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@iis.sinica.edu.tw

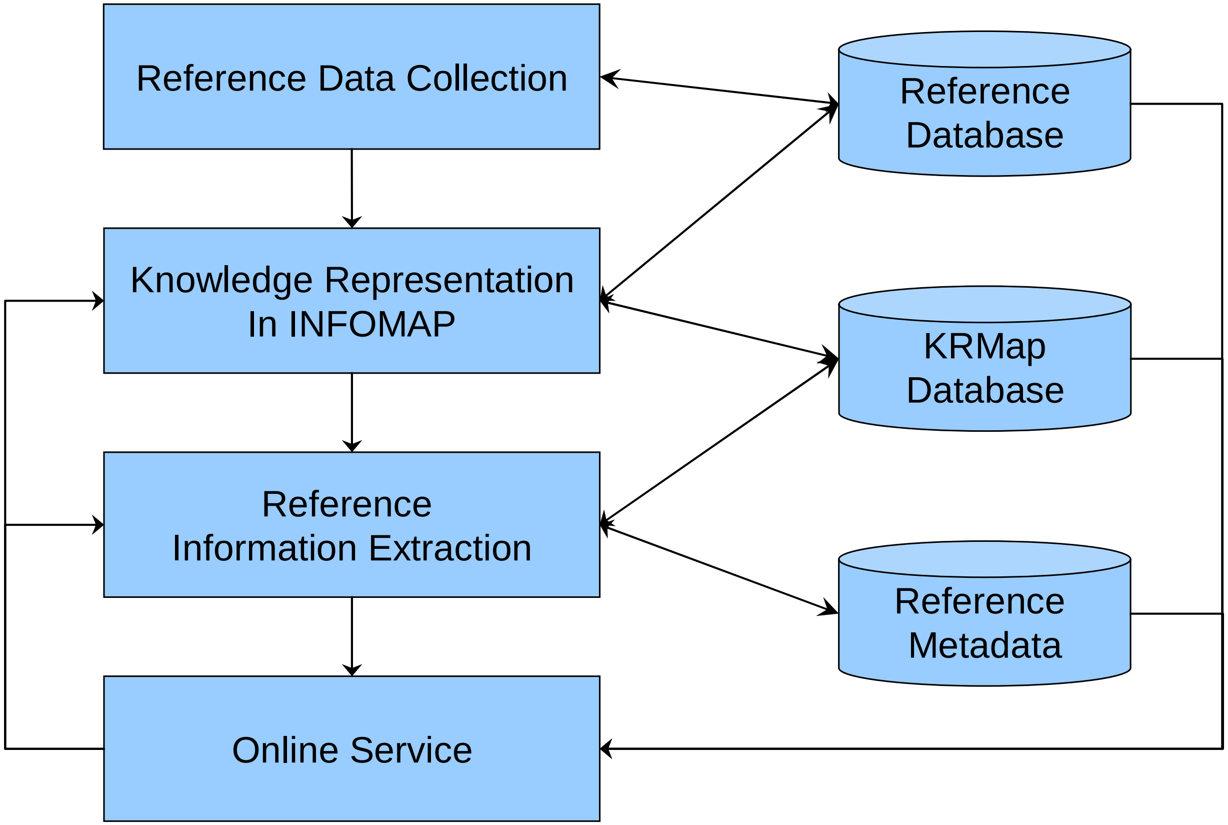
Outline

- Introduction
- Proposed Approach
- Experimental Results and Discussion
- Related Works
- Conclusions and Future Research

Introduction

- Integration of the 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.
- We propose a knowledge-based approach to literature mining and focus on reference metadata extraction methods for scholarly publications.
  - INFOMAP: ontological knowledge representation framework
  - Automatically extract the reference metadata.

