FireCite: Lightweight real-time reference string extraction from web pages

Ching Hoi Andy Hong Jesse Prabawa Gozali Min-Yen Kan

School of Computing

National University of Singapore

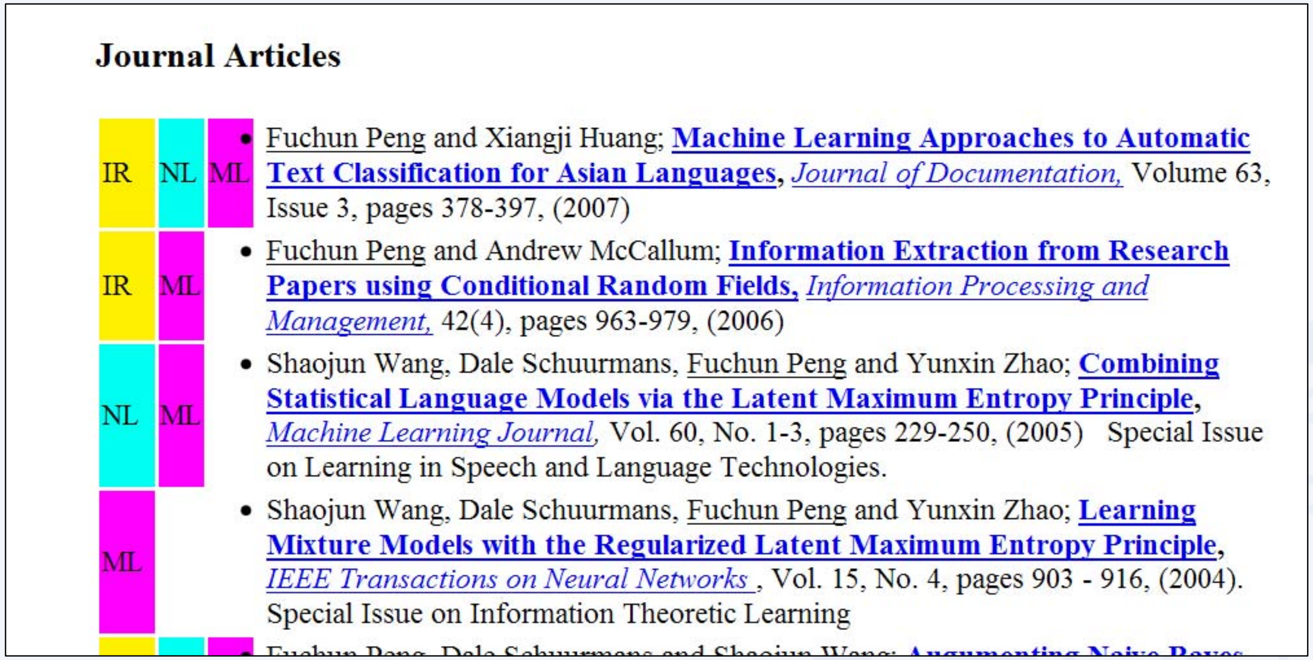
Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion

Introduction: The Problem

- Recognition and parsing of references found on the Internet

- Criteria:
  - Accurate
  - Fast



Journal Articles

IR NL ML
- Fuchun Peng and Xiangji Huang; **Machine Learning Approaches to Automatic Text Classification for Asian Languages**, *Journal of Documentation,* Volume 63, Issue 3, pages 378-397, (2007)

IR ML
- Fuchun Peng and Andrew McCallum; **Information Extraction from Research Papers using Conditional Random Fields,** *Information Processing and Management,* 42(4), pages 963-979, (2006)

NL ML
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Combining Statistical Language Models via the Latent Maximum Entropy Principle**, *Machine Learning Journal,* Vol. 60, No. 1-3, pages 229-250, (2005) Special Issue on Learning in Speech and Language Technologies.

