, combined with rules to resolve the ambiguity that arises during the template matching process. For example, we adopt two rules to prevent ambiguity between the author field and the title field: 1) statistical information from the family name list and a general English dictionary, and 2) punctuation features. The alignment reference citation agent aligns the attribute (e.g., “Author”) and value candidates (e.g., “Davenport, T., DeLong, D., & Beers, M.”) from INFOMAP with a field string sequence, e.g., “ATJYVIP” (Author, Title, Journal, Year, Volume, Issue, Page).

3.4 Template-based Reference Metadata Extraction - Online Service

The online web system of template-based RME for scholarly publications is comprised of three parts: the system input area for journal publication references, the system output of RME, and the BibTeX format output for data exchange and integration, as shown in Fig. 4, 5, and 6 respectively.

Users can input the plain text of a journal publication reference into the system. Fig. 7 shows the online service of template-based RME.
(http://bioinformatics.iis.sinica.edu.tw/CitationAgent/).

4. Experimental Test Bed

We now present the experimental test bed for template-based metadata extraction.

Journal Spider was used to retrieve citations from publicly available indexing and abstracting databases (ISI Web of Science, DBLP, PubMed) in HTML format. A total of 160,000 reference records were collected from digital libraries on the Web, and reference test data was generated for each of six reference styles (APA, IEEE, ACM, MISQ, JMIS, and ISR). We then randomly selected 10,000 records from each of the six reference styles for testing.

In this experiment, we consider a field to be correctly extracted only when the field values are correctly extracted from the reference test data. 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 [17].

5. Experimental Results and Discussion

5.1 Experimental results of citation extraction from six reference styles

Table 2 and Fig. 8 summarize the experimental results of citation extraction from the six different reference styles. The overall average accuracy for the six styles was 92.39%, while the best individual average accuracy was 99.31% for the MISQ style. Specifically, the average accuracy of the journal field for the six styles was 96.85%, and the individual accuracy of the MISQ style was 100%. These results indicate that our method is very reliable.

Fig. 9 shows the results of the selected reference database experiment. We observe that, by using INFOMAP, author, title, journal, volume, number (issue), year, and page information can be extracted from different reference styles with a high degree of precision.

5.2 Analysis of the structure of reference styles

There are punctuation and spacing differences between field separators in the structure of reference styles, as shown by the analysis in Table 3. For example, the APA and IEEE styles differ in the structure of the volume and issue fields, which are separated by parentheses in the APA style, while the IEEE style uses a comma.

To evaluate the generality of our approach, we randomly selected 30 other styles to test the system. Table 4 summarizes the analysis of field relation structures. There are two possible field relation structures for the author field: “<Author><Year>”, which accounted for 54.29% of our test data; and “<Author><Title>”, which accounted for another 42.8%. However, the most common sequence structure of the reference styles was “<Author> <Year> <Title> <Journal> <Volume> <Issue> <Pages>”.

We also conducted experiments on the above 30 styles without additional knowledge editing and obtained an average accuracy for reference extraction of 85%. Specifically, the results show that the overall average accuracy for the MLA style is 88.20%, thereby demonstrating that our knowledge-based method is reliable for different kinds of unseen reference styles.

Thirteen types of punctuation are used in some fields of different reference styles, as shown by the analysis in Table 5. For example, a comma can be used in the author, volume, issue, and page fields. It is also used as the major field separator in the APA, IEEE, ACM, MISQ, and JMIS reference styles. In ISR, however, a period is the major field separator.

We use the templates in INFOMAP to represent the different types of punctuation used as field separators in various reference styles. Fig. 10 shows the sample templates for structure and punctuation in INFOMAP. For example, the highlighted template "[[Journal]]:[[Comma]]:[[Parentheses]]:[[Digit]]:[[Parentheses]]:[[Comma]]" indicates that an identified journal reference followed by a comma, a parenthesis, a digit number, another parenthesis and another comma, can be interpreted as an INSTANCE of Issue. The sequence structure of "Journal" and "Issue" in the above example is <Journal><Issue>.

5.3 Analysis of Reference Words

We analyzed the words in an article's title and author's name using a general English dictionary and a family name list with a citation agent. The general English dictionary contained 240,597 entries, and the family name list contained 71,475 entries collected from the Internet; the latter was obtained from "http://www.last-names.net/" and "http://genforum.genealogy.com/surnames/". Table 6 and Table 7, respectively, show that title words in the general dictionary account for 84.82% of the total, while title words not in the family name list account for 93.35% of the total. Table 8 and Table 9, respectively, show that authors' names not in the general dictionary account for 66.41% of the total, while authors' names in the family name list account for 23.63%.

