Disambiguating Web Appearances of People in a Social Network

Ron Bekkerman
Andrew McCallum
University of Massachusetts at Amherst

WWW2005 Tutorials and Workshop - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites

Address http://www2005.org/tutorials/ Go Links

Google Search Web 406 blocked AutoFill Options

Contact

- Contact

Cancelled	Practice			
TP04 *Cancelled*	MDA Standards for Ontology Development	Dragan Gašević, Dragan Djurić, and Vladan Devedžić		
TA05	Introduction to RDF Query with SPARQL	Dave Beckett, Steve Harris, Eric Prud'hommeaux and Andy Seaborne		101A
TP05 *Cancelled*	Web-based Interactive Collaboration using Semantic Web Technology - Introduction of the Annotea and its Comparison to the Blog	Nobuhisa Shiraishi		
TA06 *Cancelled*	Networked Arts - Methods and Tools to create and maintain Virtual Museums	Alfredo Ronchi	INFO	
TP06	Matching Words and Pictures - Problems, Applications and Progress	Latifur Khan		301A
TA07	Web Content Mining	Bing Liu		301A
TP07	Location-based Services in Mobile Information Systems - Architectures, Description, and Systems	Ling Liu		101B

Full Day Tutorials

Please refer Description of Full Day Tutorials for the details.

	TITLE	PRESENTERS	INFO	ROOM

Done Internet

start 3 X-Win32 X-... presentation 2 SSH Secure... 3 Internet Ex... Inbox - Microso... www05.ppt 6:54 PM

WWW2005 Tutorials and Workshop - Microsoft Internet Explorer

http://www2005.org/tutorials/

Contact

- Contact

Cancelled	Practice	Motomu Naito		
TP04 *Cancelled*	MDA Standards for Ontology Development	Dragan Gašević, Dragan Djurić, and Vladan Devedžić		
TA05	Introduction to RDF Query with SPARQL	Dave Beckett, Steve Harris, Eric Prud'hommeaux and Andy Seaborne		101A
TP05 *Cancelled*	Web-based Interactive Collaboration using Semantic Web Technology - Introduction of the Annotea and its Comparison to the Blog	Nobuhisa Shiraishi		
TA06 *Cancelled*	Networked Arts - Methods and Tools to create and maintain Virtual Museums	Alfredo Ronchi	INFO	
TP06	Matching Words and Pictures - Problems, Applications and Progress	Latifur Khan		301A
TA07	Web Content Mining	Bing Liu		
TP07	Location-based Services in Mobile Information Systems - Architectures, Description, and Systems	Ling Liu		

Full Day Tutorials

Please refer Description of Full Day Tutorials for the details.