Proposed Approach



Reference Data Collection

Knowledge Representation
In INFOMAP

Reference
Information Extraction

Online Service

Reference
Database

KRMap
Database

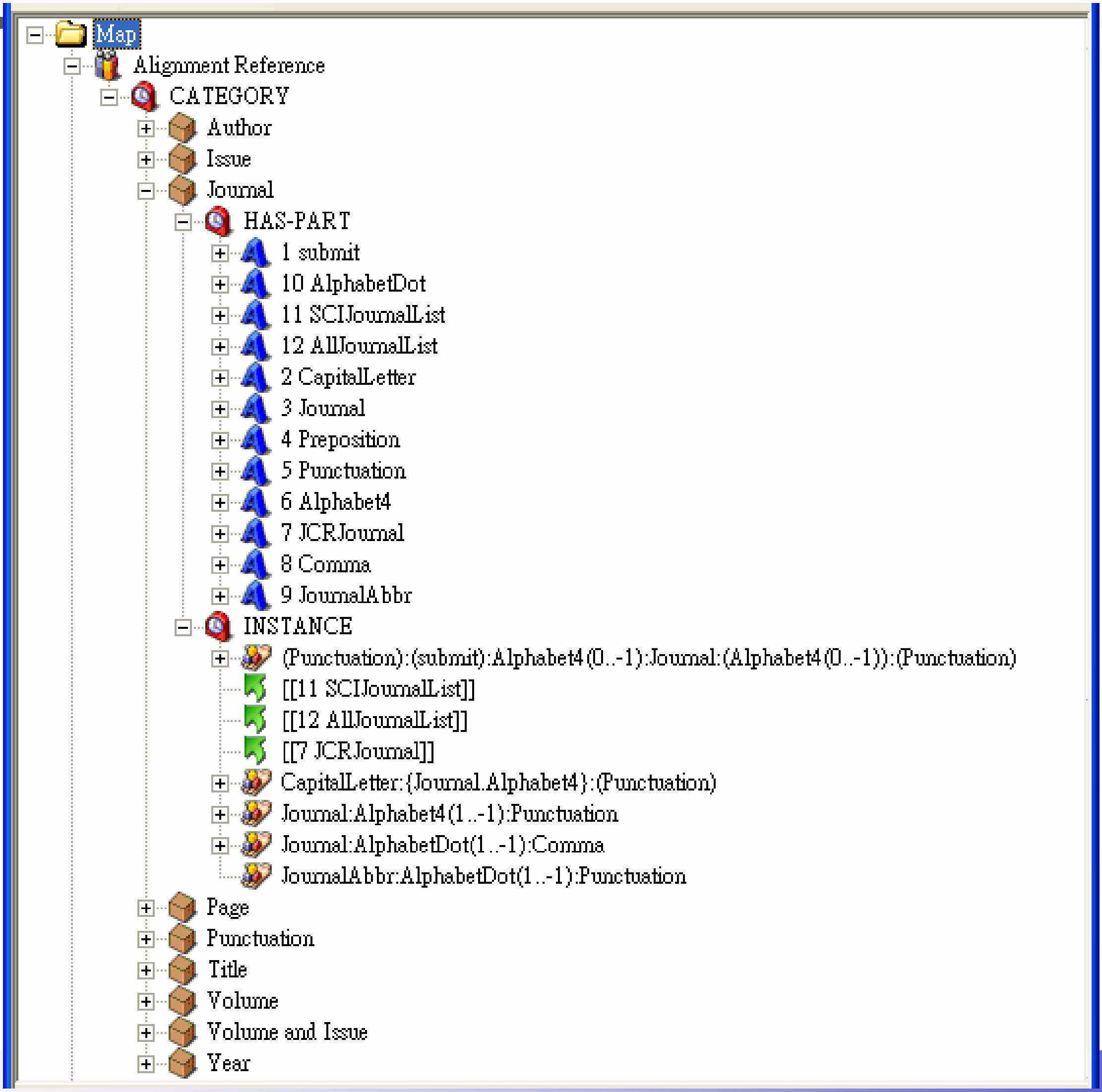
Reference
Metadata

Reference Data Collection

- Journal Spider (journal agent)
  - collect journal data from the Journal Citation Reports (JCR) indexed by the ISI and digital libraries on the Web.
- Citation data source
  - ISI web of science
  - DBLP
  - Citeseer
  - PubMed

Phase 2

Knowledge Representation in INFOMAP



```
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
```

INFOMAP

- INFOMAP as ontological knowledge representation framework
  - extracts important citation concepts from a natural language text.
- Feature of INFOMAP
  - represent and match complicated template structures
    - hierarchical matching
    - regular expressions
    - semantic template matching
    - frame (non-linear relations) matching
    - 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.