ML
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Learning Mixture Models with the Regularized Latent Maximum Entropy Principle**, *IEEE Transactions on Neural Networks* , Vol. 15, No. 4, pages 903 - 916, (2004). Special Issue on Information Theoretic Learning

- Fuchun Peng, Dale Schuurmans and Shaojun Wang; **Augmenting Naive Bayes**

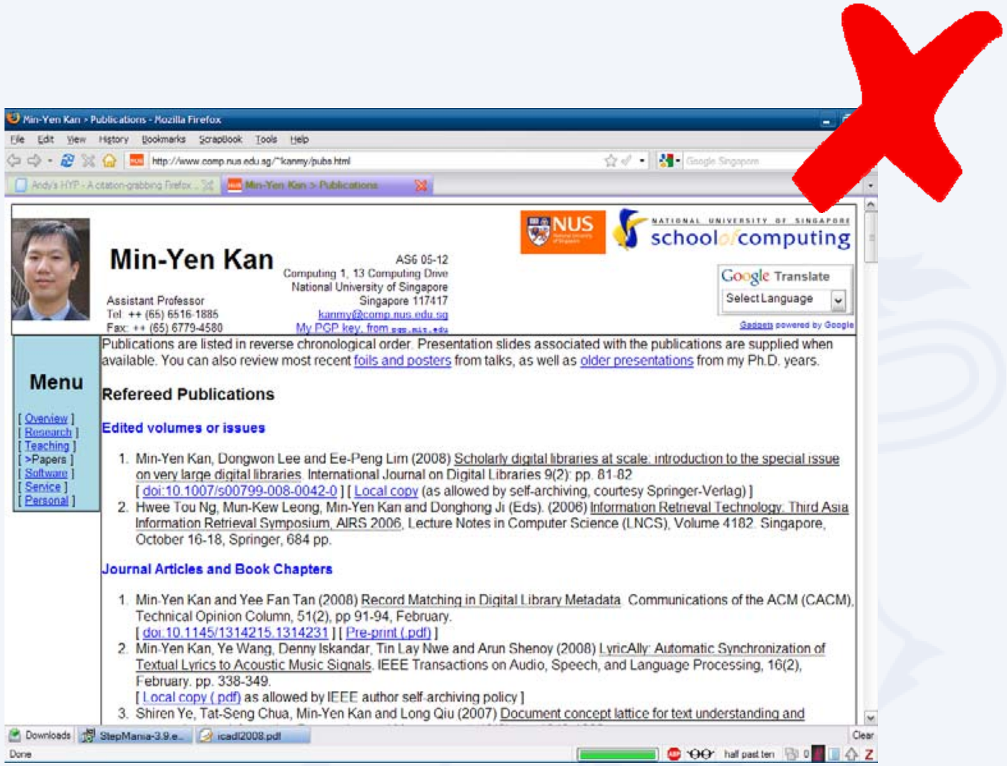
Introduction: Related Work

zotero

MENDELEY
RESEARCH NETWORKS BETA 0.6

✓ Google scholar beta

✓ CiteSeerx beta



Min-Yen Kan > Publications - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
http://www.comp.nus.edu.sg/~kanmy/pubs.html
Google Singapore
Andy's HYP - A citation-grabbing Firefox ...
Min-Yen Kan > Publications
NUS
NATIONAL UNIVERSITY OF SINGAPORE
school of computing
Google Translate
Select Language
Gadgets powered by Google
Downloads
StepMania-3.9.e...
icadl2008.pdf
Clear
Done
half past ten

Min-Yen Kan

Assistant Professor
Tel: ++ (65) 6516-1885
Fax: ++ (65) 6779-4580

AS6 05-12
Computing 1, 13 Computing Drive
National University of Singapore
Singapore 117417
kanmy@comp.nus.edu.sg
My PGP key, from pgp.mit.edu

Menu

[Overview]
[Research]
[Teaching]
[>Papers]
[Software]
[Service]
[Personal]

Publications are listed in reverse chronological order. Presentation slides associated with the publications are supplied when available. You can also review most recent foils and posters from talks, as well as older presentations from my Ph.D. years.

Refereed Publications

Edited volumes or issues

1. Min-Yen Kan, Dongwon Lee and Ee-Peng Lim (2008) Scholarly digital libraries at scale: introduction to the special issue on very large digital libraries. International Journal on Digital Libraries 9(2): pp. 81-82
   [doi:10.1007/s00799-008-0042-0] [Local copy (as allowed by self-archiving, courtesy Springer-Verlag)]
2. Hwee Tou Ng, Mun-Kew Leong, Min-Yen Kan and Donghong Ji (Eds). (2006) Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006, Lecture Notes in Computer Science (LNCS), Volume 4182. Singapore, October 16-18, Springer, 684 pp.

