Information Extraction

Lecture #19

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

Today's Main Points

- Why IE?
- Components of the IE problem and solution
- Approaches to IE segmentation and classification
  - Sliding window
  - Finite state machines
- IE for the Web
- Semi-supervised IE

- Later: relation extraction and coreference
- …and possibly CRFs for IE & coreference

Query to General-Purpose Search Engine:

+camp +basketball "north carolina" "two weeks"

File Edit View Go Communicator Help

Bookmarks Location: http://www.altavista.com/cgi-bin/query?pg=q&kl=XX&q=%2

AltaVista found about 94867 Web pages for you. Refine your search

1. Parent Information - Camp High Rocks
   Parent Guide - Camp High Rocks. This page is for Parents of Camp High Rocks campers, but portions can be useful for any parents of children that attend...
   URL: highrocks.com/hrp71.htm
   Last modified 20-Jun-98 - page size 21K - in English [Translate]

2. [7/15/98] Tim Stevens: North Carolina players shine at ABCD basketball camp
   The N&O: Sports
   URL: www.news-observer.com/daily/1998/07/15/sports11.html
   Last modified 22-Sep-98 - page size 14K - in English [Translate]

3. Rasheed Wallace signs six-year, $80 million contract
   Rasheed Wallace signs six-year, $80 million contract. Copyright © 1997 Nando.net Copyright © 1997 The Associat Press. For a complete roundup of...
   URL: wedge.nando.net/newsroom/ntn/sports/0929..._2173_body.html
   Last modified 17-Aug-98 - page size 5K - in English [Translate]

4. As WNBA rookie, Reid the right fit for Charlotte (05/13/98, CHN)
   nbsp; Home | Classifieds | News | Sports | Arts | Calendar. Chapel Hill News Real Estate Section | Opinion | Around Town. As WNBA rookie, Reid the right...
   URL: www.chapelhillnews.com/Issues/1998/05/13/sports05.html
   Last modified 9-Aug-98 - page size 5K - in English [Translate]

5. May 22 1998 Newsletter - St. Andrew's Church
   A-Spire. St. Andrew's Presbyterian Church, Kitchener, Ontario. Vol. 6 No. 7. Friday, May 22,1998. In this Issue: Twenty-two new elders ordained Camp...
   URL: wchat.on.ca/public/standrews/aspire/aspiremay98.htm
   Last modified 21-Jun-98 - page size 20K - in English [Translate]

6. North Carolina State:1997 Fan's Guide to ACC BASKETBALL
   The Players. ISHUA JABASS BENJAMIN. Herb Sendek says he knows what it's like to edure 'humbling, ego-bruisin times as a basketball player. The...
   URL: www.nando.net/SportServer/FansGuide/ncsu/players.html
   Last modified 27-Aug-98 - page size 41K - in English [Translate]

Amazon.com suggests
Books about +camp +bas
Amazon.com Bestsellers

Domain-Specific Search Engine

File Edit View Go Communicator Help

Bookmarks Location: http://www.camp.ca/cgi-shl/dbml.exe?template=/camp/

Internet Lookup New&Cool Netcaster

CAMP SEARCH .com

THE SEARCH ENGINE FOR CAMPS

Empty fields are ignored in the search.

Where in United States of America do you want to go to camp?

North Carolina

What type of Camp do you want to go to?

GENDER: Co-ed

ORGANIZATION: Doesn't Matter

STAY TYPE:

- Day Camp
- Residential Camp
- Specialty Camp
- Family Camp
- Tours & Adventures
- Outdoor Education
- Conference Site
- Adult Camp

How old are you?

How much do you want to spend? (per week)

Doesn't Matter

Are you looking for a specialty camp?

Football
Wind Surfing
Basketball
Canoeing
Ice Hockey
Aquatic

How long do you want to go for?

2 weeks

File Edit View Go Communicator Help

Bookmarks Location: http://www.campsearch.com/

CAMP SEARCH.com

THE SEARCH ENGINE FOR CAMPS

FEATURED CAMPS

Day Camps

Residential Camps

Specialty Camps

Family Camps

Tours & Adventures

Outdoor Education

Conference Sites

Adult Camps

GUIDE

- Search Options
- Add Your Camp
- Update Your Listing
- Camp Job Network
- Camp Mall
- Help
- Contact Us

CLICK HERE TO SEARCH 2000+ CAMPS

http://www.campsearch.com/residential

File Edit View Go Communicator Help

Bookmarks Location: http://www.mrqe.com/lookup?^Lost+World%3a+Jurassic+P

Lost World: Jurassic Park, The (1997)

Reviews *130 articles*

The New York Times review *registration required*
Scott Renshaw <renshaw@inconnect.com> review
ReelViews (James Berardinelli) review [3/4]
Boxoffice Magazine review
Chicago Reader (Bill Stamets) review
Addicted to Noise (Cynthia Fuchs) review
Girls on Film (Chicks, flicks, politicks!) review
Roger Ebert, Chicago Sun-Times review [2/4]
Microsoft Cinemania Online review
Empire Magazine review [UK]
Washington Post review
Dr. Daniel's Movie Emergency review
Deseret News, Salt Lake City review
Cinemaven Online (Doug Thomas) review [2.5/4]
The Flick Filosopher (MaryAnn Johanson) review
New Jersey Online review
Harvey O'Brien (Harvey's Movie Review) review
Movie Magazine International review
Images (Gary Johnson) review

http://movie-reviews.colossus.net/movies/j/jurassic2.

Example: The Problem

Advanced Search Preferences Language Tools Search Tips

baker job opening | Google Search

Web | Images | Groups | Directory | News-New!

Searched the web for **baker job opening**. Results

Job Opening - Find ANY **Job**! - Search by Type, Industry & Geography
www.careerbuilder.com Post Your RESUME Here to Reach Thousands of Employers - It's FREE!

Job Opening At Flipdog.Com
www.FlipDog.com Fetch your next **job** at FlipDog.com!

Softimage::Community::Discussion Groups::ds.archive.0004
... Le Rudulier; Drive space Ken Skaggs; Help about rendering denis.courtot; **JOB OPENING** ... Tony Cacciarelli; RE: ALE Karim Arbaoui; RE: omf to timeline Martin **Baker**; Re ...
www.softimage.com/community/xsi/discuss/ Archives/ds.archive.0004/default.htm - 49k - Cached - Similar pages

Softimage::Community::Discussion Groups::ds.archive.0004
... Re: **JOB OPENING** Philip Herring - 2000/04/28 22:35. ... RE: omf to timeline Martin **Baker** - 2000/04/26 17:33: Re: omf to timeline adam - 2000/04/26 18:11. ...
www.softimage.com/community/xsi/discuss/Archives/ ds.archive.0004/ThreadIndex.htm - 50k - Cached - Similar pages
[More results from www.softimage.com]

CGI : **Job Opening**
www.genomics.cornell.edu/jobs/view_job.cfm?id=10 - 15k - Cached - Similar pages

Information Activist **Job Opening** - May 2001
www.igc.org/datacenter/job.html - 6k - Cached - Similar pages

Post an Employee Benefits **Job Opening** (Help Wanted) Ad
... edit the ad to add a new **job opening** ... as possible when it is emailed to 2,985 **job** ... jobs/posthelpwanted.shtml · Webmaster: webmaster@BenefitsLink.com (Dave **Baker** ...
www.benefitslink.com/jobs/posthelpwanted.shtml - 24k - Cached - Similar pages

Post an Employee Benefits **Job Opening** (Help Wanted) Ad
Employee Benefits Jobs! Brought to you by BenefitsLink (tm) and its EmployeeBenefitsJobs.com (tm) division.
www.benefitslink.com/jobs/pricinginfo.shtml - 7k - Cached - Similar pages
[More results from www.benefitslink.com]

***Martin Baker*, a person**

Genomics job

Employers job posting form

Example: A Solution

job search find employment careers @ FlipDog.com free! - Microsoft Internet Explorer

Address http://www.flipdog.com/home.html

FlipDog.com

Home | Find Jobs | Your Account | Resource Center | Support | Employers

Job Search at FlipDog.com: Employment & Career Management

647,514 Job Opportunities from 53,641 Employers

Find a Job!

Post Your Resume

Employers click here for Products & Services

Job Seekers: Find your dream job!

- Check our 'Best Places to Find a Job' January report.
- Open your FREE account and put your resume online.
- Search 24x7 with our FREE automatic JobHunters™.
- Research our database of over 50,000 employers.
- Get expert advice at our new Resource Center .
- Access salary surveys/calculators, relocation tools, networking opportunities, & training/testing tools.
- Use FlipDog.com to search jobs right from your desktop! Download Snippets today!

Pigskin Places

Health Care in NY	2,770
Health Care in MD	1,262
Sales in NY	3,751
Sales in MD	958
Computing in NY	8,050
Computing in MD	4,114

Jobs for Sports Fans

- Head Football Coach
- Football Coach
- Asst. Football Coach
- High School Football Coach
- Univ. Asst. Football Coach

Job Seeker Newsletter

Enter your e-mail address:

Sign Me Up!

"Top 100 Web Sites" PC Magazine, Nov. 2000

"Top 10 Career Web Site" Media Metrix, Sept. 2000

"Top 10 Job Site"

Showcase Jobs

MRI Management Recruiters of Charlotte North

We provide total staffing solutions in the areas of Human Resources, Compensation, Web-based HR self-service, and Customer Management Systems.

Learn More

Looking for a Vice President of Academic Affairs to oversee planning, operation and evaluation of the college's academic programs.