5.4 Analysis of Types of Error

Here, we give some examples to illustrate the types of error that might be generated by our template-based RME method.

(1) Author

In the following example, the reference does not contain any author information.

"From the Thoughtful Business," Harvard Business Review (43:1), Jan/Feb 1965, pp. 42-48.

Another type of author error is caused by the ambiguity of the author name boundary, as in the following:

Lau, Hon-Shiang; Wingender, John R.; Hing-Ling Lau, Amy "ON ESTIMATING SKEWNESS IN STOCK RETURNS," Management Science (35:9), September 1989, pp. 1139-1142.

The system mistakenly extracted “Hing-Ling Lau” for the author name “Hing-Ling Lau, Amy”.

(2) Year

In the following example, our system reads page “1607” as the year “1607”

Schultz, Kenneth L.; Juran, David C. "Modeling and Worker Motivation in JIT Production Systems," Management Science (44:12), December 98 Part 1 of 2, pp. 1595-1607.

(3) Title

Errors in titles can be caused by punctuation problems (e.g., “:”) or HTML encoding problems (e.g., " R&D " stands for “R&D”). For example:

Helfat, Constance E. "Evolutionary trajectories in petroleum firm R&D," Management Science (40:12), December 1994, pp. 1720-1746.

(4) Journal

An error could be caused by missing journal information in the reference.

(5) Volume

An error could be caused by missing volume information in the reference.

(6) Issue

An error could be caused by missing issue information in the reference.

(7) Pages

Errors in page information are caused by pages not in single page format. For example:

"Managerial Economics," Sloan Management Review (20:3), Spring 1979, pp. 82, 1/6.

The system only extracted “82” for page information, but the correct reference should be “82, 1/6”.

In the future, we will conduct more extensive error analysis; and add more templates and knowledge to enhance our system.

6. Comparison of Machine Learning and Rule-based Models

Numerous works on extracting citation information from digital document references are reported in the literature [6, 9, 10, 11, 12, 13, 15, 17, 19, 20]. In this section, we compare related works on machine learning models and rule-based models.

First, machine learning approaches, such as Citeseer [11, 12, 15], take advantage of probabilistic estimation based on the training sets of tagged bibliographic data to boost performance. Citeseer’s citation parsing technique can identify titles and authors with more than 80% accuracy and page numbers with approximately 40% accuracy in all the different citation formats found on the Web [11]. Seymore et al. [19] used 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. [17] employed Conditional Random Fields (CRF) to extract various common fields from the headers and citations of research papers and achieved an overall word accuracy of 98.3% when extracting fields from paper headers. They used the Cora reference dataset [17], which contains 500 references covering 13 fields: author, title, editor, book_title, date, journal, volume, tech, institution, pages, location, publisher, and note. Peng et al. [17] achieved an overall word accuracy between 85.1% (HMM) and 95.37% (CRF) and an overall instance accuracy between 10% (HMM) and 77.33% (CRF) for paper references.

Second, rule-based models, such as those developed by Chowdhury [6] and Ding et al. [9], use a template mining approach to extract citations from digital documents. Ding et al. 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 processes references from tagged text in one style (e.g., references formatted in HTML), whereas our method processes references from plain text in more than six reference styles (cf. Table 1).

In contrast to the above approaches, which use small test datasets, our proposed template-based RME method for scholarly publications can extract reference information from 160,000 records in various styles with a high degree of precision (The overall average field accuracy was 92.39% for the six major styles tested, 99.31% for the MISQ style, and 85% for the other 30 randomly selected styles).

For the author field, machine learning approaches, such as Citeseer, yield 82% accuracy, whereas the accuracy of our approach is 90.18% for six reference styles.

We conducted an experiment on the same testbed (Cora dataset) in order to compare our approach with Citeseer. The experimental results show that the overall average field accuracy was 73.34%, while that of the author field was 87.40% on the Cora dataset, which contains 500 records with book and journal references. Specifically, the overall average field accuracy was 84.94% for the 166 selected journal reference records from the Cora dataset, and the author field accuracy was 93.37%. It should be noted that the Cora dataset comprises multiple styles that are difficult to differentiate. Thus, the comparison made here may not be completely fair.

The strength of our approach is that, unlike machine learning approaches, there is no need for training. However, its shortcoming is that it requires a domain expert to design and maintain a number of templates.

7. Conclusions and Future Research

RME is a challenging problem due to the diverse nature of reference styles. In this paper, we have proposed a template-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.

INFOMAP is more powerful than typical rule-based approaches, since it provides an integrated hierarchical template editing environment and a more flexible template matching engine. In addition, it achieves better accuracy by using a smaller number of labeled datasets than typical machine learning approaches. This reduction in the required number of annotated trained datasets is crucial in real-world scenarios, because it facilitates more rapid deployment of reference metadata extraction.