Journal Articles and Book Chapters

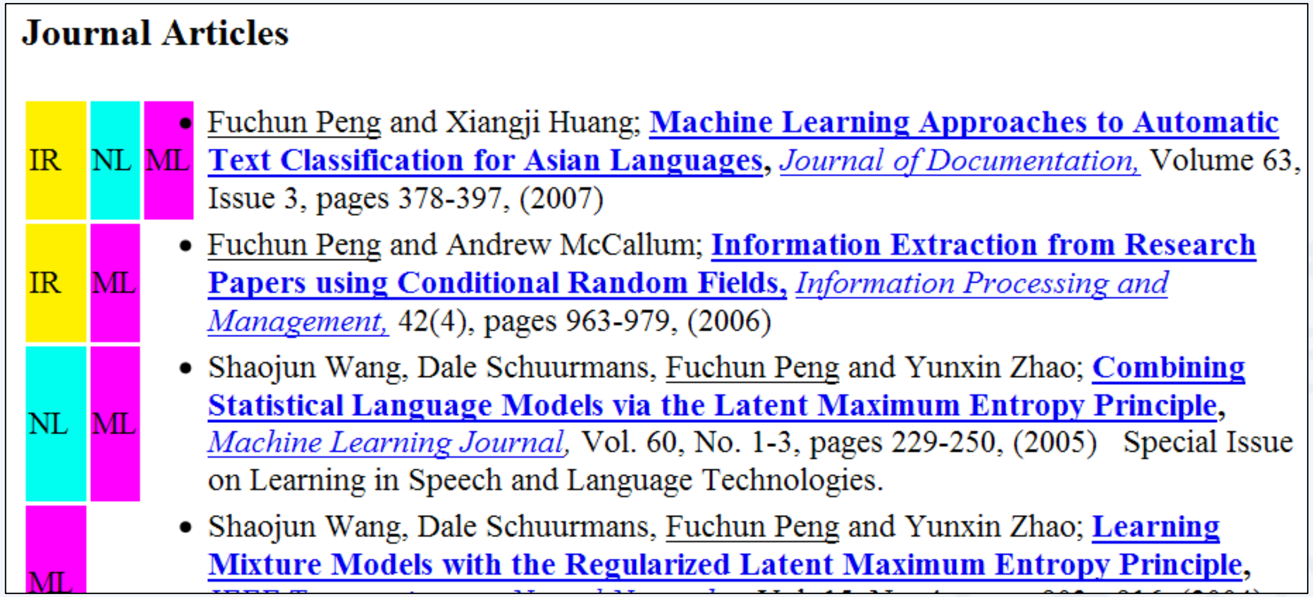
1. Min-Yen Kan and Yee Fan Tan (2008) Record Matching in Digital Library Metadata. Communications of the ACM (CACM), Technical Opinion Column, 51(2), pp 91-94, February.
   [doi:10.1145/1314215.1314231] [Pre-print (.pdf)]
2. Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe and Arun Shenoy (2008) LyricAlly: Automatic Synchronization of Textual Lyrics to Acoustic Music Signals. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), February. pp. 338-349.
   [Local copy (.pdf) as allowed by IEEE author self-archiving policy]
3. Shiren Ye, Tat-Seng Chua, Min-Yen Kan and Long Qiu (2007) Document concept lattice for text understanding and

Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion

Reference String Recognition: Definition

- Are there reference strings?
- Where are the reference strings?