	TITLE	PRE

Bing Liu

Associate Professor

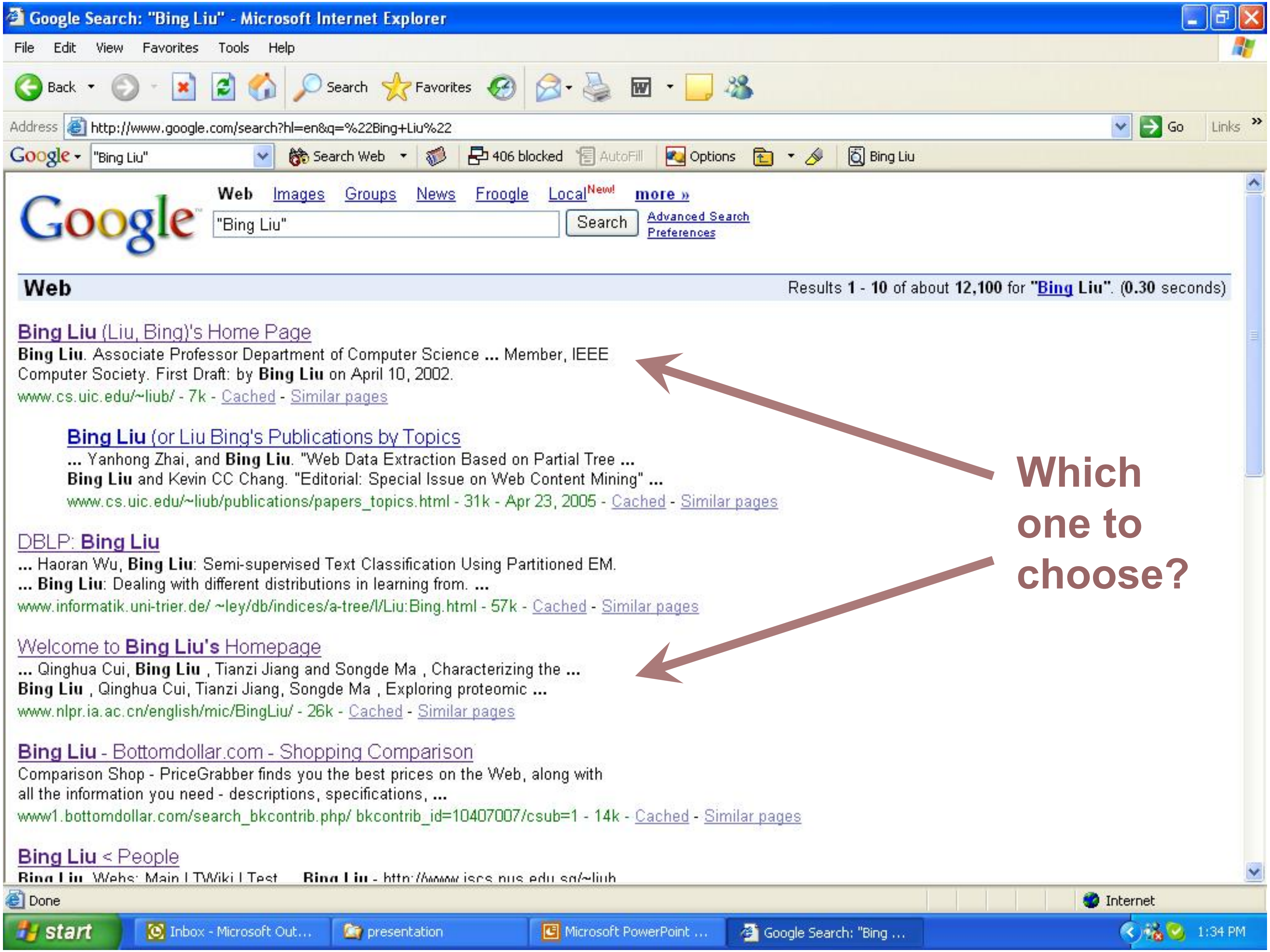
Department of Computer Science

University of Illinois at Chicago

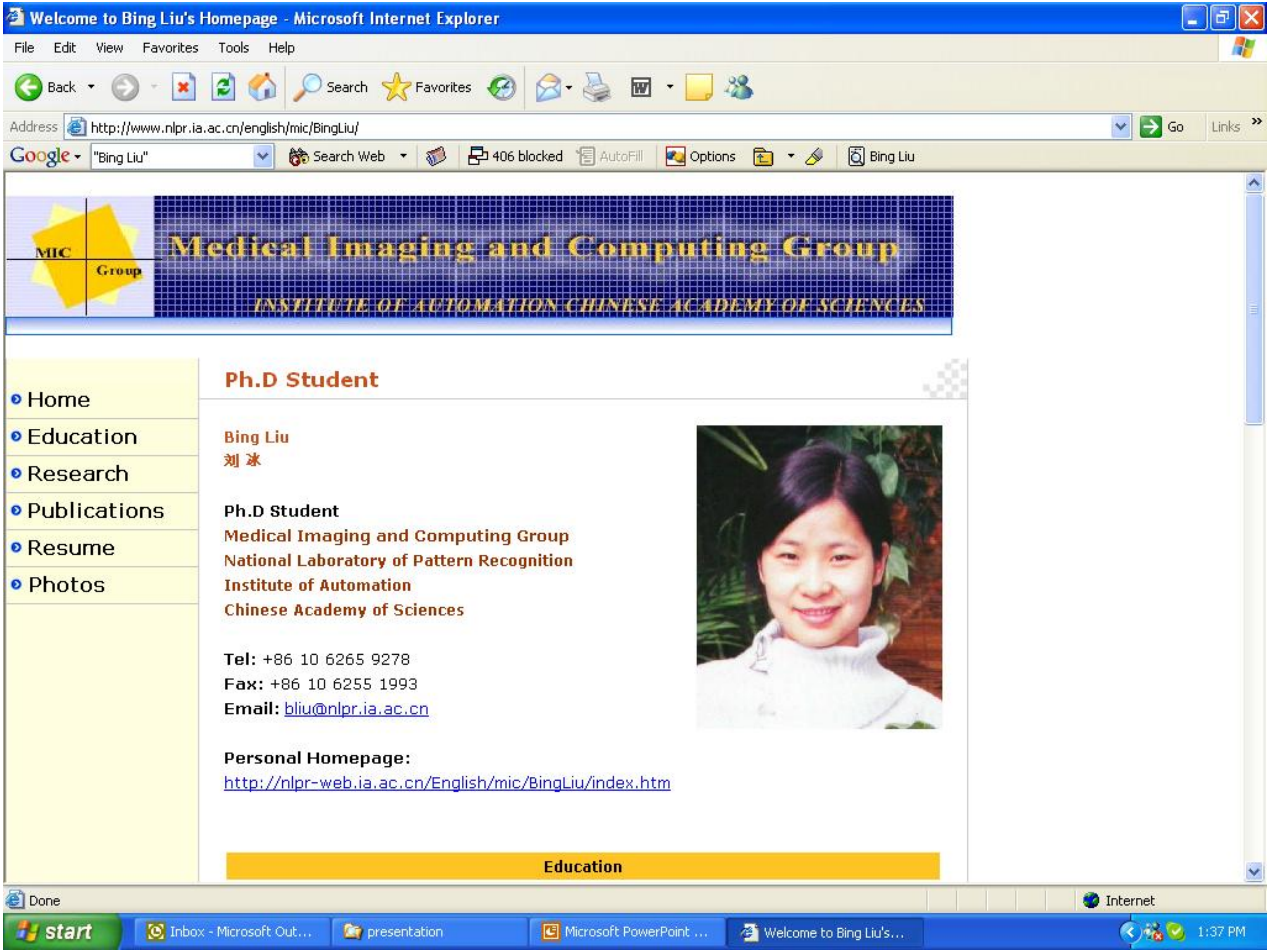
PhD, 1989, University of Edinburgh

Interests: Data Mining, Machine Learning

Done



Which one to choose?


Welcome to Bing Liu's Homepage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.nlpr.ia.ac.cn/english/mic/BingLiu/
Go Links
Google "Bing Liu" Search Web 406 blocked AutoFill Options Bing Liu
MIC Group
Medical Imaging and Computing Group
INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES
Home
Education
Research
Publications
Resume
Photos
Ph.D Student
Bing Liu
刘 冰
Ph.D Student
Medical Imaging and Computing Group
National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Tel: +86 10 6265 9278
Fax: +86 10 6255 1993
Email: bliu@nlpr.ia.ac.cn
Personal Homepage:
http://nlpr-web.ia.ac.cn/English/mic/BingLiu/index.htm
Education
Done
Internet
start
Inbox - Microsoft Out...
presentation
Microsoft PowerPoint ...
Welcome to Bing Liu's...
1:37 PM

Looking for a person on the Web

- Web appearance disambiguation
  - Which pages refer to the particular person
- Web page content filtering
  - Which section is relevant to the person
- Decision making
  - E.g., whether there are *no* relevant pages

Looking for a person on the Web

- Web appearance disambiguation
  - Which pages refer to the particular person
- Web page content filtering
  - Which section is relevant to the person
- Decision making
  - E.g., whether there are *no* relevant pages

Objectives

- High precision
  - Retrieve information only about the desired person
  - Precision error can be a disaster
- High recall
  - Retrieve as much relevant info as possible
  - High recall is the major goal