The major research direction for the future will be the integration of knowledge acquisition with machine learning techniques, which will enhance knowledge acquisition. We will integrate knowledge-based and machine learning approaches (such as the Maximum-Entropy Method (MEM), Hidden Markov Model (HMM), Conditional Random Fields (CRF), and Support Vector Machines (SVM)) to automate template generation, boost the performance of citation information extraction, and produce a more robust prototype that can deal with free style references as well as input that contains errors.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments, which have greatly improved the presentation of this paper. This research was supported in part by the thematic program of Academia Sinica under Grant AS 95ASIA02, the National Science Council under Grants NSC 95-2752-E-001-001-PAE and NSC 95-2416-H-002-047, the Research Center for Humanities and Social Sciences, Academia Sinica.

References

[1] E. Agichtein, 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.

[2] V.R. Borkar, K. Deshmukh, S. Sarawagi, Automatic Segmentation of Text into Structured Records, Proceedings of the ACM SIGMOD Conference, 2001, pp. 175-186.

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

[4] S.C. Bradford, Sources of Information on Specific Subjects, Engineering 137 (1934) 85-86.

[5] 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) (1999) 1209-1217.

[6] G. Chowdhury, Template mining for information extraction from digital documents, Library Trends 48 (1) (1999) 182-208.

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

[8] M.-Y. Day, C.-H. Lu, J.-T.D. Yang, G.-F. Chiou, C.-S. Ong, W.-L. Hsu, Designing an Ontology-based Intelligent Tutoring Agent with Instant Messaging, Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2005), Kaohsiung, Taiwan, 2005, pp. 318-320.

[9] 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, 1999, pp. 47-62.

[10] J. Geng, J. Yang, AutoBib: Automatic Extraction of Bibliographic Information on the Web, Proceedings of the International Database Engineering and Applications Symposium (IDEAS '04), 2004, pp. 193-204.

[11] C.L. Giles, K.D. Bollacker, S. Lawrence, CiteSeer: An Automatic Citation Indexing System, Proceedings of the Third ACM Conference on Digital Libraries (Digital Libraries 98) 1998, pp. 89-98.

[12] A. Goodrum, K. McCain, S. Lawrence, C. Giles, Scholarly publishing in the Internet age: a citation analysis of computer science literature, Information Processing & Management 37 (5) (2001) 661-675.

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

[14] H. Han, L. Giles, H. Zha, C. Li, K. Tsioutsiouliklis, Two supervised learning approaches for name disambiguation in author citations, Proceedings of the 4th ACM/IEEE-CS joint

conference on Digital libraries, 2004, pp. 296-305.
[15] S. Lawrence, C.L. Giles, K. Bollacker, Digital libraries and autonomous citation indexing, Computer 32 (6) (1999) 67-71.
[16] C.-H. Lu, S.-H. Wu, L.Y. Tu, W.-L. Hsu, The Design of An Intelligent Tutoring System Based on the Ontology of Procedural Knowledge, Proceedings of IEEE International Conference on Advanced Learning Technologies (ICALT 2004), Finland, 2004, pp. 525-529.
[17] F. Peng, 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.
[18] A. Sen, Metadata management: past, present and future, Decision Support Systems 37 (1) (2004) 151-173.
[19] K. Seymore, A. McCallum, R. Rosenfeld, Learning hidden Markov model structure for information extraction, AAAI-99 Workshop on Machine Learning for Information Extraction, 1999, pp. 37-42.
[20] 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, 2003, pp. 49-60.
[21] P. Vinkler, The origin and features of information referenced in pharmaceutical patents, Scientometrics 30 (1) (1994) 283-302.
[22] L.A. West, T.J. Hess, Metadata as a knowledge management tool: supporting intelligent agent and end user access to spatial data, Decision Support Systems 32 (3) (2002) 247-264.
[23] S.-H. Wu, M.-Y. Day, W.-L. Hsu, FAQ-centered Organizational Memory, Proceedings of the Knowledge Management and Organizational Memory workshop on the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001, pp. 112-120.
[24] S.-H. Wu, T.-H. Tsai, 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.