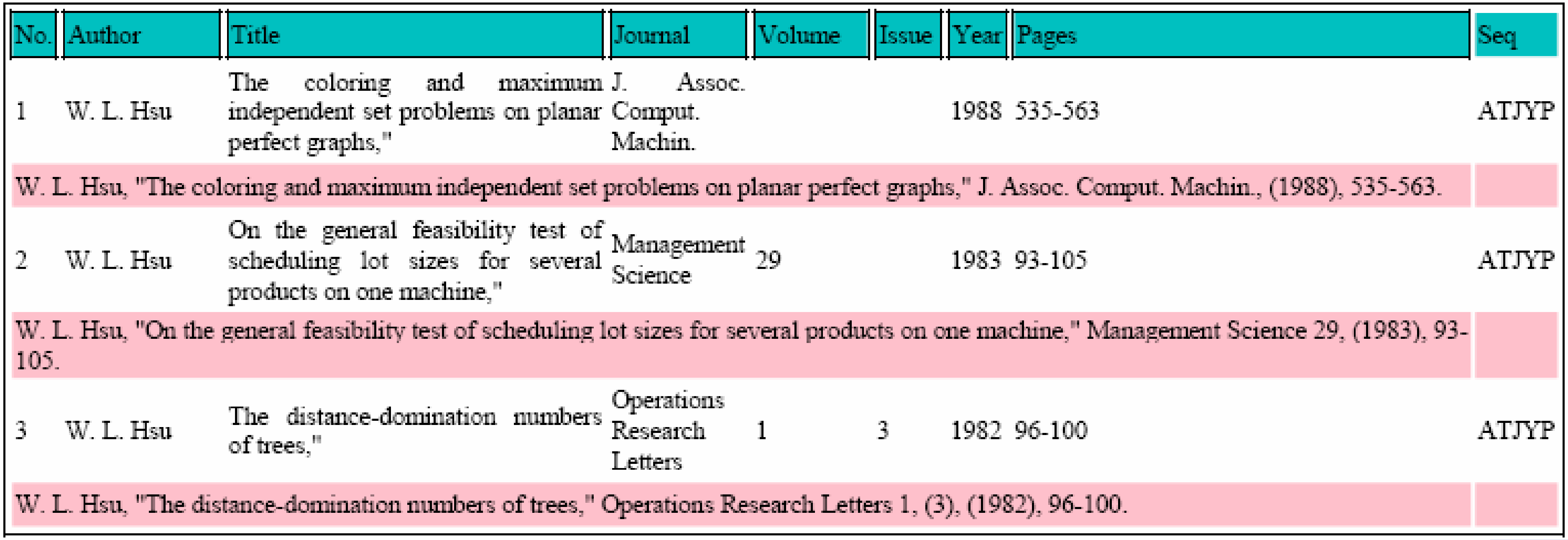
Phase 3

Reference Metadata Extraction

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.

Table 1. Examples of different journal reference styles

Knowledge-based Reference Metadata Extraction - Online Service



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.								