Bing Liu

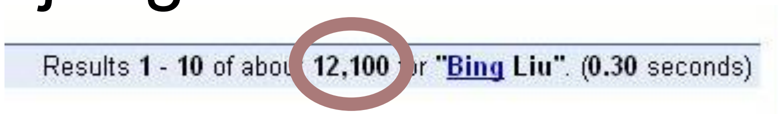
Associate Professor

Department of Computer Science

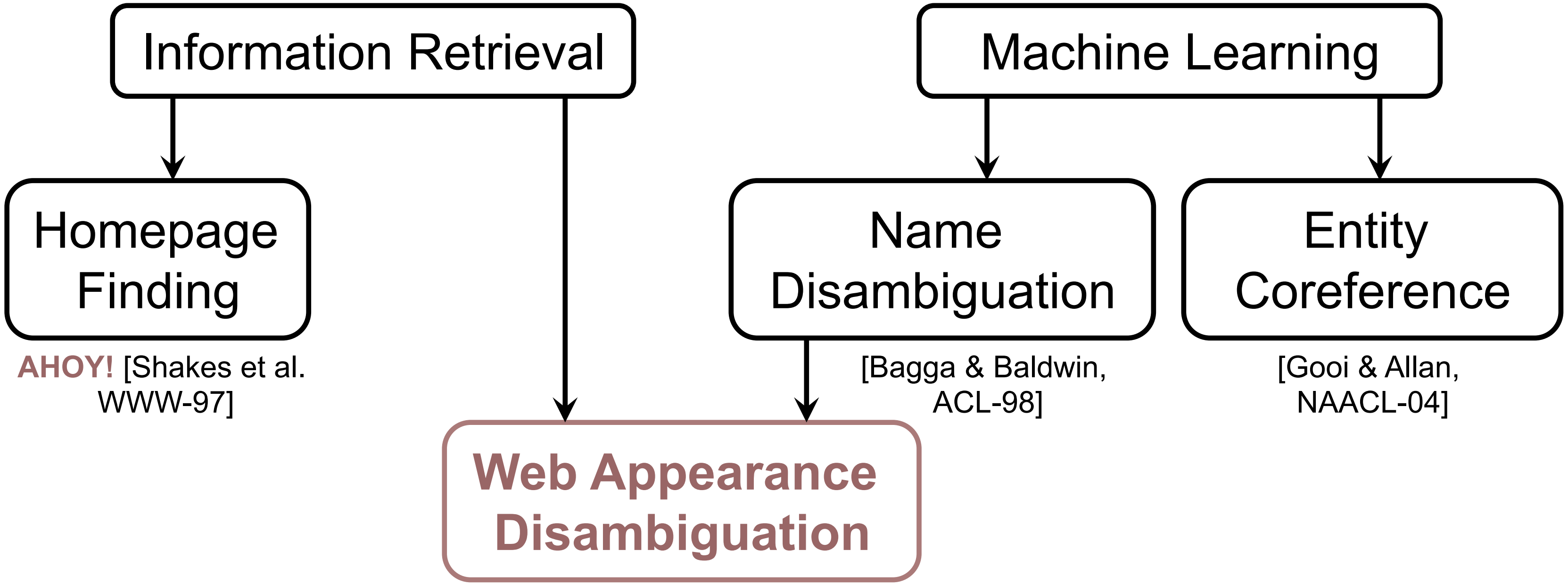
University of Illinois at Chicago

PhD, 1989, University of Edinburgh

Interests: Data Mining, Machine Learning



Localization



- ***Coreference***: unite <u>different</u> mentions of the <u>same</u> entity
- ***Disambiguation***: distinguish between <u>identical</u> mentions of <u>different</u> entities

Literature on name disambiguation

- Bagga & Baldwin, ACL-98
- Mann & Yarowsky, CoNLL-03
- Fleischman & Hovy, ACL-04
- Pedersen et al., CICLing-05
- Han et al., JCDL-05

All these perform clustering

- Increasing interest!

Requirements

- Asymmetry
  - Accept pages of the desired person
  - Reject pages of his/her namesakes
- Unsupervised approach
  - No training set can model the Web
- Single-link preferred over average-link

- How to represent the *desired* person?

Name itself tells nothing

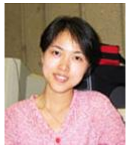
- Given just a name, the task is ill-defined

- Additional information required!
  - Keywords? (e.g. “professor”, “student”)
    - Nelken et.al., WWW 2003
    - But how to obtain them?!