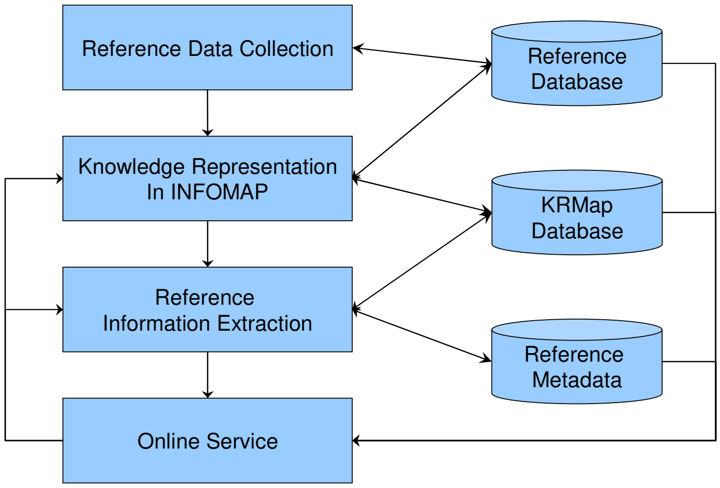

Fig. 1. The system framework of template-based RME

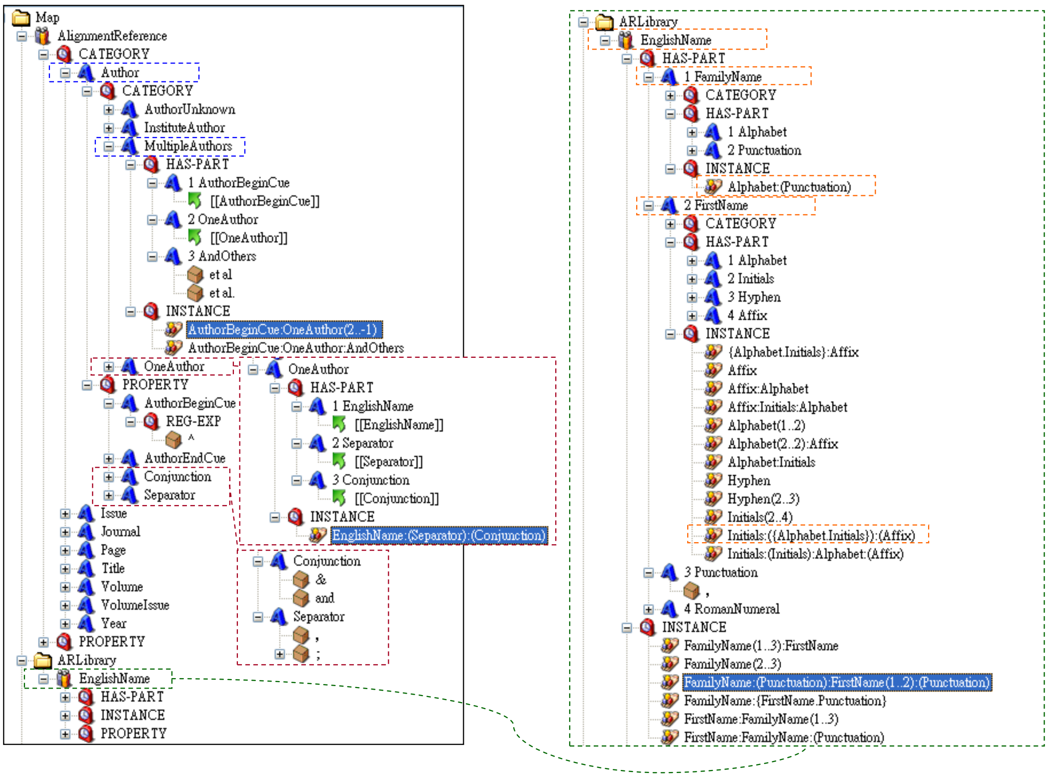

Fig. 2. An example of knowledge representation for template-based RME in INFOMAP

No.	Author	Title	Journal	Volume	Issue	Year	Pages
1	Davenport, T., DeLong, D., and Beers, M.	Successful knowledge management projects	Sloan management review	39	2	1998	43-57

Davenport, T., DeLong, D., and Beers, M. "Successful knowledge management projects," Sloan management review (39:2) 1998, pp 43-57.

No.	Author	Title	Journal	Volume	Issue	Year	Pages
1	Davenport, T.; DeLong, D.; and Beers, M. Successful	Successful knowledge management projects	Sloan management review	39	2	1998	43-57

1.Davenport, T.; DeLong, D.; and Beers, M. Successful knowledge management projects. Sloan management review, 39, 2, (1998), 43-57

Fig. 3. Results of reference information extraction

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.

Fig. 4. System input of template-based RME

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.								