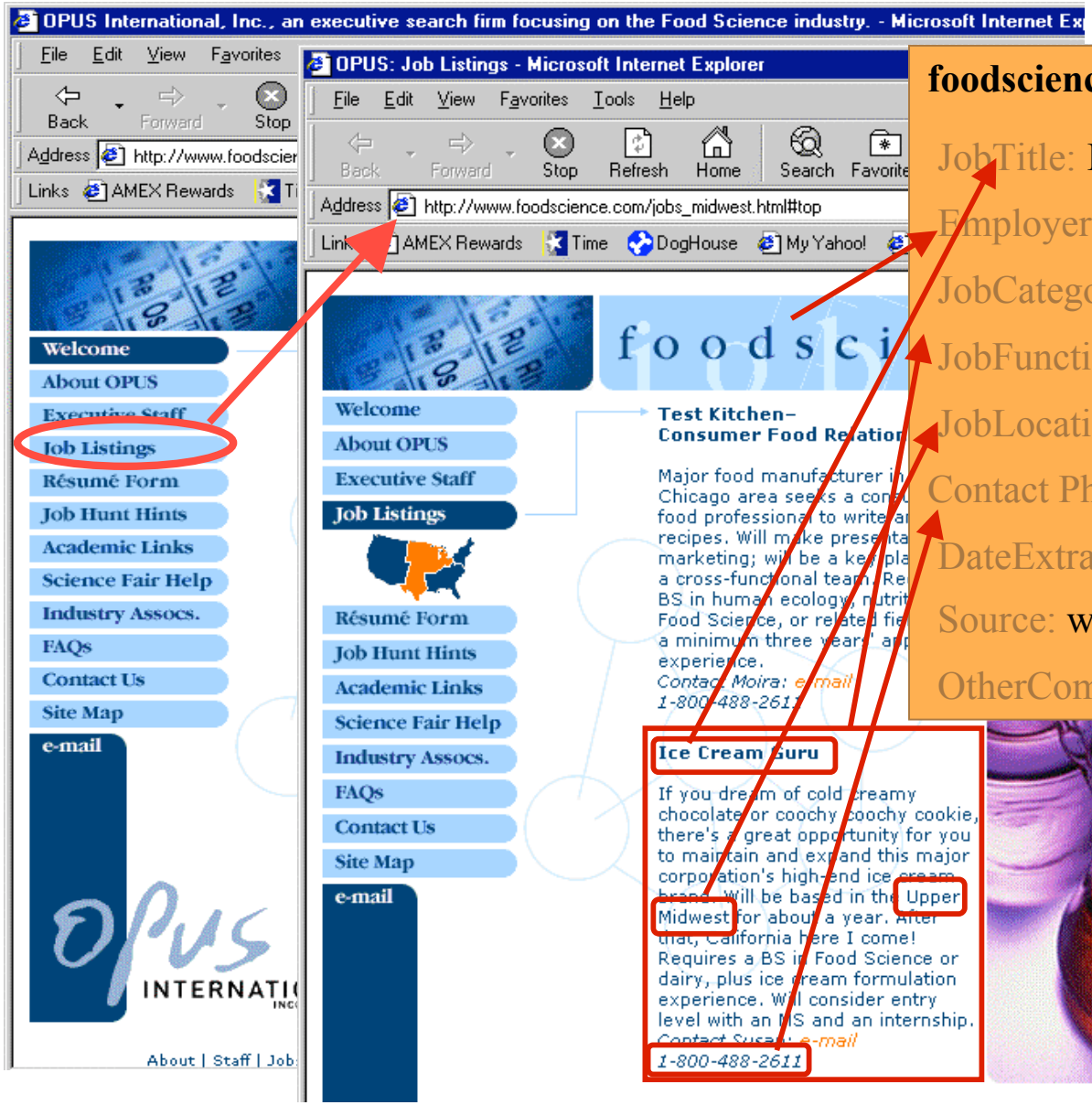
Learn More

powered by WhizBang!

Internet

Start | Microsoft PowerPoint - [sta... | job search find employmen... | 12:12 AM

Extracting Job Openings from the Web



foodscience.com-Job2

JobTitle: Ice Cream Guru

Employer: foodscience.com

JobCategory: Travel/Hospitality

JobFunction: Food Services

JobLocation: Upper Midwest

Contact Phone: 800-488-2611

DateExtracted: January 8, 2001

Source: www.foodscience.com/jobs_midwest.htm

OtherCompanyJobs: foodscience.com-Job1

Job Openings:
Category = *Food Services*
Keyword = *Baker*
Location = *Continental U.S.*

FlipDog.com — Fetch Your Next Job Here™

• Employers • Support

Home | Find Jobs | Your Account | Resource Center

Return to Results | Modify Search | New Search

The University Alliance — A BISK EDUCATION NETWORK — Degrees Online
Learn While You Earn **MBA, BA, AA** Degrees Online & **Project Mgt**.

Click here to e-mail your resume to 1000's of Head Hunters with **ResumeZapper.com**

how to easily DOUBLE your chances when applying FOR JOBS!
Breakthrough ebook shows why most people are WRONG about how to apply for jobs.

1 - 25 of **47** jobs shown below

1 2 Next

Search these results for: GO! Search tips **Show Jobs Posted:** For all time periods

View: Brief | Detailed

Web Jobs: FlipDog technology has found these jobs on thousands of employer Web sites.

Food Pantry Workers at Lutheran Social Services	October 11, 2002	Archbold, OH
Cooks at Lutheran Social Services	October 11, 2002	Archbold, OH
Bakers Assistants at Fine Catering by Russell Morin	October 11, 2002	Attleboro, MA
Baker's Helper at Bird-in-Hand	October 11, 2002	United States
Assistant Baker at Gourmet To Go	October 11, 2002	Maryland Heights, MO
Host/Hostess at Sharis Restaurants	October 10, 2002	Beaverton, OR
Cooks at Alta's Rustler Lodge	October 10, 2002	Alta, UT
Line Attendant at Sun Valley Coporation	October 10, 2002	Huntsville, UT
Food Service Worker II at Garden Grove Unified School District	October 10, 2002	Garden Grove, CA
Night Cook / Baker at SONOCO	October 10, 2002	Houma, LA
Cooks/Prep Cooks at GrandView Lodge	October 10, 2002	Nisswa, MN
Line Cook at Lone Mountain Ranch	October 10, 2002	Big Sky, MT
Production Baker at Whole Foods Market	October 08, 2002	Willowbrook, IL
Cake Decorator/Baker at Mandalay Bay Hotel and Casino	October 08, 2002	Las Vegas, NV
Shift Supervisors at Brueggers Bagels	October 08, 2002	Minneapolis, MN

Data Mining the Extracted Job Information

Job Opportunity Index - Microsoft Internet Explorer provided by WhizBang! Labs

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://joi.flipdog.com/joi/ Links

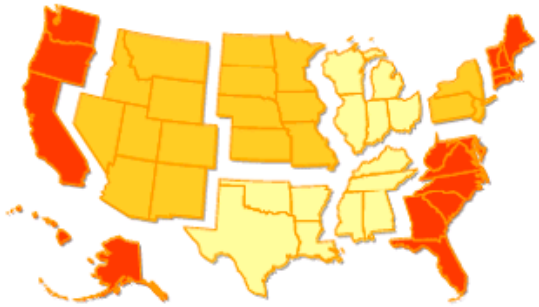
HOME SUBSCRIBE DOWNLOAD ABOUT JOI

FlipDog.com

Job Opportunity Index®

November 2001

Welcome -- Tuesday, May 7, 2002



U.S. JOB SUPPLY BY REGION

Above Average Average Below Average

UNITED STATES

November 2001 JOI: 28.4 (October: 27.7)
September Unemployment Rate: 5.4% (August: 4.9%)

Click on a region to see individual reports.

See printable version

U.S. Job Supply Increases Amid Rising Unemployment

The Job Opportunity Index™ (JOI) increased for the first time in three months in October – climbing 0.7 point to 28.4 and signifying a slight increase in U.S. job supply. However, numerous factors, including a dramatic half-point increase in the national unemployment rate, made October anything but normal.

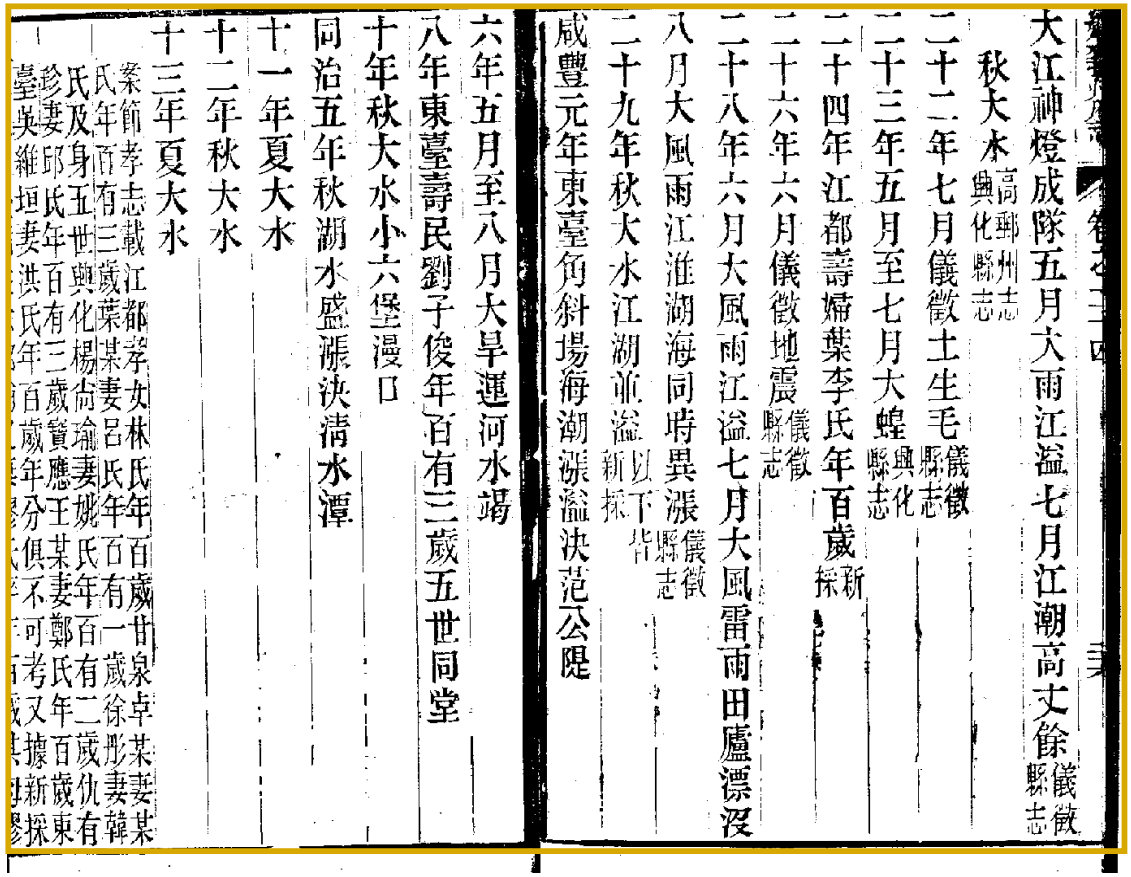
Subscribe now

Special Offer! Find out how you can earn a free subscription to the *JOI Report on U.S. Labor Markets* through a limited-time JOI Subscriber Referral Program!

Done Internet

IE from
Chinese Documents regarding Weather

Department of Terrestrial System, Chinese Academy of Sciences


大江神燈成隊五月大雨江溢七月江潮高丈餘儀徵縣志
秋大水高郵州志興化縣志
二十二年七月儀徵土生毛儀徵縣志
二十三年五月至七月大蝗興化縣志
二十四年江都壽婦葉李氏年百歲新採
二十六年六月儀徵地震儀徵縣志
二十八年六月大風雨江溢七月大風雷雨田廬漂沒
八月大風雨江淮湖海同時異漲儀徵縣志
二十九年秋大水江湖並溢以下皆新採
咸豐元年東臺角斜場海潮漲溢決范公隄
六年五月至八月大旱運河水竭
八年東臺壽民劉子俊年百有三歲五世同堂
十年秋大水小六堡漫口
同治五年秋湖水盛漲決清水潭
十一年夏大水
十二年秋大水
十三年夏大水
案節孝志載江都孝女林氏年百歲甘泉卓某妻某氏年百有三歲葉某妻呂氏年百有一歲徐彤妻韓氏及身五世與化楊尚瑜妻姚氏年百有二歲仇有珍妻邱氏年百有三歲寶應王某妻鄭氏年百歲東臺吳維垣妻洪氏年百歲年分俱不可考又據新採

200k+ documents
several millennia old

- **Qing Dynasty Archives**
- **memos**
- **newspaper articles**
- **diaries**

IE from Research Papers

[McCallum et al '99]

Reinforcement Learning: A Survey

Leslie Pack Kaelbling LPK@CS.BROW
Michael L. Littman MLITTMAN@CS.BROW
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore AWM@CS.CM
Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

This paper surveys the field of reinforcement learning from a computer-science p
spective. It is written to be accessible to researchers familiar with machine learning. Bo
the historical basis of the field and a broad selection of current work are summariz
Reinforcement learning is the problem faced by an agent that learns behavior throu
trial-and-error interactions with a dynamic environment. The work described here has
resemblance to work in psychology, but differs considerably in the details and in the u
of the word "reinforcement." The paper discusses central issues of reinforcement learni
including trading off exploration and exploitation, establishing the foundations of the fi
via Markov decision theory, learning from delayed reinforcement, constructing empiri
models to accelerate learning, making use of generalization and hierarchy, and coping wi
hidden state. It concludes with a survey of some implemented systems and an assessmo
of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in sta
psychology, neuroscience, and computer science. In the last five to ten years, it has att
rapidly increasing interest in the machine learning and artificial intelligence commu
Its promise is beguiling—a way of programming agents by reward and punishment w
needing to specify *how* the task is to be achieved. But there are formidable comput
obstacles to fulfilling the promise.