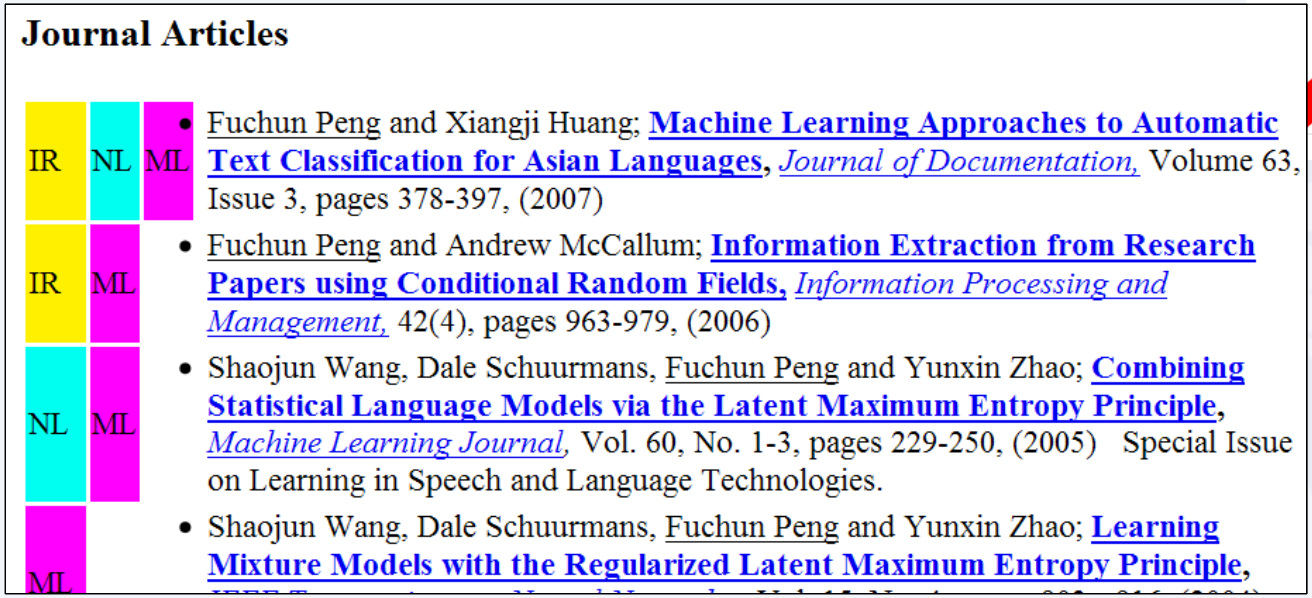
Journal Articles

- IR NL ML — Fuchun Peng and Xiangji Huang; **Machine Learning Approaches to Automatic Text Classification for Asian Languages**, *Journal of Documentation,* Volume 63, Issue 3, pages 378-397, (2007)
- IR ML — Fuchun Peng and Andrew McCallum; **Information Extraction from Research Papers using Conditional Random Fields,** *Information Processing and Management,* 42(4), pages 963-979, (2006)
- NL ML — Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Combining Statistical Language Models via the Latent Maximum Entropy Principle**, *Machine Learning Journal,* Vol. 60, No. 1-3, pages 229-250, (2005) Special Issue on Learning in Speech and Language Technologies.
- ML — Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Learning Mixture Models with the Regularized Latent Maximum Entropy Principle**,

Reference String Recognition: Methodology

1. Web page exclusion

- Keyword + URL Matching



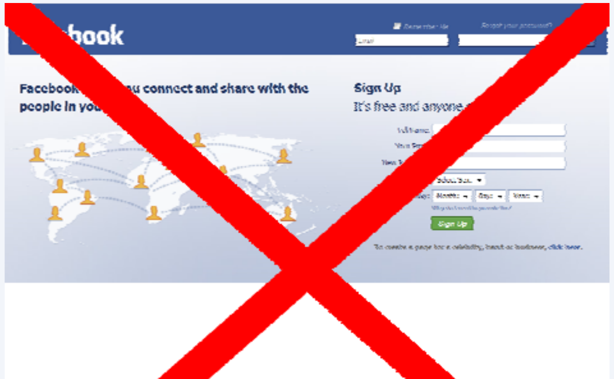
Journal Articles

IR NL ML
- Fuchun Peng and Xiangji Huang; **Machine Learning Approaches to Automatic Text Classification for Asian Languages**, *Journal of Documentation,* Volume 63, Issue 3, pages 378-397, (2007)

IR ML
- Fuchun Peng and Andrew McCallum; **Information Extraction from Research Papers using Conditional Random Fields,** *Information Processing and Management,* 42(4), pages 963-979, (2006)

NL ML
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Combining Statistical Language Models via the Latent Maximum Entropy Principle**, *Machine Learning Journal,* Vol. 60, No. 1-3, pages 229-250, (2005) Special Issue on Learning in Speech and Language Technologies.