Fig. 5. System Output of template-based RME

```
@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 }}
```

Fig. 6. System output of the BibTex Format

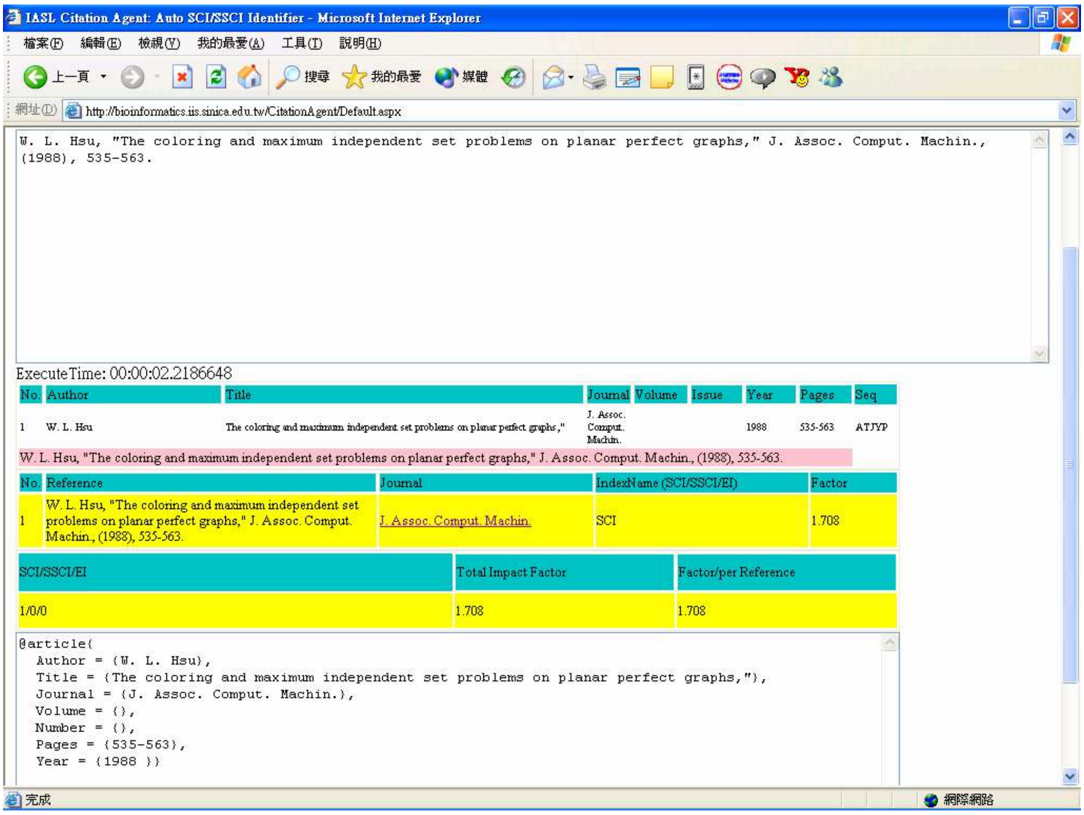

Fig. 7. The online service of template-based RME
http://bioinformatics.iis.sinica.edu.tw/CitationAgent/