This paper surveys the historical basis of reinforcement learning and some of the c
work from a computer science perspective. We give a high-level overview of the field
taste of some specific approaches. It is, of course, impossible to mention all of the imp
work in the field; this should not be taken to be an exhaustive account.

IE from Research Papers

A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation - Peter, Wi - Microsoft Internet Explorer p

File Edit View Favorites Tools Help

Back Search Favorites History

Address http://citeseer.nj.nec.com/peter90critical.html Links

A Critical Evaluation of Commensurable Abduction Models for Semantic Interpretation (1990) (Correct) (5 citations)

Peter Norvig Robert Wilensky University of California, Berkeley Computer...

Thirteenth International Conference on Computational Linguistics, Volume 3

NEC ResearchIndex Bookmark Context Related

Download: norvig.com/coling.ps

Cached: PS.gz PS PDF DjVu Image Update Help

From: norvig.com/resume (more)

Home: R.Wilensky HPSearch (Correct)

(Enter summary)

Rate this article: 1 2 3 4 5 (best)

Comment on this article

Abstract: this paper we critically evaluate three recent abductive interpretation models, those of Charniak and Goldman (1989); Hobbs, Stickel, Martin and Edwards (1988); and Ng and Mooney (1990). These three models add the important property of commensurability: all types of evidence are represented in a common currency that can be compared and combined. While commensurability is a desirable property, and there is a clear need for a way to compare alternate explanations, it appears that a single scalar measure is not enough to account for all types of processing. We present other problems for the abductive approach, and some tentative solutions. (Update)

Context of citations to this paper: More

.... **(break slight modification of the one given in [Ng and Mooney, 1990] The new definition remedies the anomaly reported in [Norvig and Wilensky, 1990] of occasionally preferring spurious interpretations of greater depths.** Table 1: Empirical Results Comparing Coherence and...

.... **costs as probabilities, specifically within the context of using abduction for text interpretation, are discussed in Norvig and Wilensky (1990).** The use of abduction in disambiguation is discussed in Kay et al. 1990) We will assume the following: 13) a. Only literals...

Cited by: More

Translation Mismatch in a Hybrid MT System - Gawron (1999) (Correct)

Abduction and Mismatch in Machine Translation - Gawron (1999) (Correct)

Interpretation as Abduction - Hobbs, Stickel, Appelt, Martin (1990) (Correct)

Active bibliography (related documents): More All

0.1: Critiquing: Effective Decision Support in Time-Critical Domains - Gertner (1995) (Correct)

0.1: Decision Analytic Networks in Artificial Intelligence - Matzkevich, Abramson (1995) (Correct)

Internet

Mining Research Papers

Most cited authors in Computer Science - June 2004 (CiteSeer.IST)

Generated from documents in the CiteSeer.IST database. This list does not inclu
where one or more authors of the citing and cited articles match, or citations whe
relevant author is an editor. An entry may correspond to multiple authors (e.g. J.
list is automatically generated and may contain errors. Citation counts may differ
results because this list is generated in batch mode whereas the database is contin
updated. A total of 703686 authors were found.

1. D. Johnson: 13216
2. J. Ullman: 11724
3. A. Gupta: 8968
4. R. Milner: 8464
5. R. Rivest: 7552
6. M. Garey: 7295
7. R. Tarjan: 7106
8. J. Dongarra: 7007
9. V. Jacobson: 6937
10. L. Lamport: 6780
11. J. Smith: 6563
12. S. Shenker: 6411
13. D. Knuth: 6352
14. E. Clarke: 6272
15. S. Floyd: 6133
16. A. Aho: 5795
17. J. Hennessy: 5759
18. R. Agrawal: 5702
19. C. Papadimitriou: 5690
20. R. Johnson: 5613
21. A. Pnueli: 5598
22. L. Zhang: 5438
23. D. Goldberg: 5414

[Rosen-Zvi, Griffiths, Steyvers, Smyth, 2004]

TOPIC 19	
WORD	**PROB.**
LIKELIHOOD	0.0539
MIXTURE	0.0509
EM	0.0470
DENSITY	0.0398
GAUSSIAN	0.0349
ESTIMATION	0.0314
LOG	0.0263
MAXIMUM	0.0254
PARAMETERS	0.0209
ESTIMATE	0.0204
AUTHOR	**PROB.**
Tresp_V	0.0333
Singer_Y	0.0281
Jebara_T	0.0207
Ghahramani_Z	0.0196
Ueda_N	0.0170
Jordan_M	0.0150
Roweis_S	0.0123

TOPIC 24	
WORD	**PROB.**
RECOGNITION	0.0400
CHARACTER	0.0336
CHARACTERS	0.0250
TANGENT	0.0241
HANDWRITTEN	0.0169
DIGITS	0.0159
IMAGE	0.0157
DISTANCE	0.0153
DIGIT	0.0149
HAND	0.0126
AUTHOR	**PROB.**
Simard_P	0.0694
Martin_G	0.0394
LeCun_Y	0.0359
Denker_J	0.0278
Henderson_D	0.0256
Revow_M	0.0229
Platt_J	0.0226