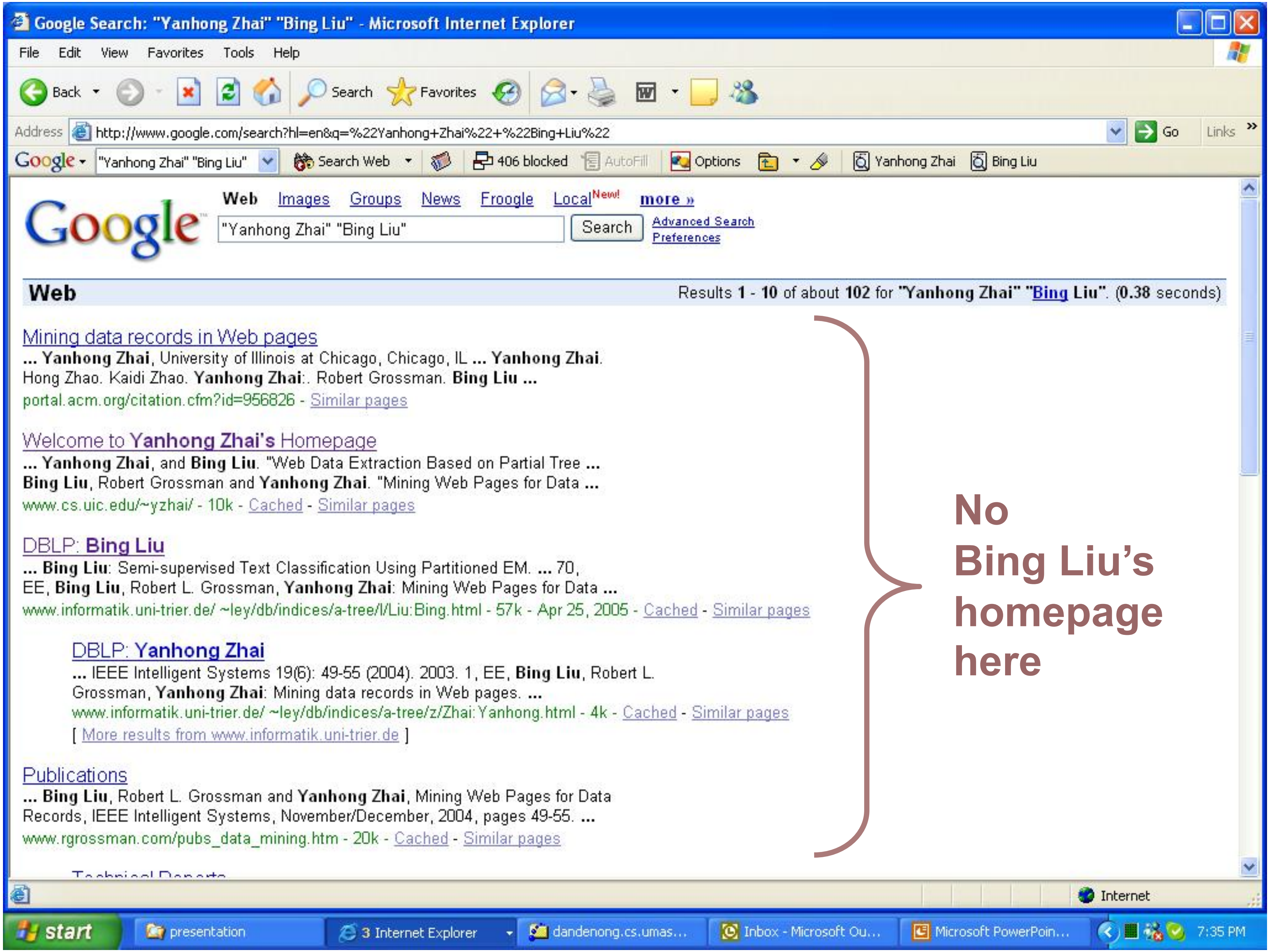
The idea

- Consider a *list* of names!
  - Of people in *one* social network

Yanhong Zhai* + *Bing Liu (I) | ***Bing Liu (II)***



- Not burdensome to obtain such list
  - Any name appears in context of other names