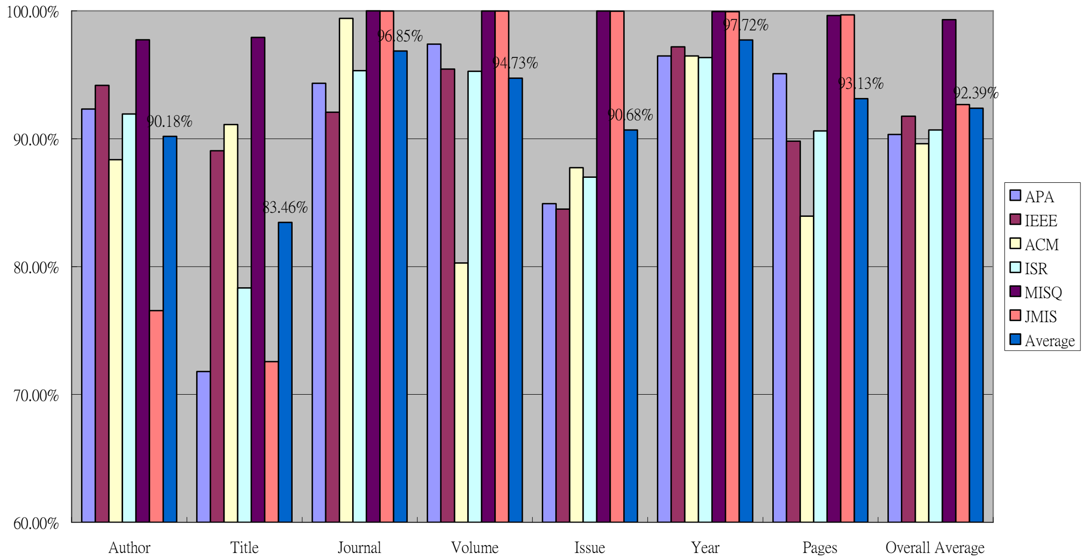

Fig. 8. Experimental results of citation extraction from six reference styles

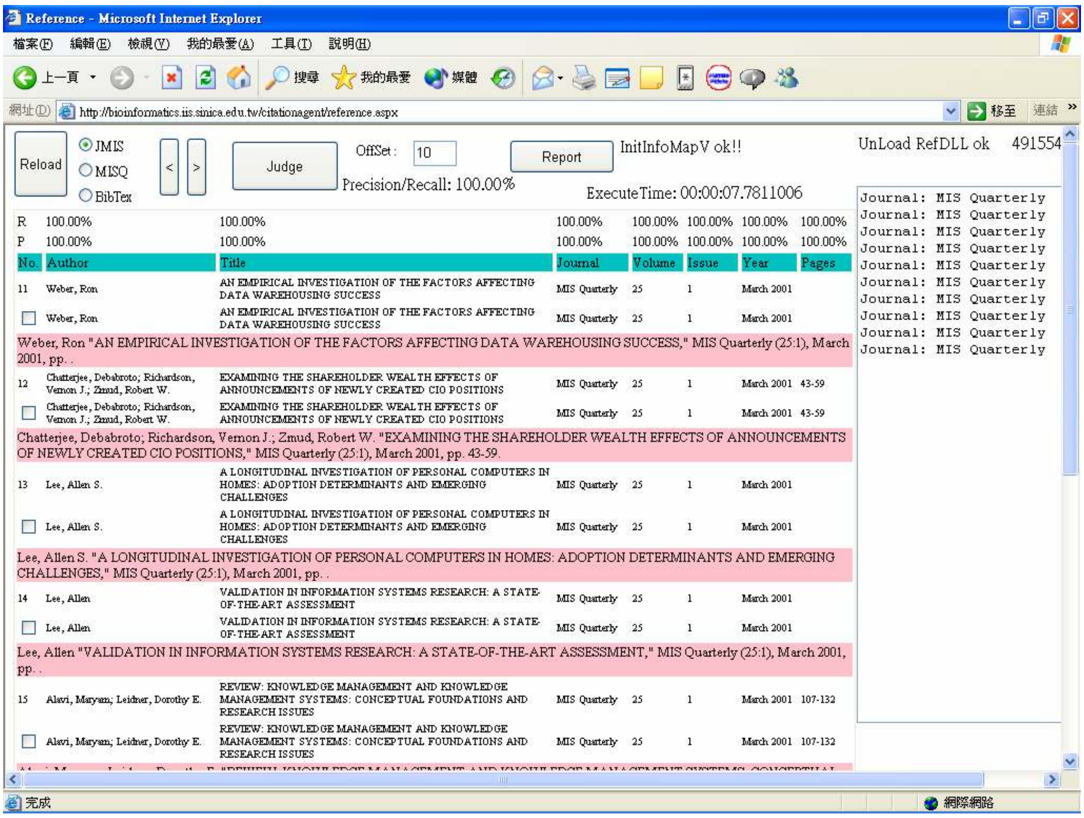

Fig. 9. Results of the selected reference database experiment

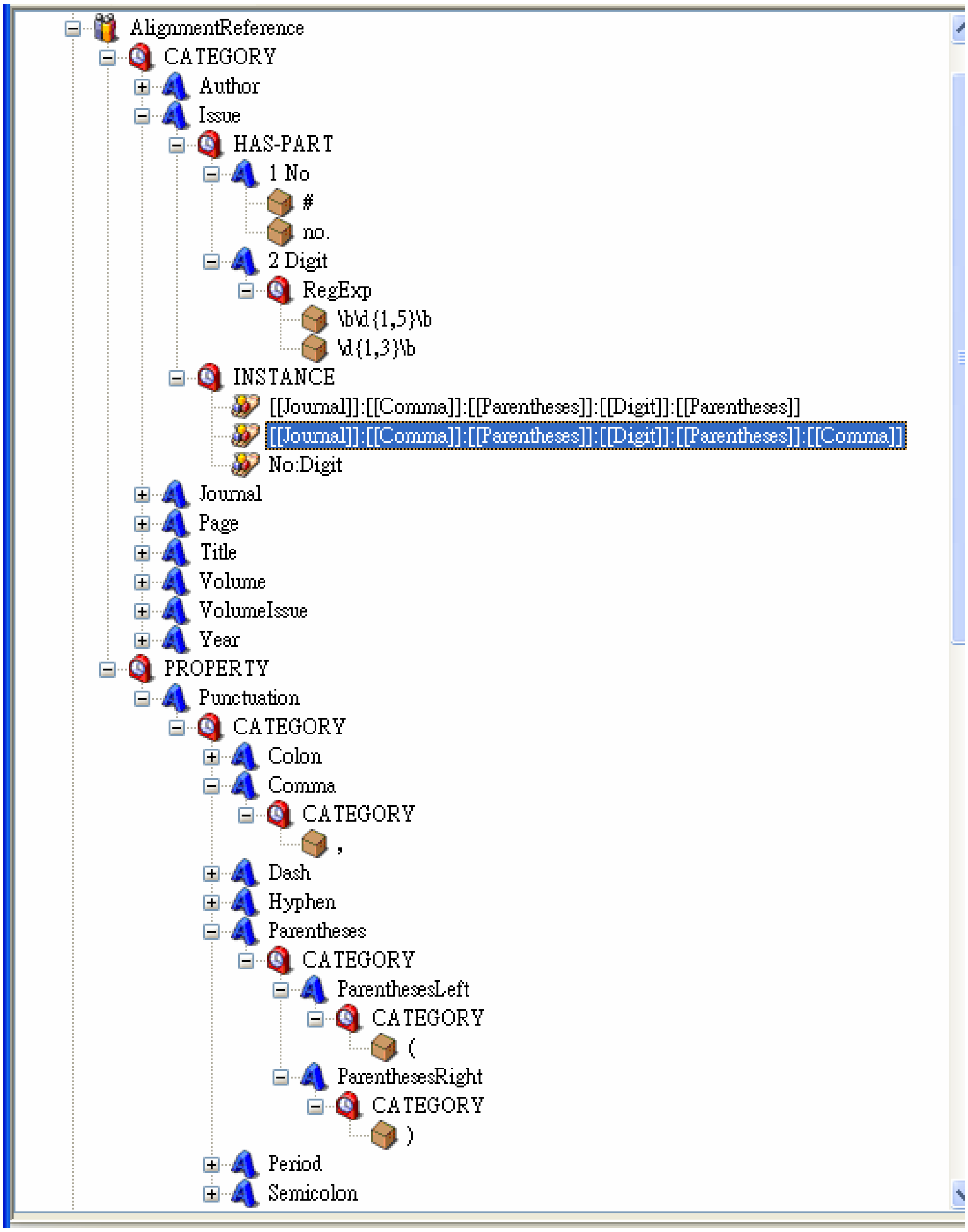

Fig. 10. Sample templates of reference structure and punctuation in INFOMAP

Table 1
Examples of different journal reference styles

Journal Reference styles	Reference style example
APA style	Davenport, T., DeLong, D., & Beers, M. (1998). Successful knowledge management projects. *Sloan management review, 39*(2), 43-57.
IEEE style	[1] T. Davenport, D. DeLong, and M. Beers, "Successful knowledge management projects," Sloan management review, vol. 39, no. 2, pp. 43-57, 1998.
ACM style	1. Davenport, T., DeLong, D. and Beers, M. 1998. Successful knowledge management projects. Sloan management review, 39 (2). 43-57.