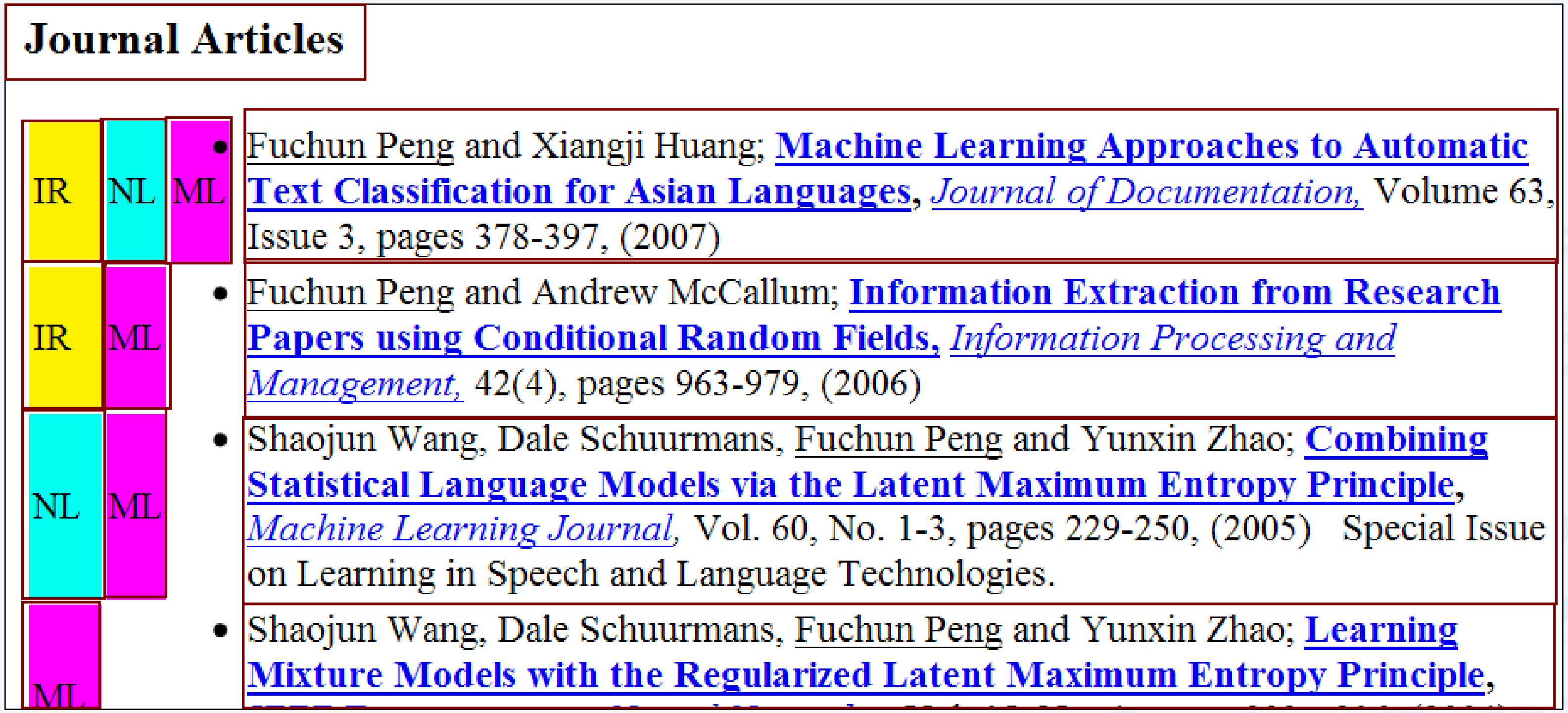
ML
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Learning Mixture Models with the Regularized Latent Maximum Entropy Principle**,


book
Facebook ... u connect and share with the people in you...
Sign Up
It's free and anyone c...

Reference String Recognition: Methodology

2. Splitting

- Split web page text into segments
- GOAL: Each segment to contain at most 1 reference string, and each reference string to exist in only 1 segment.