Named Entity Recognition

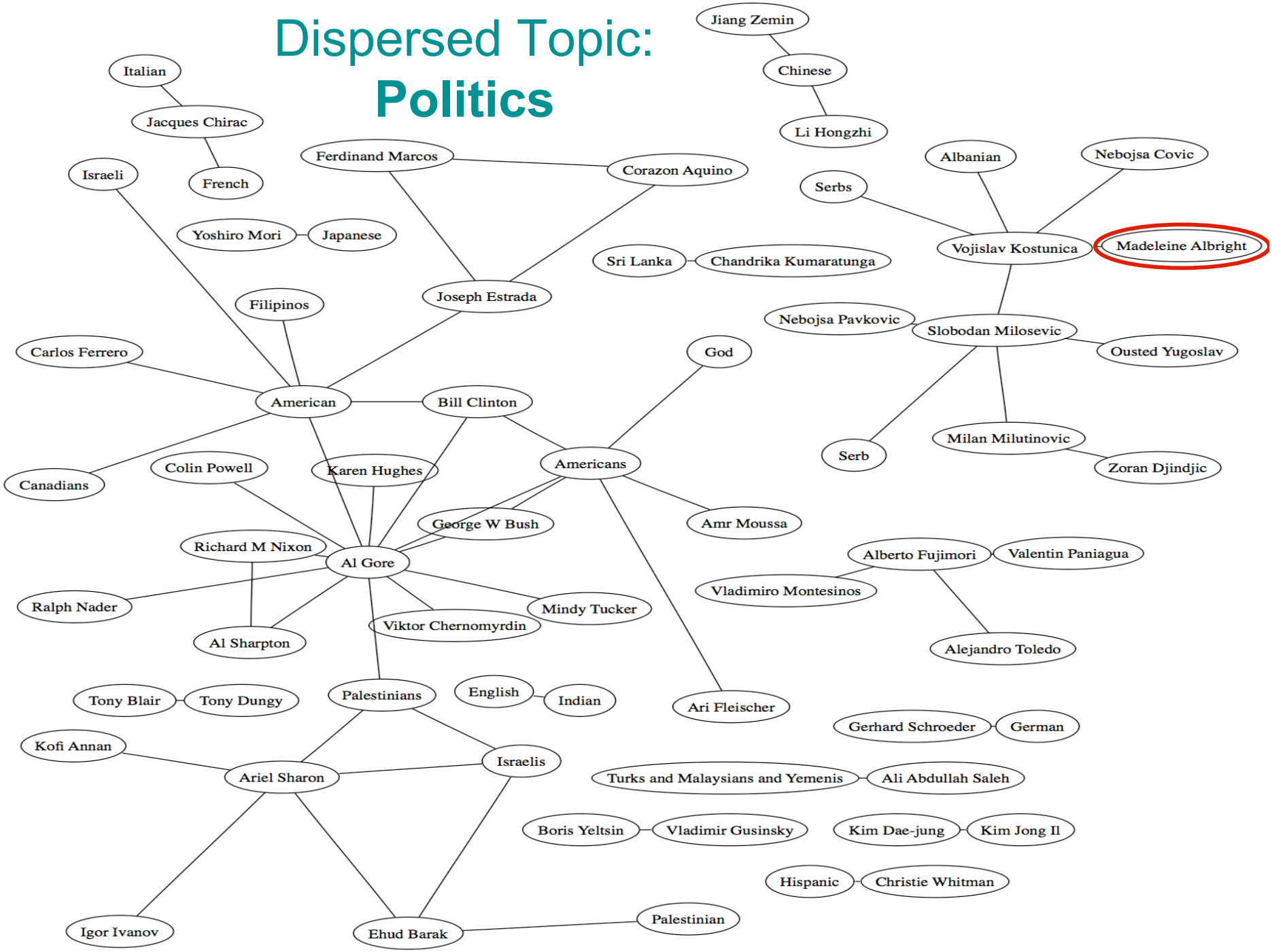
CRICKET -
MILLNS SIGNS FOR BOLAND

CAPE TOWN 1996-08-22

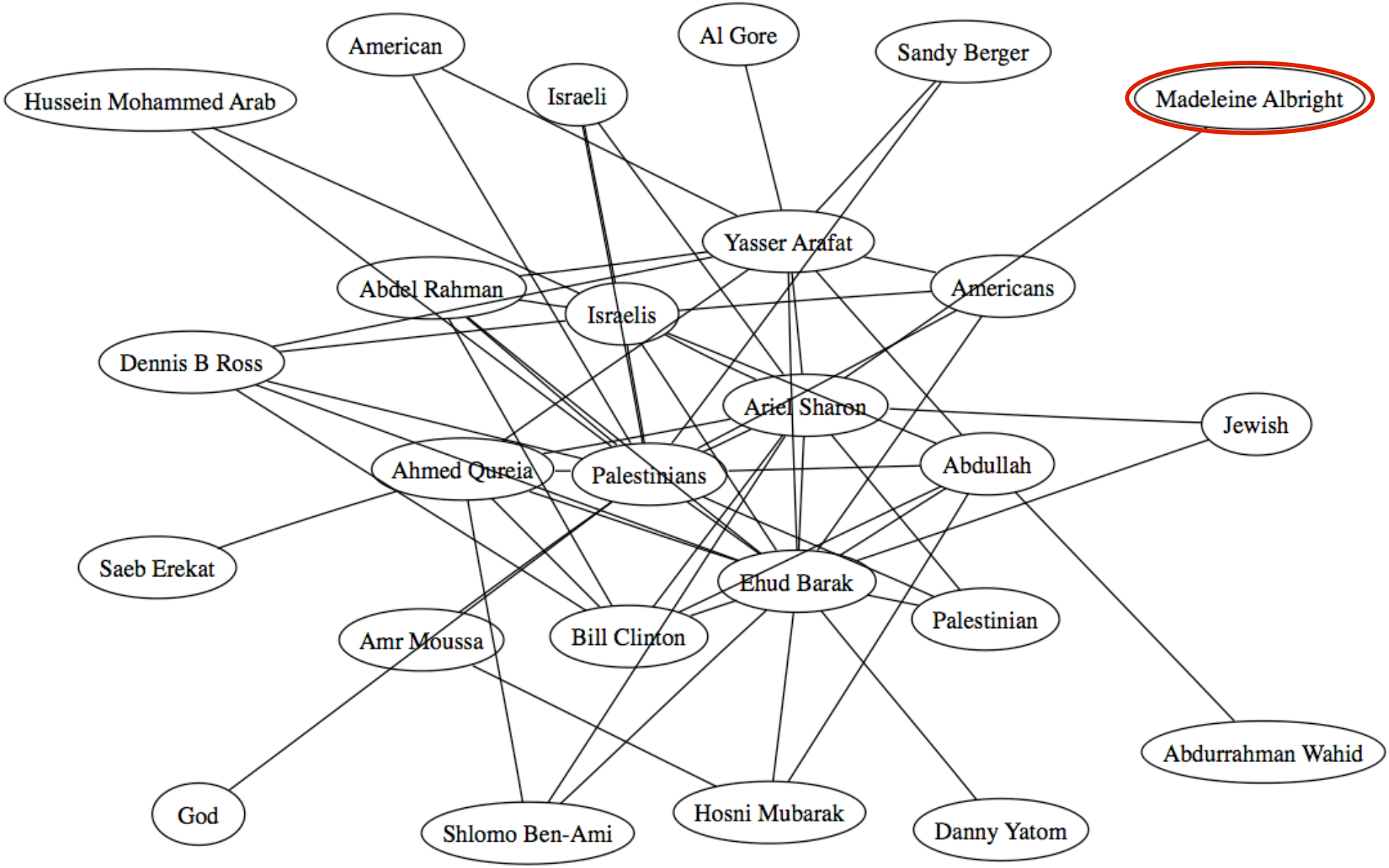
South African provincial side Boland said on Thursday they had signed Leicestershire fast bowler David Millns on a one year contract.
Millns, who toured Australia with England A in 1992, replaces former England all-rounder Phillip DeFreitas as Boland's overseas professional.

Labels:	Examples:
PER	Yayuk Basuki Innocent Butare
ORG	3M KDP Cleveland
LOC	Cleveland Nirmal Hriday The Oval
MISC	Java Basque 1,000 Lakes Rally

Dispersed Topic: **Politics**

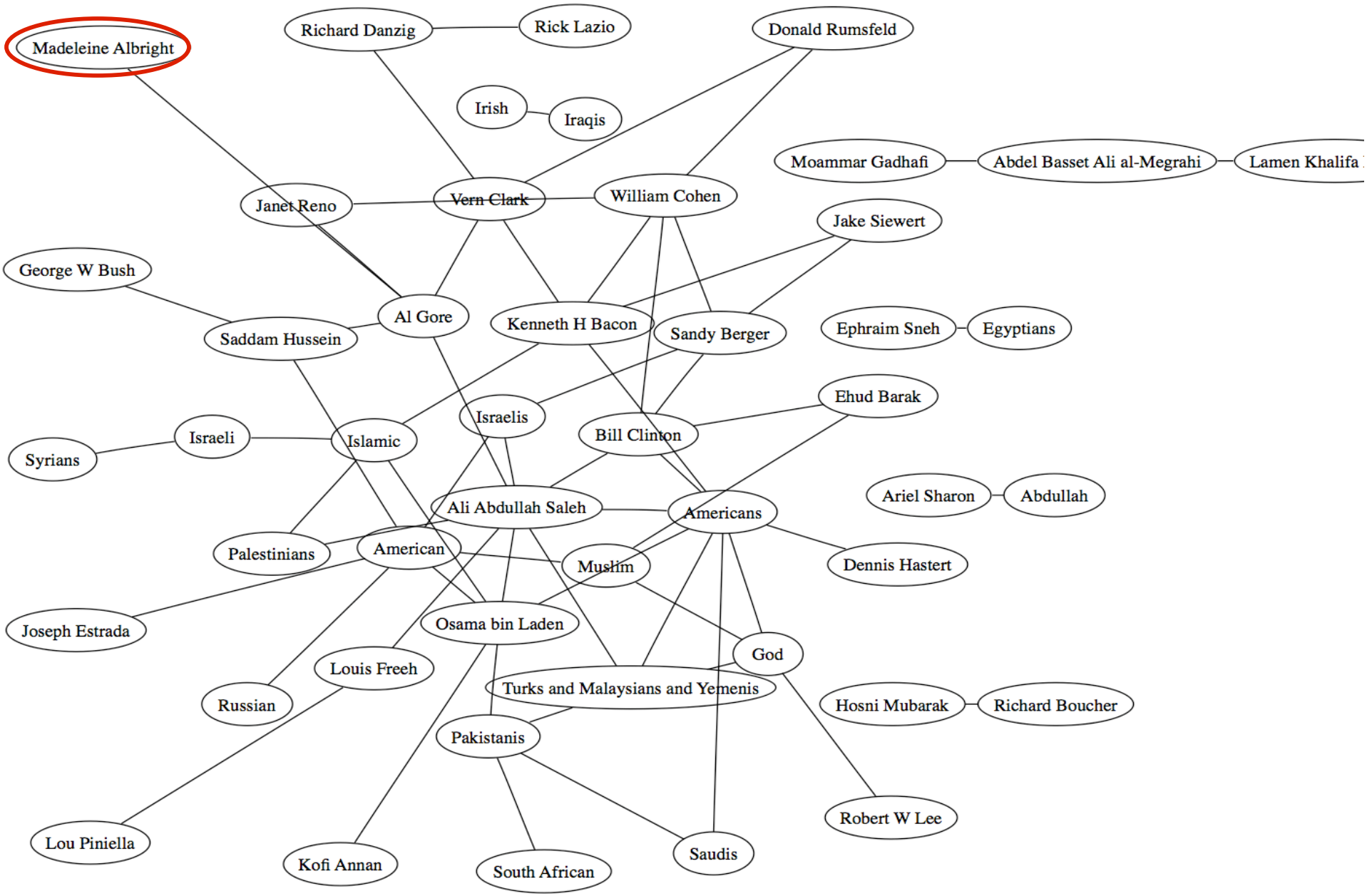


Densely Linked Topic: **Israel/Palestine**



palestinian israeli israel palestinians barak arafat peace violence minister clinton west bank jerusalem

USS Cole attack



officials attack cole navy yemen iraq military al bin saudi bombing laden security

Entities that co-occur with *Madeleine Albright*, by topic

Middle East	Serbia	Korea	Deal making
Ariel Sharon Sandy Berger Ehud Barak Abdel Rahman Dennis B Ross Al Gore Amr Moussa	Slobodan Milosevic Terry Madonna Vojislav Kostunica Serbs Radovan Karadic Jacques Chirac Sandy Berger	Al Gore Americans Colin Powell Kim Jong Il Chinese Jake Siewert George W Bush	Americans Sandy Berger Ariel Sharon Abdel Rahman Alberto Fujimori Edmond Pope Chinese

What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

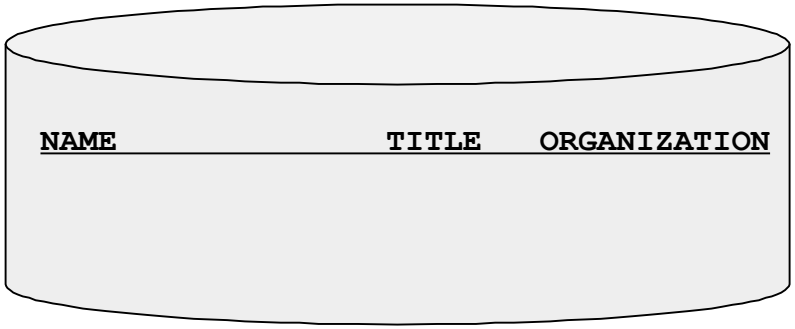
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a task: **Filling slots in a database from sub-segments of text.**

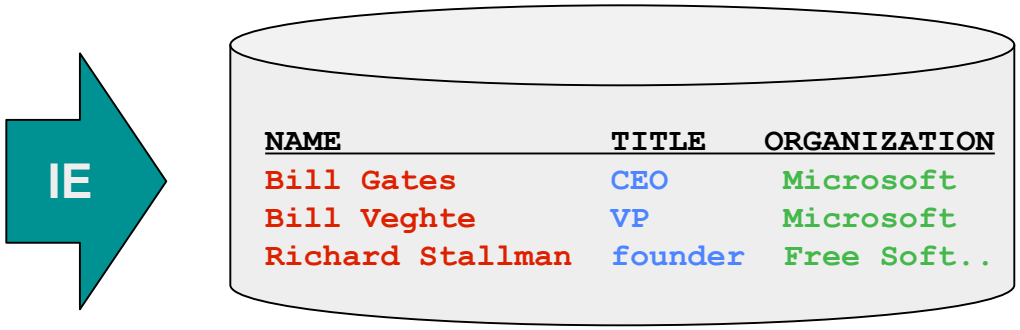
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft
Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

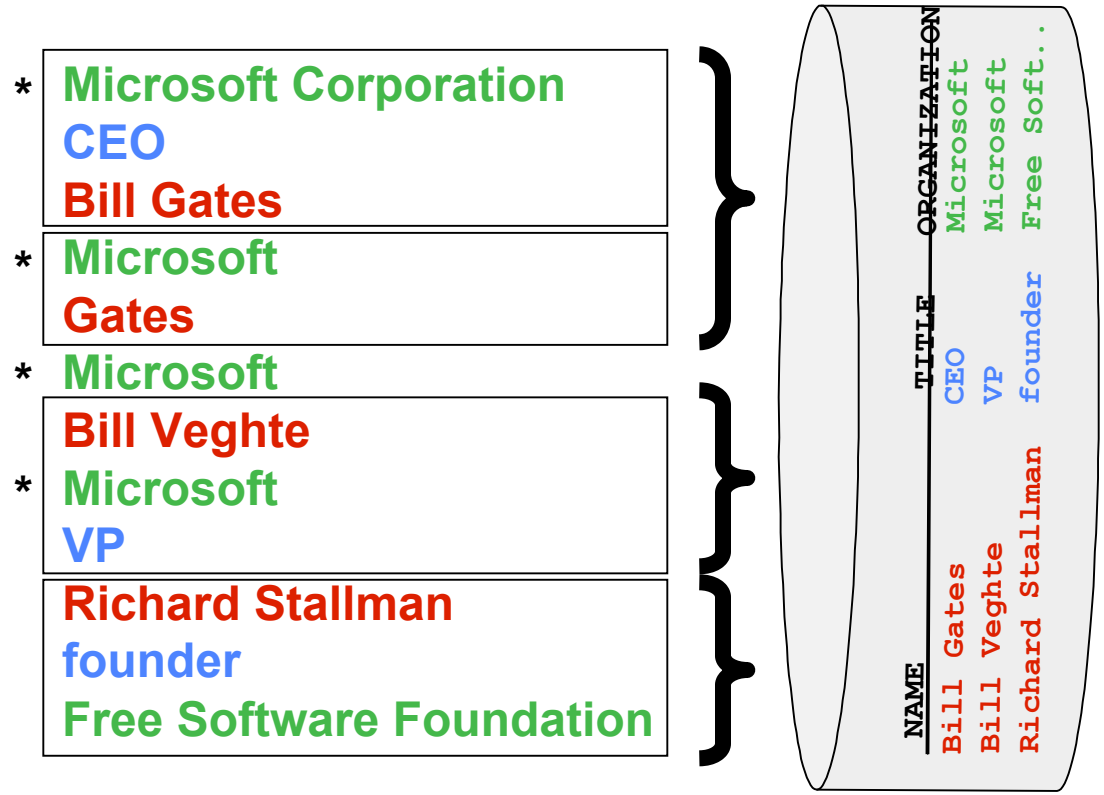
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