No Bing Liu's homepage here

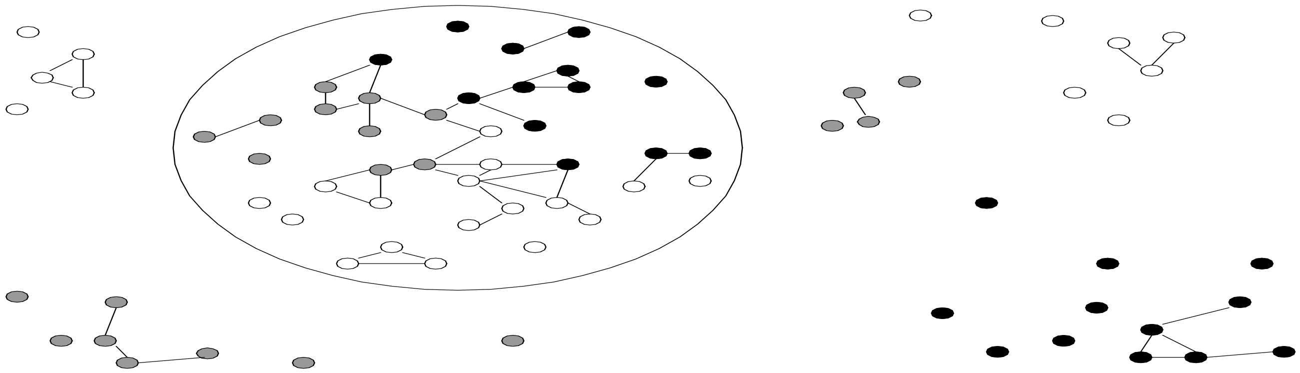
Approaches

- Pages of acquaintances are interconnected
- Link Structure Model (LS)
  - Build a core of interconnected pages
    - Of *different* people!
  - Add proximate pages to the core

Approaches

- Pages of acquaintances are interconnected
- Link Structure Model (LS)
  - Build a core of interconnected pages
    - Of *different* people!
  - Add proximate pages to the core
- Distributional Clustering Model (DC)
  - Simultaneously cluster pages and their words
    - Double clustering is usually more accurate
  - Pick cluster with most interconnected pages

Link Structure model (LS)



- Nodes are pages, edges are hyperlinks
- This picture shows pages of 3 people

Technical details of LS

- Pages are *interconnected* if they share a hyperlink
  - URL's domain & first dir:

    http://**www.cs.umass.edu/~ronb**/enron_dataset.html
  - And the domain is not too common

Use Google's *link:* operator

Technical details of LS

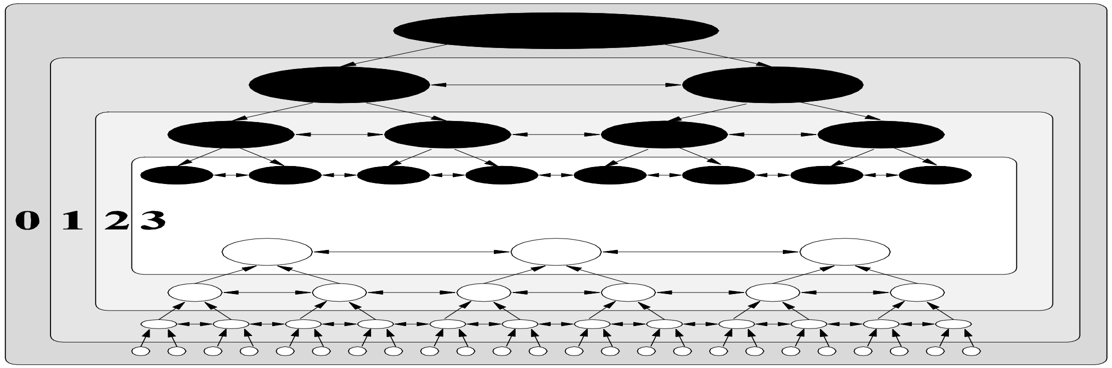
- Pages are *interconnected* if they share a hyperlink
  - URL's domain & first dir:

    http://**www.cs.umass.edu/~ronb**/enron_dataset.html
  - And the domain is not too common
- *Cosine similarity* between page & the core
  - With novel *google_tfidf* weighting:

$$\text{google\_tfidf}(w) = \frac{\text{tf}(w)}{\log(\text{google\_df}(w))}$$

Use Google's *total results count*

Distributional Clustering model (DC)