MISQ style	Davenport, T., DeLong, D., and Beers, M. "Successful knowledge management projects," Sloan management review (39:2) 1998, pp 43-57.
JMIS style	1.Davenport, T.; DeLong, D.; and Beers, M. Successful knowledge management projects. Sloan management review, 39, 2 (1998), 43-57.
ISR style	Davenport, Thomas, David DeLong and Michael Beers, "Successful knowledge management projects," Sloan management review, 39, 2, (1998), 43-57.

Table 2
Experimental results of citation extraction from six reference styles

Accuracy	Author	Title	Journal	Volume	Issue	Year	Pages	Overall Average
APA	92.32%	71.80%	94.33%	97.39%	84.92%	96.48%	95.09%	90.33%
IEEE	94.17%	89.05%	92.07%	95.45%	84.49%	97.18%	89.81%	91.75%
ACM	88.36%	91.10%	99.41%	80.28%	87.73%	96.47%	83.95%	89.61%
ISR	91.93%	78.33%	95.32%	95.28%	87.00%	96.34%	90.61%	90.69%
MISQ	97.73%	97.92%	100.00%	99.99%	99.98%	99.94%	99.64%	99.31%
JMIS	76.55%	72.57%	99.99%	99.98%	99.97%	99.93%	99.69%	92.67%
Average	90.18%	83.46%	96.85%	94.73%	90.68%	97.72%	93.13%	92.39%