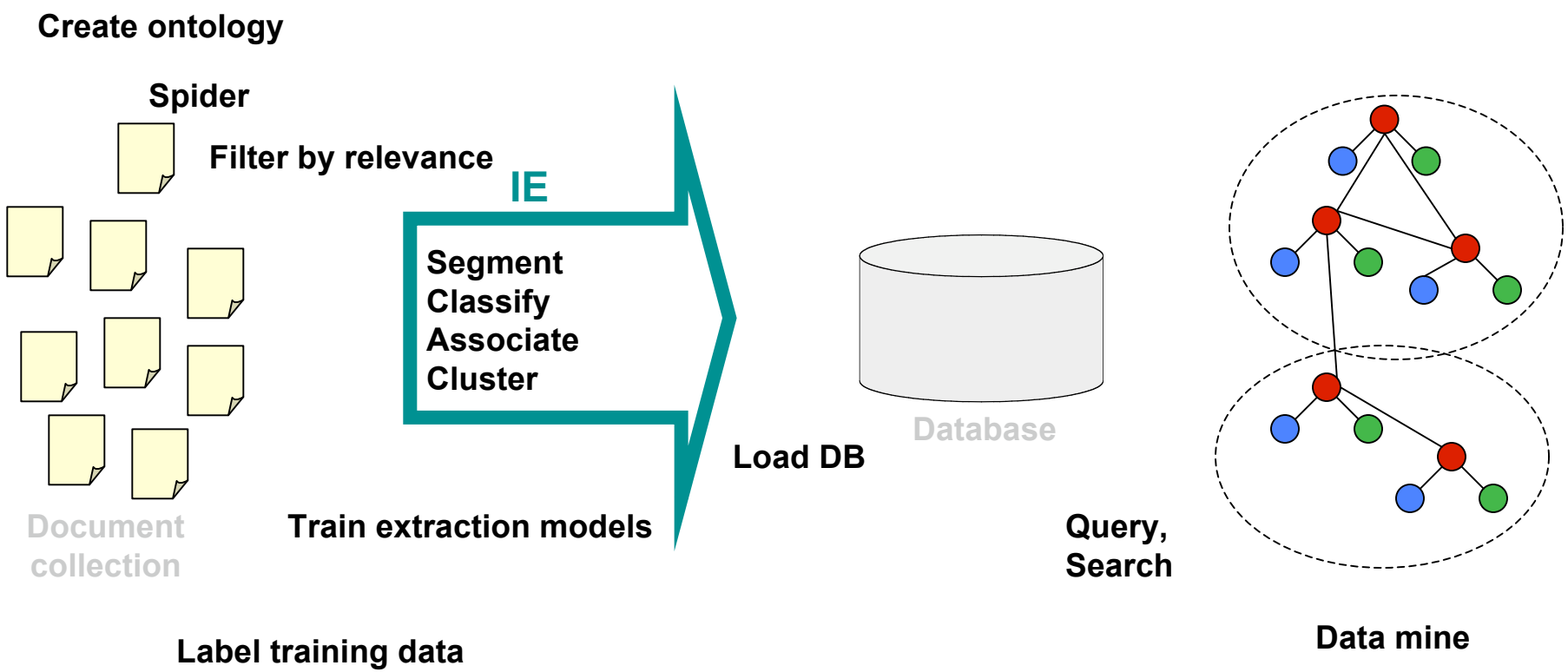
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



IE in Context



IE History

Pre-Web

- Mostly news articles
  - De Jong's *FRUMP* [1982]
    - Hand-built system to fill Schank-style "scripts" from news wire
  - *Message Understanding Conference (MUC)* DARPA ['87-'95], *TIPSTER* ['92-'96]
- Most early work dominated by hand-built models
  - E.g. SRI's *FASTUS*, hand-built FSMs.
  - But by 1990's, some machine learning: Lehnert, Cardie, Grishman and then HMMs: Elkan [Leek '97], BBN [Bikel et al '98]

Web

- AAAI '94 Spring Symposium on "Software Agents"
  - Much discussion of ML applied to Web. Maes, Mitchell, Etzioni.
- Tom Mitchell's WebKB, '96
  - Build KB's from the Web.
- Wrapper Induction
  - Initially hand-build, then ML: [Soderland '96], [Kushmeric '97],…

What makes IE from the Web Different?

Less grammar, but more formatting & linking

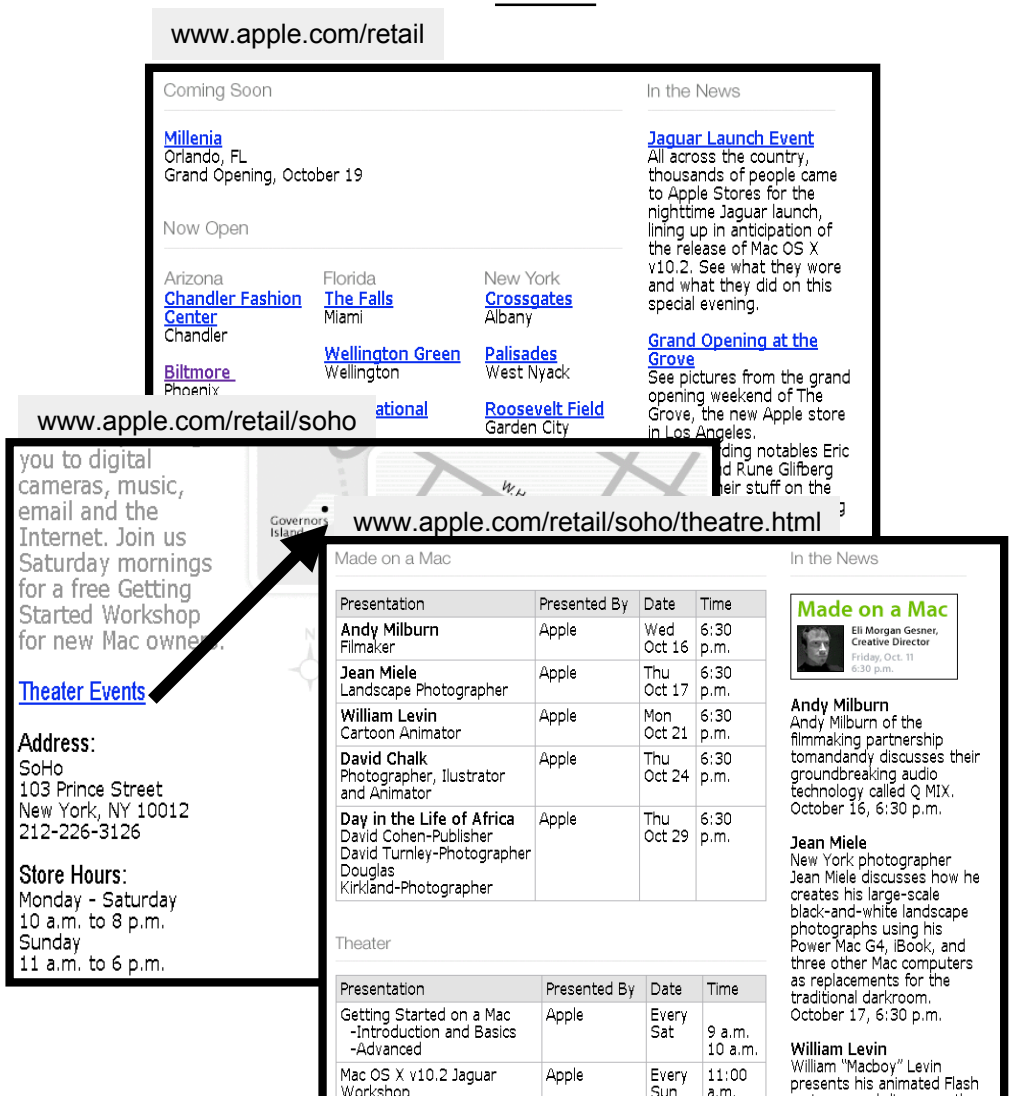
Newswire

Apple to Open Its First Retail Store in New York City

MACWORLD EXPO, NEW YORK--July 17, 2002--Apple's first retail store in New York City will open in Manhattan's SoHo district on Thursday, July 18 at 8:00 a.m. EDT. The SoHo store will be Apple's largest retail store to date and is a stunning example of Apple's commitment to offering customers the world's best computer shopping experience.

"Fourteen months after opening our first retail store, our 31 stores are attracting over 100,000 visitors each week," said Steve Jobs, Apple's CEO. "We hope our SoHo store will surprise and delight both Mac and PC users who want to see everything the Mac can do to enhance their digital lifestyles."

Web



The directory structure, link structure, formatting & layout of the Web is its own new grammar.

Landscape of IE Tasks (1/4): Pattern Feature Domain

Text paragraphs without formatting

Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

Grammatical sentences and some formatting & links

Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

- Press

Contact

- General information
- Directions maps

Non-grammatical snippets, rich formatting & links

Barto, Andrew G. (413) 545-2109 barto@cs.umass.edu CS276
Professor.
Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.

Berger, Emery D. (413) 577-4211 emery@cs.umass.edu CS344
Assistant Professor.

Brock, Oliver (413) 577-0334 oli@cs.umass.edu CS246
Assistant Professor.

Clarke, Lori A. (413) 545-1328 clarke@cs.umass.edu CS304
Professor.
Software verification, testing, and analysis; software architecture and design.

Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

Tables

8:30 - 9:30 AM	**Invited Talk: Plausibility Measures: A General Approach for Representing Uncertainty** *Joseph Y. Halpern, Cornell University*				
9:30 - 10:00 AM	Coffee Break				
10:00 - 11:30 AM	Technical Paper Sessions:				
Cognitive Robotics	**Logic Programming**	**Natural Language Generation**	**Complexity Analysis**	**Neural Networks**	**Games**
739: A Logical Account of Causal and Topological Maps *Emilio Remolina and Benjamin Kuipers*	116: A-System: Problem Solving through Abduction *Marc Denecker, Antonis Kakas, and Bert Van Nuffelen*	758: Title Generation for Machine-Translated Documents *Rong Jin and Alexander G. Hauptmann*	417: Let's go Nats: Complexity of Nested Circumscription and Abnormality Theories *Marco Cadoli, Thomas Eiter, and Georg Gottlob*	179: Knowledge Extraction and Comparison from Local Function Networks *Kenneth McGarry, Stefan Wermter, and John MacIntyre*	71: Iterative Widening *Tristan Cazenave*
549: Online-Execution of ccGolog Plans *Henrik Grosskreutz*	131: A Comparative Study of Logic Programs with	246: Dealing with Dependencies between Content Planning and	470: A Perspective on Knowledge Compilation	258: Violation-Guided Learning for Constrained	353: Temporal Difference Learning Applied to a

Landscape of IE Tasks (2/4): Pattern Scope

Web site specific

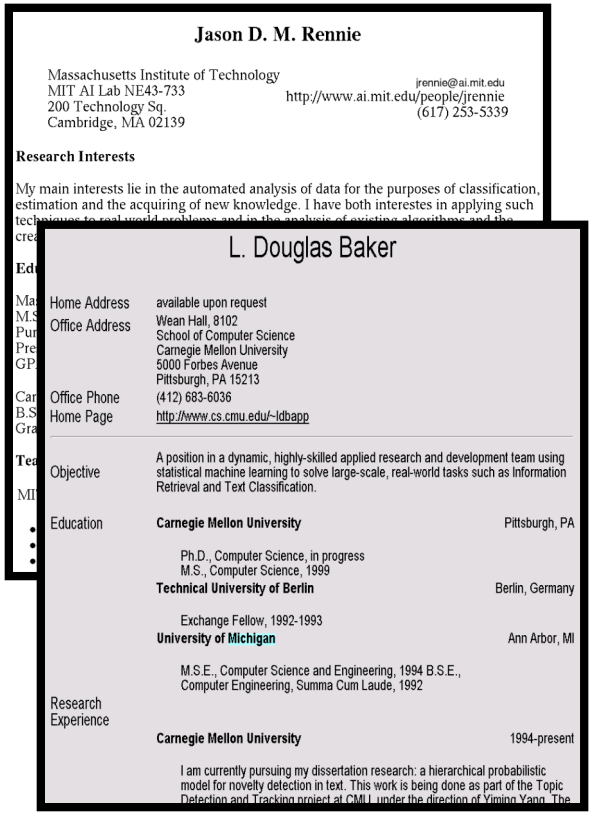
Formatting

Amazon.com Book Pages

Genre specific

Layout

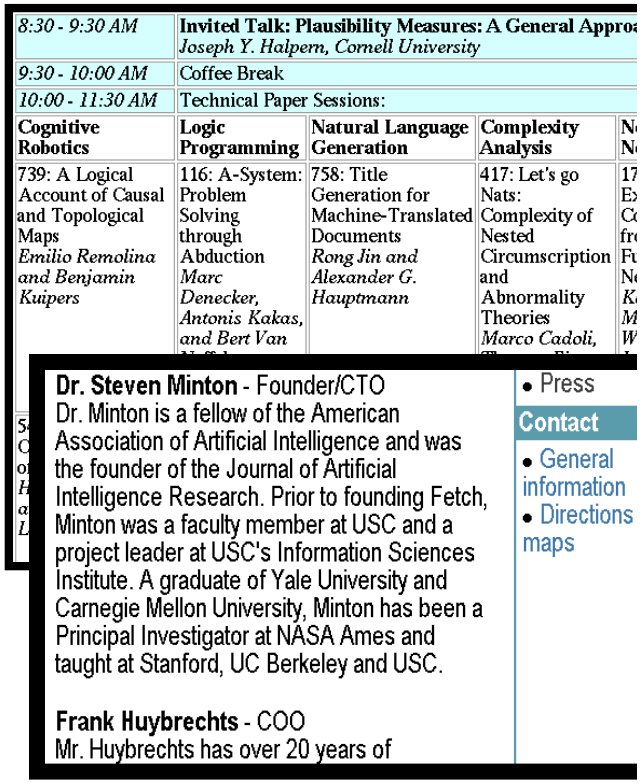
Resumes



Wide, non-specific

Language

University Names



Landscape of IE Tasks (3/4): Pattern Complexity

E.g. word patterns:

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

Person names

…was among the six houses sold by Hope Feldman that year.

Pawel Opalinski, Software Engineer at WhizBang Labs.

Landscape of IE Tasks (4/4): Pattern Combinations

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

Single entity

Person: **Jack Welch**

Person: **Jeffrey Immelt**

Location: **Connecticut**

Binary relationship

Relation: **Person-Title**
Person: **Jack Welch**
Title: **CEO**

Relation: **Company-Location**
Company: **General Electric**
Location: **Connecticut**

N-ary record

Relation: **Succession**
Company: **General Electric**
Title: **CEO**
Out: **Jack Welsh**
In: **Jeffrey Immelt**

"Named entity" extraction

Evaluation of Single Entity Extraction

TRUTH:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

PRED:

Michael Kearns and Sebastian Seung will start Monday's tutorial, followed by Richard M. Karpe and Martin Cooke.

$$\text{Precision} = \frac{\#\text{ correctly predicted segments}}{\#\text{ predicted segments}} = \frac{2}{6}$$

$$\text{Recall} = \frac{\#\text{ correctly predicted segments}}{\#\text{ true segments}} = \frac{2}{4}$$

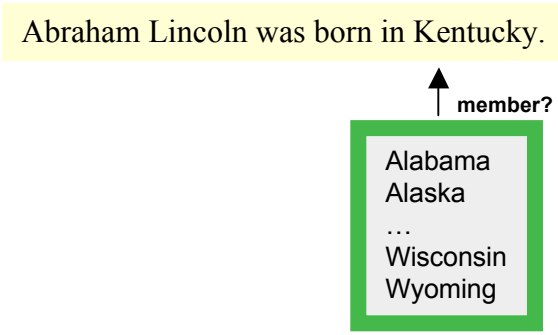
$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

State of the Art Performance

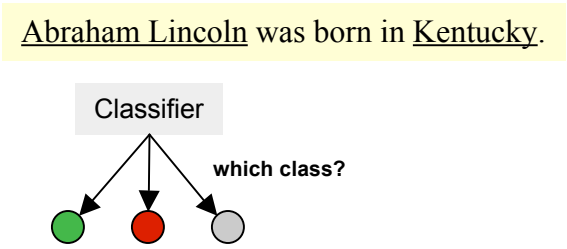
- Named entity recognition
  - Person, Location, Organization, …
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Wrapper induction
  - Extremely accurate performance obtainable
  - Human effort (~30min) required on each site

Landscape of IE Techniques (1/1): Models

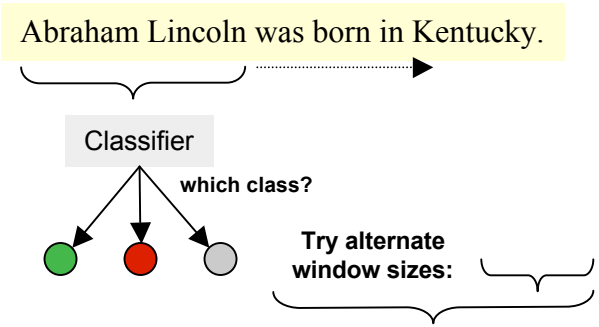
Lexicons



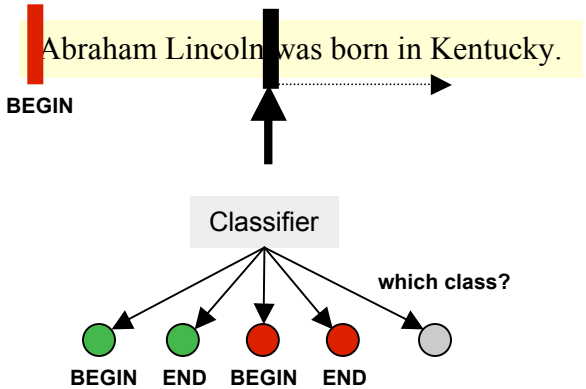
Classify Pre-segmented Candidates



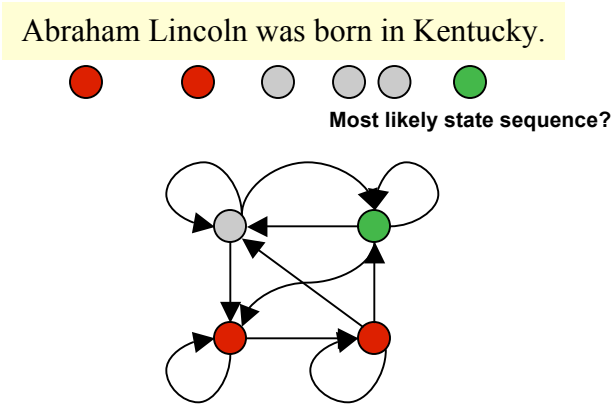
Sliding Window



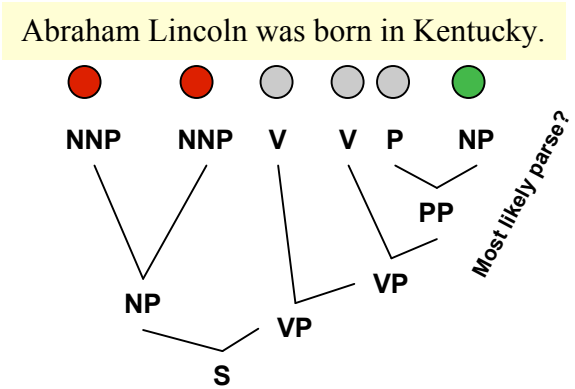
Boundary Models



Finite State Machines



Context Free Grammars



…and beyond

Any of these models can be used to capture words, formatting or both.

Sliding Windows

Extraction by Sliding Window

E.g.
Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

                3:30 pm
             7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Extraction by Sliding Window

E.g.
Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Extraction by Sliding Window

E.g.
Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

                3:30 pm
             7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

Extraction by Sliding Window

E.g.
Looking for seminar location

```
    GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

               3:30 pm
            7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence
during the 1980s and 1990s.   As a result
of its success and growth, machine learning
is evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning),
genetic algorithms, connectionist learning,
hybrid systems, and so on.
```

CMU UseNet Seminar Announcement

A “Naïve Bayes” Sliding Window Model

[Freitag 1997]

… **00 : pm Place : Wean Hall Rm 5409 Speaker : Sebastian Thrun** …

w_{t-m} … w_{t-1} w_t … w_{t+n} w_{t+n+1} … w_{t+n+m}

prefix **contents** **suffix**

P(“Wean Hall Rm 5409” = LOCATION) =

Ä

Prior probability of start position **Prior probability of length** **Probability prefix words** **Probability contents words** **Probability suffix words**

Try all start positions and reasonable lengths

Estimate these probabilities by (smoothed) counts from labeled training data.

If P(“Wean Hall Rm 5409” = LOCATION) **is above some threshold, extract it.**

Other examples of sliding window: ***[Baluja et al 2000]***
(decision tree over individual words & their context)

“Naïve Bayes” Sliding Window Results

Domain: CMU UseNet Seminar Announcements

```
    GRAND CHALLENGES FOR MACHINE LEARNING

            Jaime Carbonell
       School of Computer Science
      Carnegie Mellon University

                3:30 pm
             7500 Wean Hall

Machine learning has evolved from obscurity
in the 1970s into a vibrant and popular
discipline in artificial intelligence during
the 1980s and 1990s.   As a result of its
success and growth, machine learning is
evolving into a collection of related
disciplines: inductive concept acquisition,
analytic learning in problem solving (e.g.
analogy, explanation-based learning),
learning theory (e.g. PAC learning), genetic
algorithms, connectionist learning, hybrid
systems, and so on.
```

Field	F1
Person Name:	30%
Location:	61%
Start Time:	98%

Problems with Sliding Windows and Boundary Finders

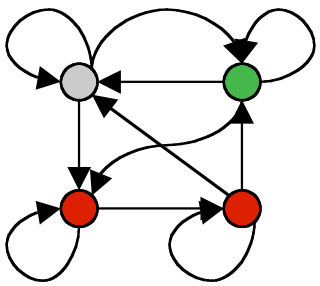
- Decisions in neighboring parts of the input are made independently from each other.
  - Naïve Bayes Sliding Window may predict a "seminar end time" before the "seminar start time".
  - It is possible for two *overlapping* windows to both be above threshold.
  - In a Boundary-Finding system, left boundaries are laid down independently from right boundaries, and their pairing happens as a separate step.