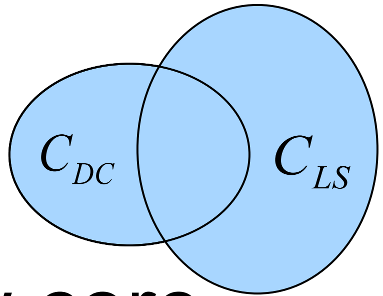
Clusters of words

Clusters of pages

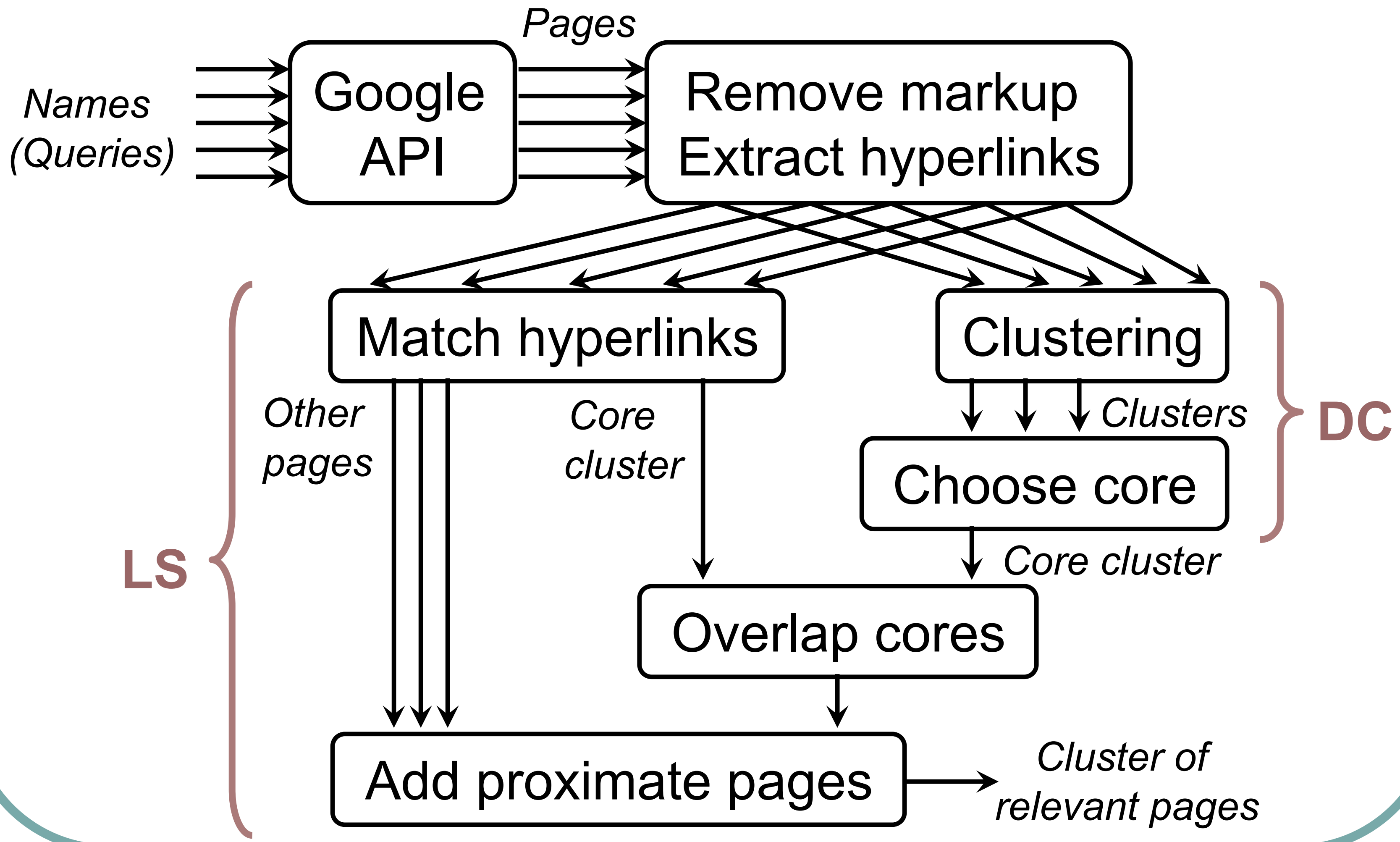
- Bekkerman, El-Yaniv & McCallum, 2005
  - Submitted to ICML

Hybrid model (LS+DC)

- DC starts with small but clean clusters
  - One of which, C_{DC}, is most interconnected
- Overlap LS's core with C_{DC}
  - Obtain larger but still clean core
- Add proximate pages to the new core
  - Just as in LS model



LS+DC system overview



Dataset

Personal name	***Position***	***Pages***	***Namesakes***	***Relevant pages***
Adam Cheyer	SRI Manag	97	2	96
William Cohen	CMU Prof	88	10	6
Steve Hardt	SRI Eng	81	6	64
David Israel	SRI Manag	92	19	20
Leslie Pack Kaelbling	MIT Prof	89	2	88
Bill Mark	SRI Manag	94	8	11
Andrew McCallum	UMass Prof	94	16	54
Tom Mitchell	CMU Prof	92	37	15
David Mulford	Stanf Undergrad	94	13	1
Andrew Ng	Stanf Prof	87	29	32
Fernando Pereira	UPenn Prof	88	19	32
Lynn Voss	SRI Eng	89	26	1
	OVERALL:	***1085***	***187***	***420***