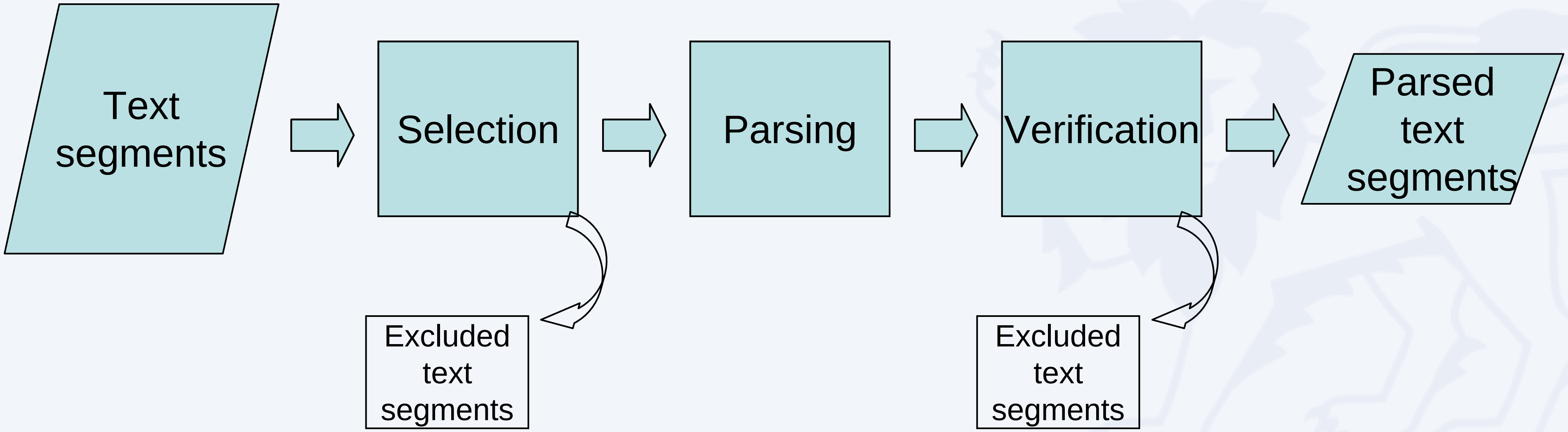
Journal Articles

- Fuchun Peng and Xiangji Huang; **Machine Learning Approaches to Automatic Text Classification for Asian Languages**, *Journal of Documentation,* Volume 63, Issue 3, pages 378-397, (2007)
- Fuchun Peng and Andrew McCallum; **Information Extraction from Research Papers using Conditional Random Fields,** *Information Processing and Management,* 42(4), pages 963-979, (2006)
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Combining Statistical Language Models via the Latent Maximum Entropy Principle**, *Machine Learning Journal,* Vol. 60, No. 1-3, pages 229-250, (2005) Special Issue on Learning in Speech and Language Technologies.
- Shaojun Wang, Dale Schuurmans, Fuchun Peng and Yunxin Zhao; **Learning Mixture Models with the Regularized Latent Maximum Entropy Principle**,

Reference String Recognition: Methodology

3. **Selection**
   - Token length and word length of segment

4. **Verification**
   - Reject segments that do not have a title or authors



Reference String Recognition: Evaluation

- Test set: 40 staff homepages from 4 universities
- Reference strings found: 364/379 (96.0%)
- False positives: 269 (42.5%)

System	Precision	Recall	F1-measure
FireCite (All 40 pages)	0.575	0.960	0.719
FireCite (Only 20 pages with reference strings)	0.655	0.960	0.779

Outline

- Introduction
- Related Work
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion

Reference String Parsing: Definition

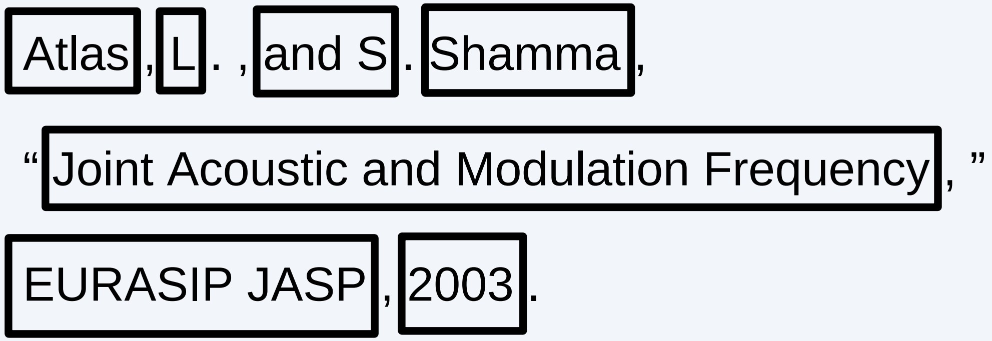
- Purpose
  - Store reference according to metadata fields
  - Assist reference string recognition

- Only identify authors, title, date

Jesse Prabawa Gozali and Min-Yen Kan (2007) A Rich OPAC User Interface with AJAX, In Proceedings of the Joint Conference on Digital Libraries (JCDL '07). Vancouver, Canada, June, pp. 329-330. Short paper.

Reference String Parsing: Methodology

- **Tokenising**



Atlas, L. , and S. Shamma,

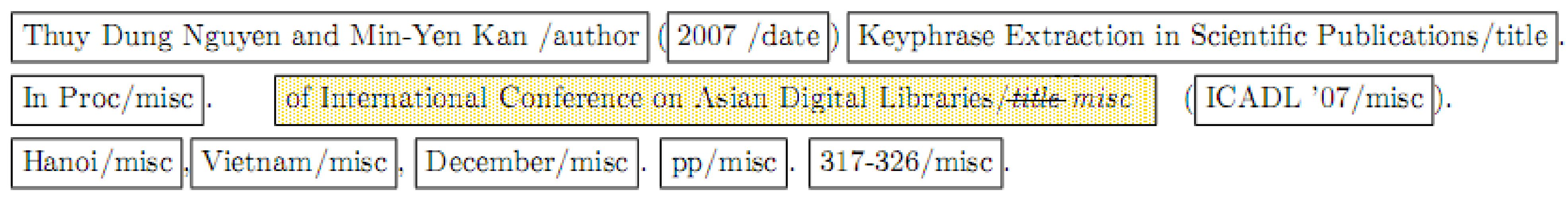
“Joint Acoustic and Modulation Frequency, ”

EURASIP JASP, 2003.

- Advantages
  - Reduce number of computations
  - Allow information-richer learning features

Reference String Parsing: Methodology

- **Labelling**
  - J48 decision tree classifier
  - CORA corpus (500 reference strings) as training corpus
- **Repairs**
  - "Title" and "Authors" fields are contiguous



Thuy Dung Nguyen and Min-Yen Kan /author (2007 /date) Keyphrase Extraction in Scientific Publications/title.

In Proc/misc. of International Conference on Asian Digital Libraries/~~title~~ *misc* (ICADL '07/misc).

Hanoi/misc, Vietnam/misc, December/misc. pp/misc. 317-326/misc.

Reference String Parsing: Evaluation

6 Faculty Staff Publication Pages

Page (No. of ref. strings)	Token-level F-measure			
	Title	Authors	Date	All Tokens
A (72)	0.902	0.893	0.988	0.708
B (52)	0.953	0.957	0.990	0.960
C (29)	0.684	0.304	0.774	0.651
D (68)	0.753	0.968	0.889	0.917
E (8)	0.692	0.875	1.000	0.889
F (45)	0.847	1.000	0.989	0.966
Overall (274)	0.836	0.916	0.948	0.878

Reference String Parsing: Evaluation

FLUX-CiM Computer Science Dataset (300 citations)

System Name	Field-level F-measure			
	Title	Authors	Date	Overall
FireCite	0.92	0.96	0.97	0.94
ParsCit	0.96	0.99	0.97	0.94
FLUX-CiM	0.93	0.95	0.98	0.97

Reference String Parsing: Evaluation

Parser	Classifier Type	Size of classifier model (KB)	Size of dictionary (KB)
FireCite	Decision Tree	12.6	0.0
FLUX-CiM	Knowledge-Based	>786 (estimated)	0.0
ParsCit	Conditional Random Fields	7339	1722