Finite State Machines

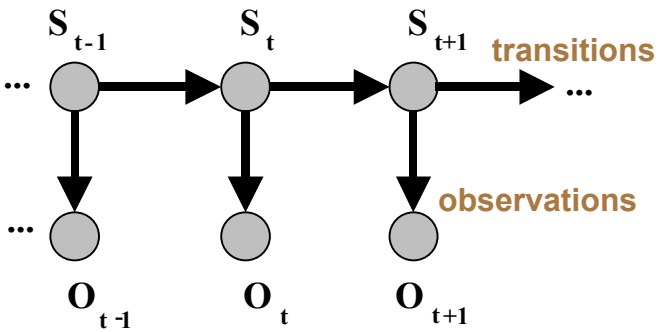
Hidden Markov Models

HMMs are the standard sequence modeling tool in genomics, music, speech, NLP, …

Finite state model

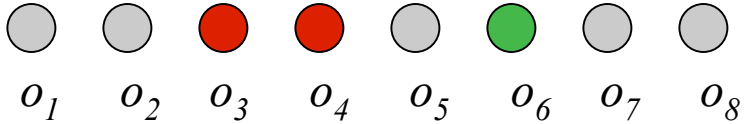


Graphical model



Generates:

State sequence
Observation sequence



$$P(\vec{s},\vec{o}) \propto \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

Parameters: for all states $S=\{s_1,s_2,\ldots\}$

Start state probabilities: $P(s_t)$

Transition probabilities: $P(s_t|s_{t-1})$

Observation (emission) probabilities: $P(o_t|s_t)$ Usually a multinomial over atomic, fixed alphabet

Training:

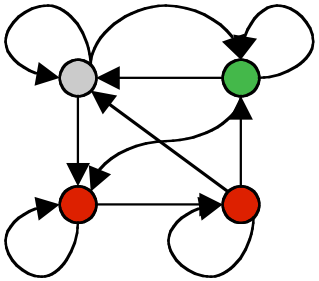
Maximize probability of training observations (w/ prior)

IE with Hidden Markov Models

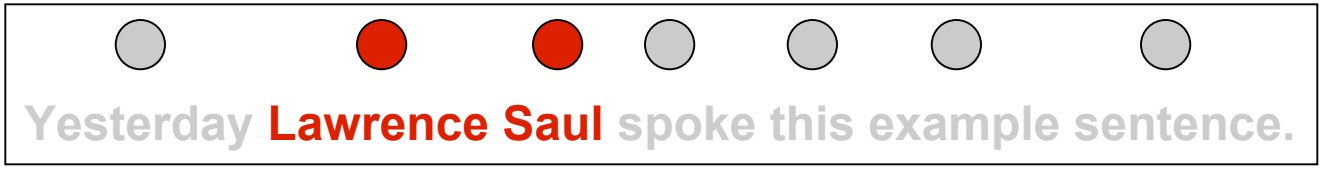
Given a sequence of observations:

Yesterday Lawrence Saul spoke this example sentence.

and a trained HMM:



Find the most likely state sequence: (Viterbi)



Any words said to be generated by the designated "person name" state extract as a person name:

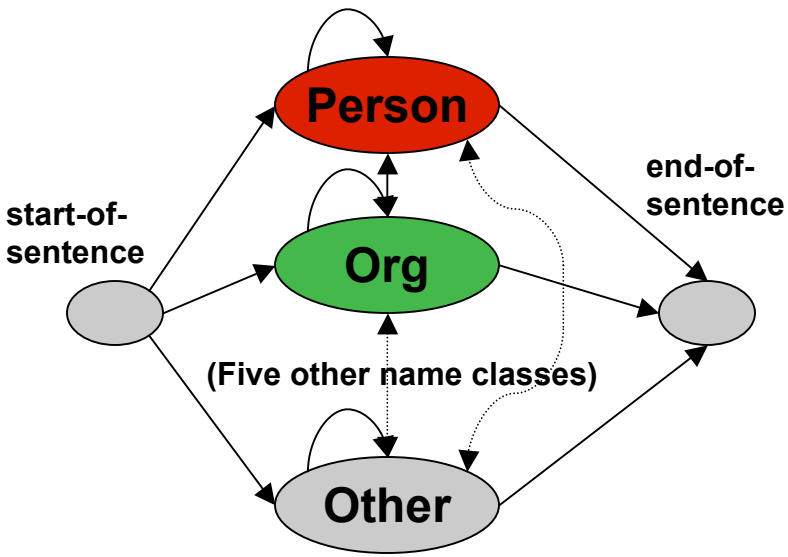
Person name: Lawrence Saul

HMMs for IE:
A richer model, with backoff

HMM Example: "Nymble"

[Bikel, et al 1998],
[BBN "IdentiFinder"]

Task: Named Entity Extraction



Transition probabilities

$$P(s_t \mid s_{t-1}, o_{t-1})$$

Back-off to:

$$P(s_t \mid s_{t-1})$$

$$P(s_t)$$

Observation probabilities

$$P(o_t \mid s_t, s_{t-1})$$

or $P(o_t \mid s_t, o_{t-1})$

Back-off to:

$$P(o_t \mid s_t)$$

$$P(o_t)$$

Train on 450k words of news wire text.

Results:

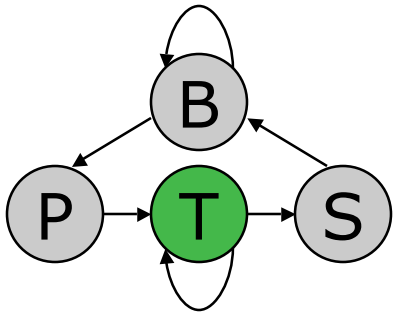
Case	Language	F1
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

Other examples of shrinkage for HMMs in IE: *[Freitag and McCallum '99]*

HMMs for IE:
Augmented finite-state structures with linear interpolation

Simple HMM structure for IE

- 4 state types:
  - **B**ackground (generates words not of interest),
  - **T**arget (generates words to be extracted),
  - **P**refix (generates typical words preceding target)
  - **S**uffix (words typically following target)



- Properties:
  - Extracts one type of target (e.g. target = person name), we will build one model for each extracted type.
  - Models different Markov-order n-grams for different predicted state contexts.
  - even thought there are multiple states for “Background”, state-path given labels is unambiguous. Therefore model parameters can all be computed using counts from labeled training data

More rich prefix and suffix structures

- In order to represent more context, add more state structure to prefix, target and suffix.
- But now overfitting becomes more of a problem.

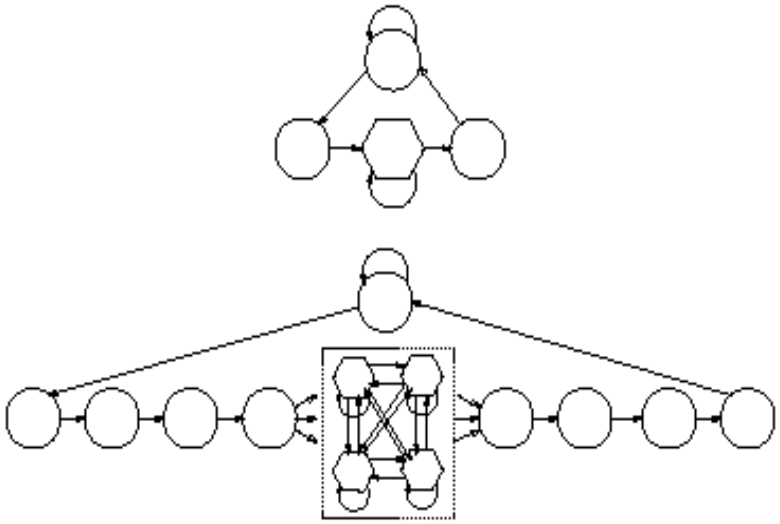
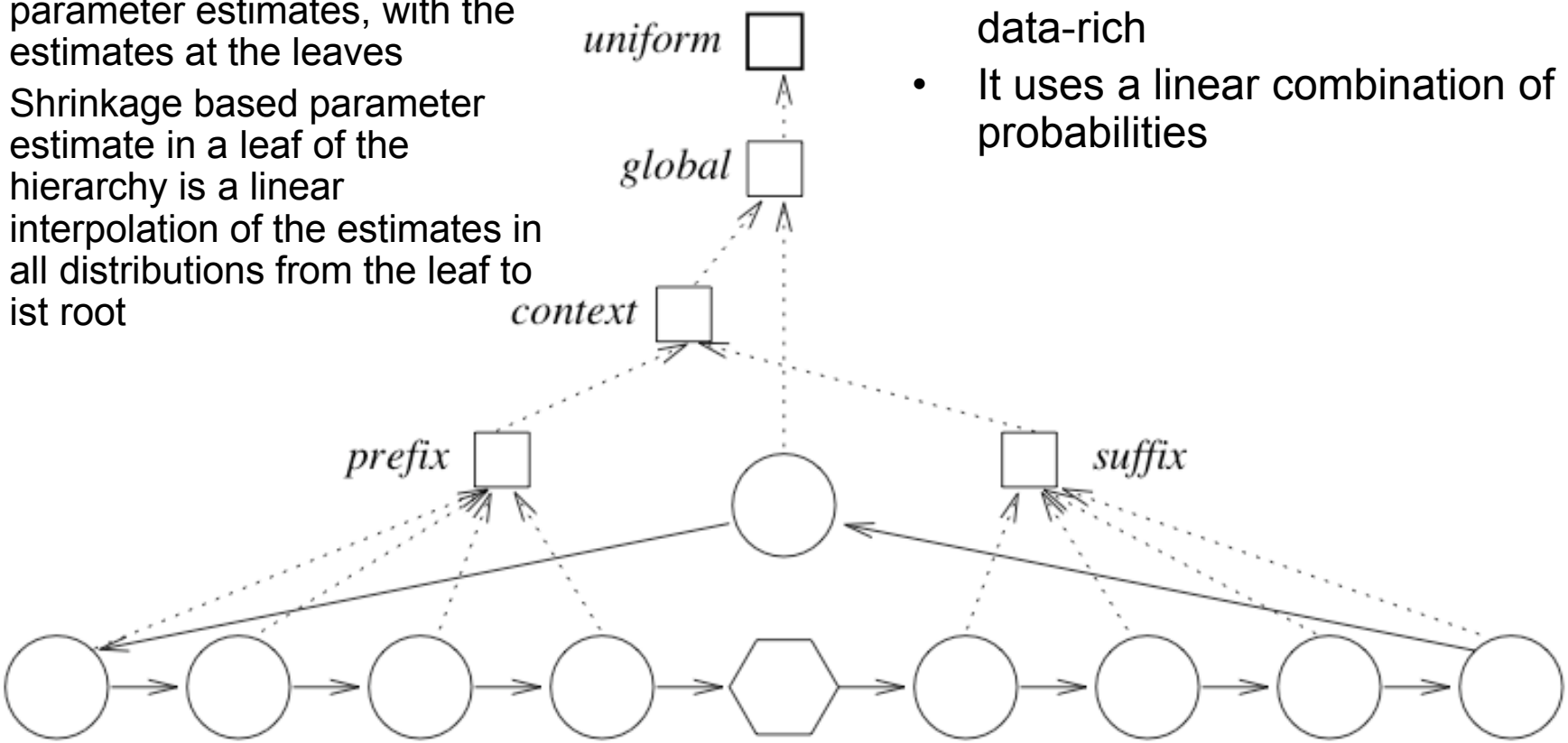


Figure 1: Two example HMM structures. Circle nodes represent non-target states; hexagon nodes represent target states.