Table 3
Analysis of the structure of reference styles

Reference Style	Reference style example	The structure of reference style
APA Style	Culnan, M. (1978). An analysis of the information usage patterns of academics and practitioners in the computer field: A citation analysis of a national conference proceedings. *Information Processing and Management, 14*(6), 395 – 404.	Author : (2Period) : 0Space : 71PerenthesesLeft : Year : 72PerenthesesRight : 2Period : 0Space : Title : 2Period : 0Space : *Journal* : 1Comma : 0Space : Volume : 71PerenthesesLeft : Issue : 72PerenthesesRight : 1Comma : Page : 2Period
IEEE Style	S. Lawrence, C.L. Giles, and K.D. Bollacker, “Digital Libraries and Autonomous Citation Indexing,” *Computer*, vol. 32, no. 6, June 1999, pp. 67–71.	Author : 1Comma : 0Space : 9Quotation : Title : 1Comma : 9Quotation: 0Space : *Journal* : 1Comma : 0Space : vol. : Volume : 1Comma : 0Space : no. : Issue : 1Comma : 0Space : Year : 1Comma : 0Space : pp. : Page : 2Period
ACM Style	XU, J. AND CROFT, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inform. Syst. 18*, 1, 79–112.	Author : (2Period) : 0Space : Year : 2Period : 0Space : Title : 2Period : 0Space : *Journal* : (2Period) : 0Space : Volume : 1Comma : 0Space : Issue : 1Comma: 0Space: Page : 2Period
MISQ Style	Alavi, A., and Leidner, D. “Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues,” *MIS Quarterly* (25:1), 2001, pp. 107-136.	Author : (2Period): 0Space : 9Quotation : Title : 1Comma : 9Quotation: 0Space : *Journal* : 0Space : 71PerenthesesLeft : Volume : 4Colon : Issue : 72PerenthesesRight : 1Comma : 0Space : Year : 1Comma : 0Space : pp. : Page : 2Period
JMIS Style	Gold, A.H.; Malhotra, A.; and Segars, A.H. Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems, 18,* 1 (Summer 2001), 185–214.	Author : (2Period): 0Space : Title : 2Period: 0Space : *Journal* : 1Comma : 0Space : Volume : 4Colon : Issue : 0Space : 71PerenthesesLeft : Year : 72PerenthesesRight :1Comma : 0Space : Page : 2Period
ISR Style	Straub, Detmar, Richard T. Watson. 2001. Transformational issues in researching IS and Net-enabled organizations. *Inform. Systems Res.* **12**(4) 337 – 345.	Author : (2Period): 0Space : Year : : 2Period: 0Space : Title : 2Period: 0Space : *Journal* : (2Period) : 0Space : Volume : 71PerenthesesLeft : Issue : 72PerenthesesRight : 0Space : Page : 2Period

Table 4
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%

Table 5
Analysis of reference punctuation

Punctuation	Fields	Styles
1 Comma ,	Author ; Volume; Issue, Page; Separator*	APA, IEEE, ACM, MISQ, JMIS
2 Period .	Author; Page; Separator*	APA, IEEE, ACM, MISQ, JMIS, ISR
3 Semicolon ;	Author Separator*	JMIS
4 Colon :	Volume Issue Page	MISQ
5 Dash -- —	Volume Issue Page	APA, IEEE, ACM, MISQ, JMIS, ISR
6 Hyphen -	Volume Issue Page	APA, IEEE, ACM, MISQ, JMIS, ISR
7 Parentheses () （）	Year Volume Issue Page	APA, MISQ, JMIS, ISR
8 Brackets []	Serial	IEEE
9 Quotation Marks " " “ ”	Title	IEEE, MISQ
10 Ellipsis …	Title	APA, IEEE, ACM, MISQ, JMIS, ISR
11 Question Mark ?	Title	APA, IEEE, ACM, MISQ, JMIS, ISR
12 Exclamation Point !	Title	APA, IEEE, ACM, MISQ, JMIS, ISR
13 Apostrophe '	Author Title	APA, IEEE, ACM, MISQ, JMIS, ISR

Table 6
Analysis of Article Title Words and Dictionary Words

Article Title Word	Number #	Percentage %
Title Word in Dictionary	**1171120**	**84.82%**
Title Word no in Dictionary	209577	15.18%
Title Word Total	1380697	100.00%

Table 7
Analysis of Article Title Words and Family Name List

Article Title Word	Number #	Percentage %
Title Word in Family Name List	91769	6.65%
Title Word not in Family Name List	**1288928**	**93.35%**
Title Word Total	1380697	100.00%

Table 8
Analysis of Author Words and Dictionary Words

Author Word	Number #	Percentage %
Author Word in English Word List	304655	33.59%
Author Word not in English Word List	**602453**	**66.41%**
Author Word Total	907108	100.00%

Table 9
Analysis of Author Words and Family Name List

Author Word	Number #	Percentage %
Author Word in Family Name List	**214366**	**23.63%**
Author Word no in Family Name List	692742	76.37%
Author Word Total	907108	100.00%