- 12 names out of Melinda Gervasio's social network

Results

Method	***Precision***	***Recall***	***F-measure***
Agglom. clust.	61.7	53.3	57.2
LS	84.2	71.8	77.5
DC	87.3±1.7	71.3±2.5	78.4±0.9
LS+DC Hybrid	86.9	74.5	**80.3**

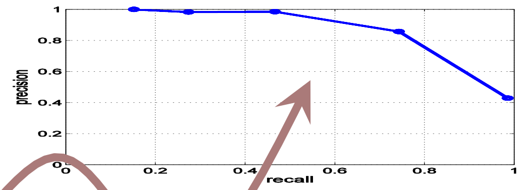
- 20% higher than the baseline
- Hybrid method increases recall
  - As we could predict

Results of LS+DC by person

Personal name	***Found correct***	***Not found***	***Found wrong***
Adam Cheyer	62	34	0
William Cohen	6	0	4
Steve Hardt	16	48	2
David Israel	19	1	4
Leslie Pack Kaelbling	84	4	1
Bill Mark	6	5	9
Andrew McCallum	54	0	2
Tom Mitchell	14	1	5
David Mulford	1	0	0
Andrew Ng	30	2	6
Fernando Pereira	21	11	14
Lynn Voss	0	1	0
OVERALL:	***313***	***107***	***47***

Additional results

- Distinguishing “doubles”
  - Intermediate iteration of DC algorithm
  - 98% precision, 45% recall
  - Doubles well discriminated

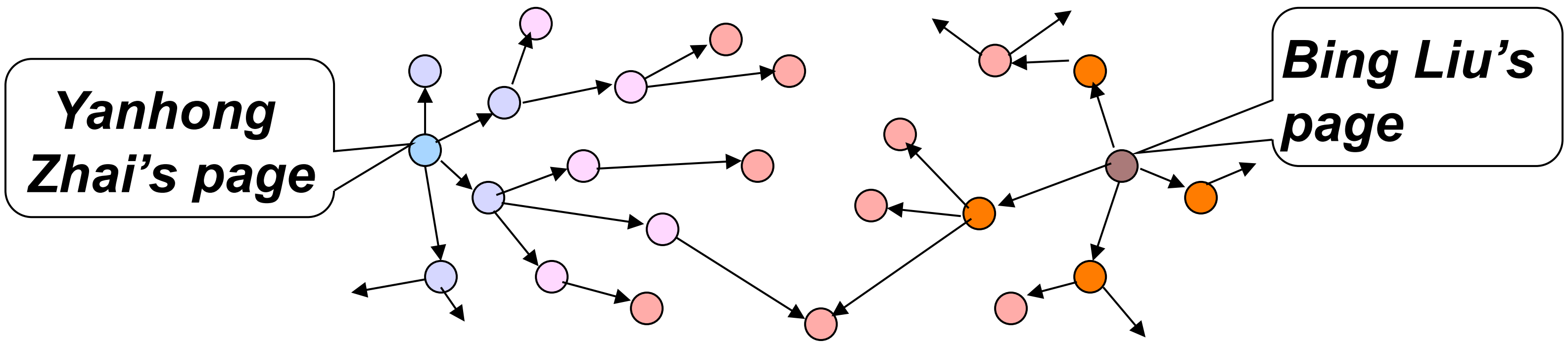


- Homepage finding
  - 9 of 10 homepages are found
  - Except for Steve Hardt’s
  - Mulford and Voss have no homepages

Ongoing research

with S. Zilberstein and J. Allan

- Heuristic search in the Web graph



- Two people are in one social network
  - If there's a path between their pages
- 89.6% precision
  - With up to 4-length paths only

Conclusion

- First attempt to tackle the problem
  - Of finding people's web appearances
- Many applications
  - Web, email, social network analysis…
- Proposed methods can be also used:
  - For acronym disambiguation
  - For word sense disambiguation

Conclusion

- First attempt to tackle the problem
  - Of finding people's web appearances
- Many applications
  - Web, email, social network analysis…
- Proposed methods can be also used:
  - For acronym disambiguation
  - For word sense disambiguation
- Thank you!