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

Citation Extraction From Text to BixTex

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

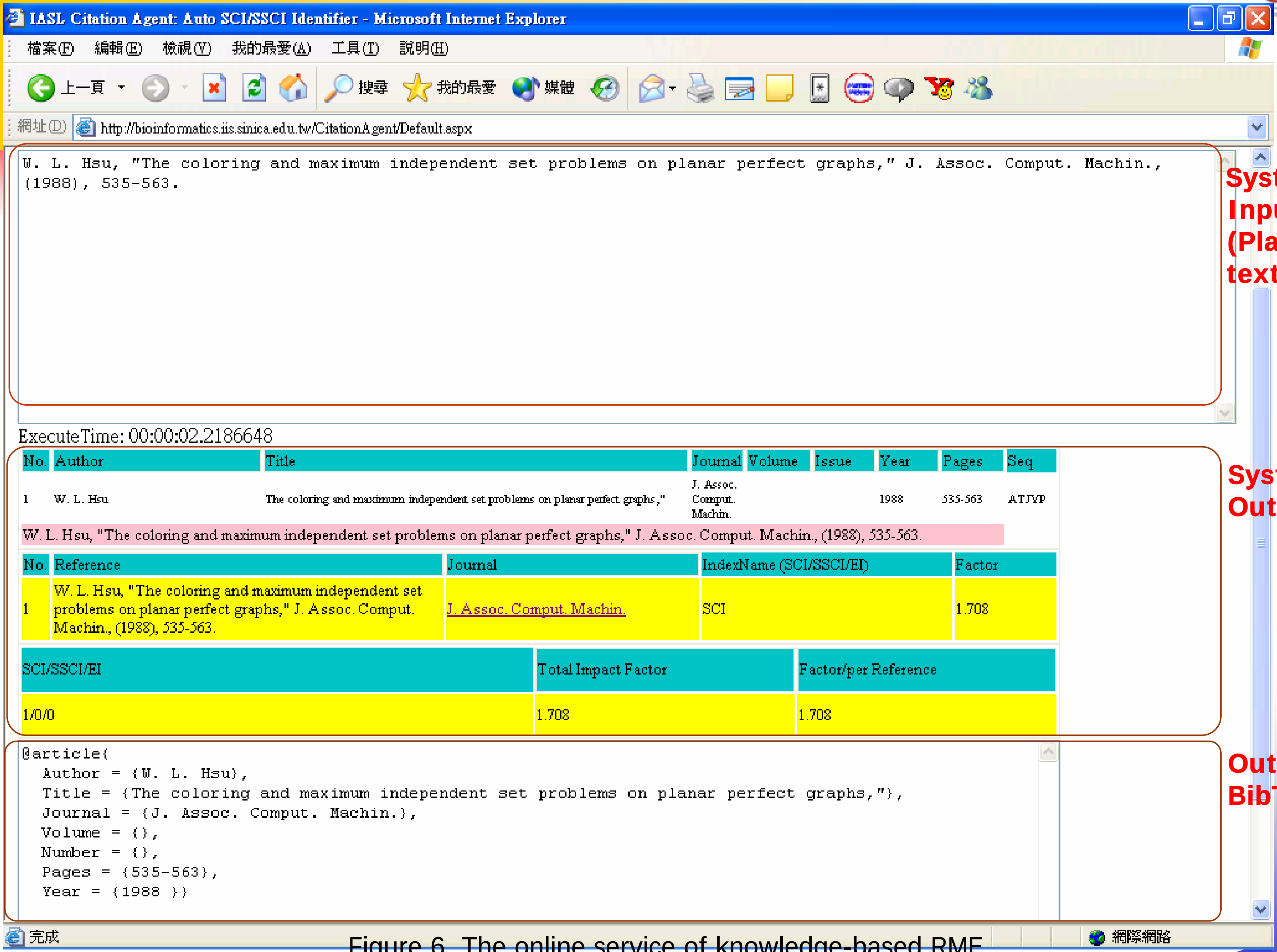
W. L. Hsu, "On the general feasibility test of scheduling lot sizes for several products on one machine," Management Science 29, (1983), 93-105.

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

Figure 3. The system input of knowledge-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 }}
```

Figure 5. The system output of BibTex Format


IASL Citation Agent: Auto SCI/SSCI Identifier - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁
搜尋
我的最愛
媒體
網址(D)
完成
網際網路

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.
```

System Input (Plain text)

ExecuteTime: 00:00:02.2186648

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.