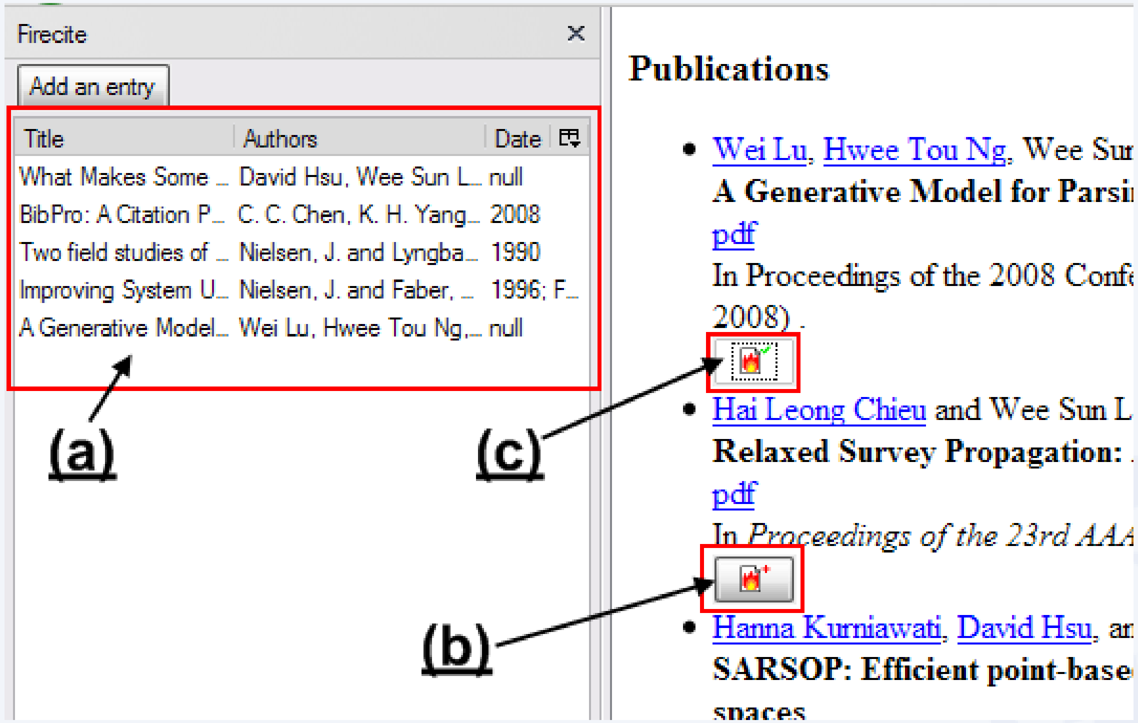
Reference String Parsing: Evaluation

Page Type	Time taken (milliseconds)		
	Minimum	Maximum	Average
Pages with reference strings	90	544	192
Pages without reference strings	6	222	74
All pages	6	544	133

Outline

- Introduction
- Reference String Recognition
- Reference String Parsing
- Firefox Extension
- Conclusion

Extension: Demo



Title	Authors	Date
What Makes Some ...	David Hsu, Wee Sun L...	null
BibPro: A Citation P...	C. C. Chen, K. H. Yang...	2008
Two field studies of ...	Nielsen, J. and Lyngba...	1990
Improving System U...	Nielsen, J. and Faber, ...	1996; F...
A Generative Model...	Wei Lu, Hwee Tou Ng,...	null

Publications

- Wei Lu, Hwee Tou Ng, Wee Sun
  A Generative Model for Parsi
  pdf
  In Proceedings of the 2008 Conf
  2008) .
- Hai Leong Chieu and Wee Sun L
  Relaxed Survey Propagation: .
  pdf
  In *Proceedings of the 23rd AAA*
- Hanna Kurniawati, David Hsu, an
  SARSOP: Efficient point-base
  spaces

Conclusion

- **Results**
  - Fast and lightweight reference string parser
  - Reference string recogniser with good recall
  - Basic, expandable Firefox extension

Conclusion

- **Future work**
  - Reference String Recognition
    - More rules to improve precision
  - Reference String Parser
    - Use web page reference strings as training data
    - Recognise implicit/common metadata
  - Firefox Extension
    - Add more features to the extension

Questions?