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

System Output

```
@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 }}
```

Output BibTex

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

Experimental Results and Discussion

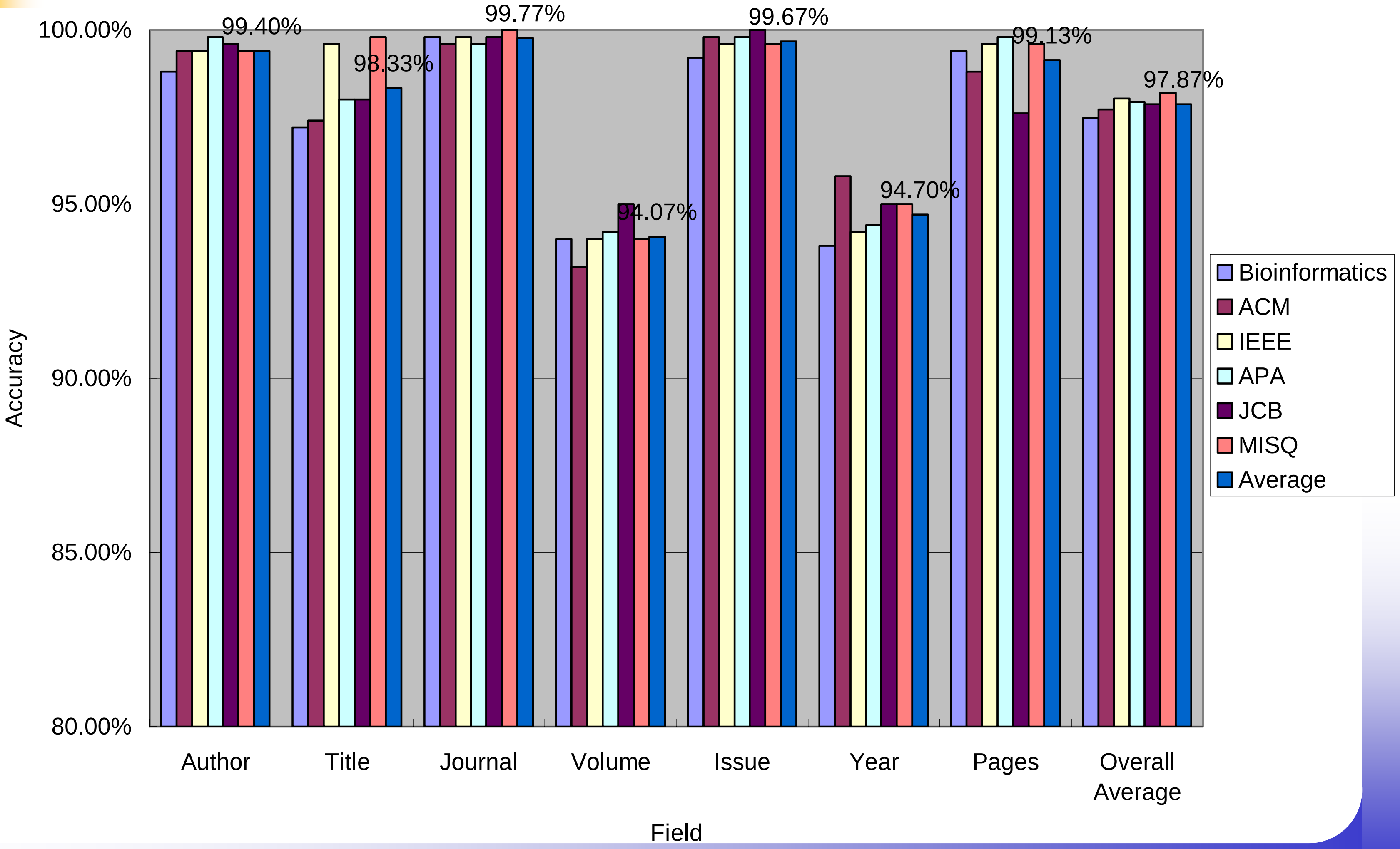
- Experimental data
  - We used EndNote to collect Bioinformatics citation data for 2004 from PubMed.
  - A total of 907 bibliography records were collected from PubMed digital libraries on the Web.
  - Reference testing data was generated for each of the six reference styles (BIOI, ACM, IEEE, APA, MISQ, and JCB).
  - Randomly selected 500 records for testing from each of the six reference styles.

Accuracy of Citation Extraction Definition:

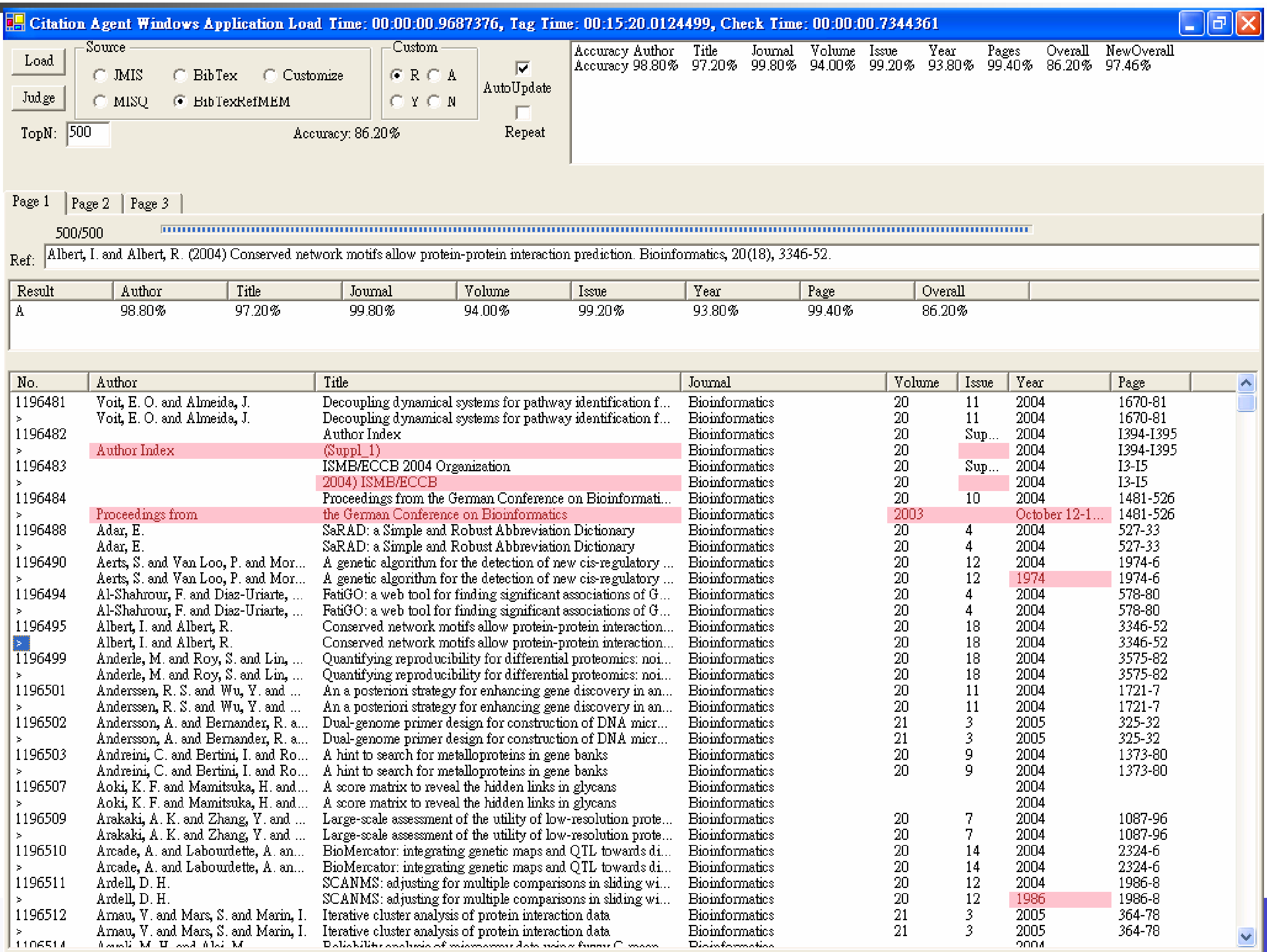
- We consider a field to be correctly extracted only when the field values in the reference testing data are correctly extracted.
- The accuracy of citation extraction is defined as follows:

$$Accuracy = \frac{Number\ of\ \ correctly\ extracted\ \ fields}{Total\ number\ of\ \ fields}$$

Experimental results of citation extraction from six reference styles



Example Results



Citation Agent Windows Application Load Time: 00:00:00.9687376, Tag Time: 00:15:20.0124499, Check Time: 00:00:00.7344361

Source: ○ JMIS ○ BibTex ○ Customize ○ MISQ ● BibTexRefMEM

Custom: ● R ○ A ○ Y ○ N

☑ AutoUpdate

□ Repeat

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	I394-I395
>	Author Index	(Suppl_1)	Bioinformatics	20		2004	I394-I395
1196483		ISMB/ECCB 2004 Organization	Bioinformatics	20	Sup...	2004	I3-I5
>		2004) ISMB/ECCB	Bioinformatics	20		2004	I3-I5
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	Anderssen, R. S. and Wu, Y. and ...	An a posteriori strategy for enhancing gene discovery in an...	Bioinformatics	20	11	2004	1721-7
>	Anderssen, 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 hint to search for metalloproteins in gene banks	Bioinformatics	20	9	2004	1373-80
>	Andreini, C. and Bertini, I. and Ro...	A hint 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

Analysis of the structure of reference styles

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%

Related Works

- Machine learning approaches
  - Citeseer [8, 9, 12] take 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.
  - Seymore et al. [15] use the Hidden Markov Model (HMM) to extract important fields from the headers of computer science research papers
    - Achieve 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.
    - 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.

Related Works (Cont.)

- Rule-based models
  - 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.

Comparison with related works

- Knowledge-based approach
  - 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 listed in Table 1
  - **98.20%** for the MISQ style
  - **87%** for other 30 randomly selected styles

Conclusions

- Citation extraction is a challenging problem
  - The diverse nature of reference styles
- We have proposed a knowledge-based citation extraction 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 major reference styles.

Future Research

- Integrate the ontological and the machine learning approaches to boost the performance of citation information extraction
  - Maximum-Entropy Method (MEM)
  - Hidden Markov Model (HMM)
  - Conditional Random Fields (CRF)
  - Support Vector Machines (SVM)

Q & A

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@iis.sinica.edu.tw