Linear interpolation across states

- Is defined in terms of some hierarchy that represents the expected similarity between parameter estimates, with the estimates at the leaves
- Shrinkage based parameter estimate in a leaf of the hierarchy is a linear interpolation of the estimates in all distributions from the leaf to ist root



- Shrinkage smoothes the distribution of a state towards that of states that are more data-rich
- It uses a linear combination of probabilities

Evaluation of linear interpolation

- Data set of seminar announcements.

	speaker	*location*	*stime*	*etime*
None	0.513	0.735	0.991	0.814
Uniform	0.614	0.776	0.991	0.933
Global	0.711	0.839	0.991	0.595
Hier.	0.672	0.850	0.987	0.584

Table 4: Effect on F1 performance of different shrinkage configurations on four seminar announcement fields, given a topology with a window size of four and four parallel length-differentiated target paths.

IE with HMMs:
Learning Finite State Structure

Information Extraction from Research Papers

References

Leslie Pack Kaelbling, Michael L. Littman and Andrew W. Moore. Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research, pages 237-285, May 1996.

Headers

Journal of Artificial Intelligence Research 4 (1996) 237-285 Submitted 9/95; published 5/96

Reinforcement Learning: A Survey

Leslie Pack Kaelbling LPK@CS.BROWN.EDU
Michael L. Littman MLITTMAN@CS.BROWN.EDU
Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA

Andrew W. Moore AWM@CS.CMU.EDU
Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

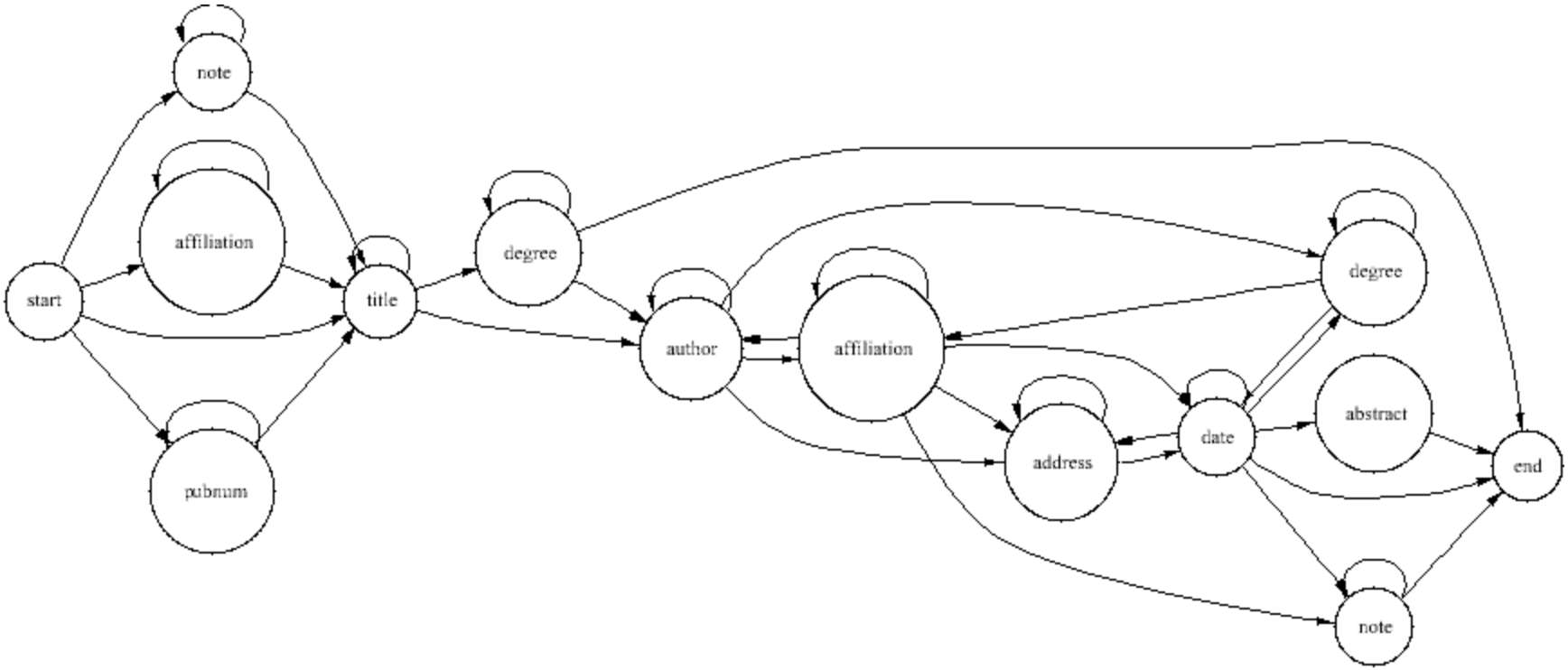
This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad selection of current work are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through trial-and-error interactions with a dynamic environment. The work described here has a resemblance to work in psychology, but differs considerably in the details and in the use of the word "reinforcement." The paper discusses central issues of reinforcement learning, including trading off exploration and exploitation, establishing the foundations of the field via Markov decision theory, learning from delayed reinforcement, constructing empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in statistics,

Information Extraction with HMMs

[Seymore & McCallum '99]



Importance of HMM Topology

- Certain structures better capture the observed phenomena in the prefix, target and suffix sequences
- Building structures by hand does not scale to large corpora
- Human intuitions don’t always correspond to structures that make the best use of HMM potential

Structure Learning

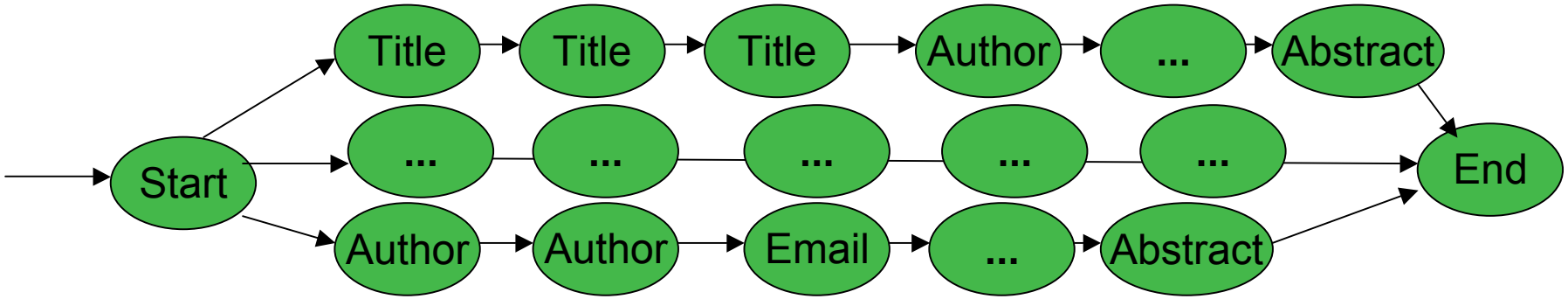
Two approaches

- Bayesian Model Merging
  - Neighbor-Merging
  - V-Merging

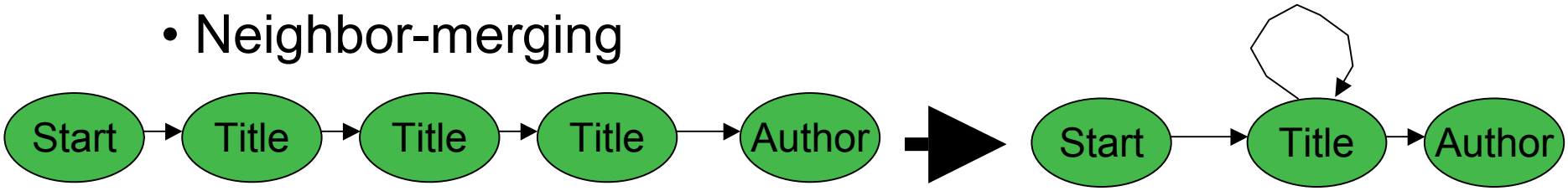
- Stochastic Optimization
  - Hill Climbing in the possible structure space by spiltting states and gauging performance on a validation set

Bayesian Model Merging

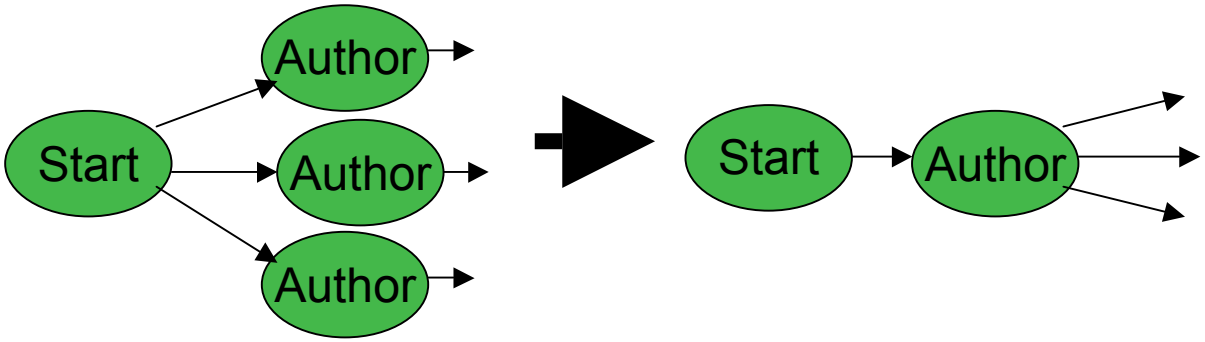
- Maximally Spesific Model



- Neighbor-merging



- V-merging

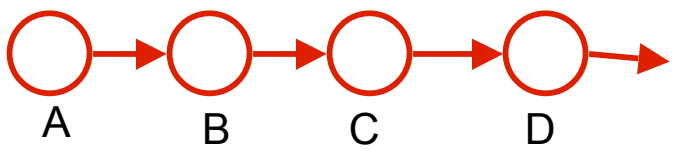


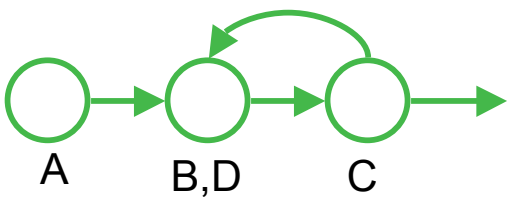
Bayesian Model Merging

- Iterates merging states until an optimal tradeoff between fit to the data and model size has been reached

$$P(M \mid D) \sim P(D \mid M)\, P(M)$$

M = Model
D = Data

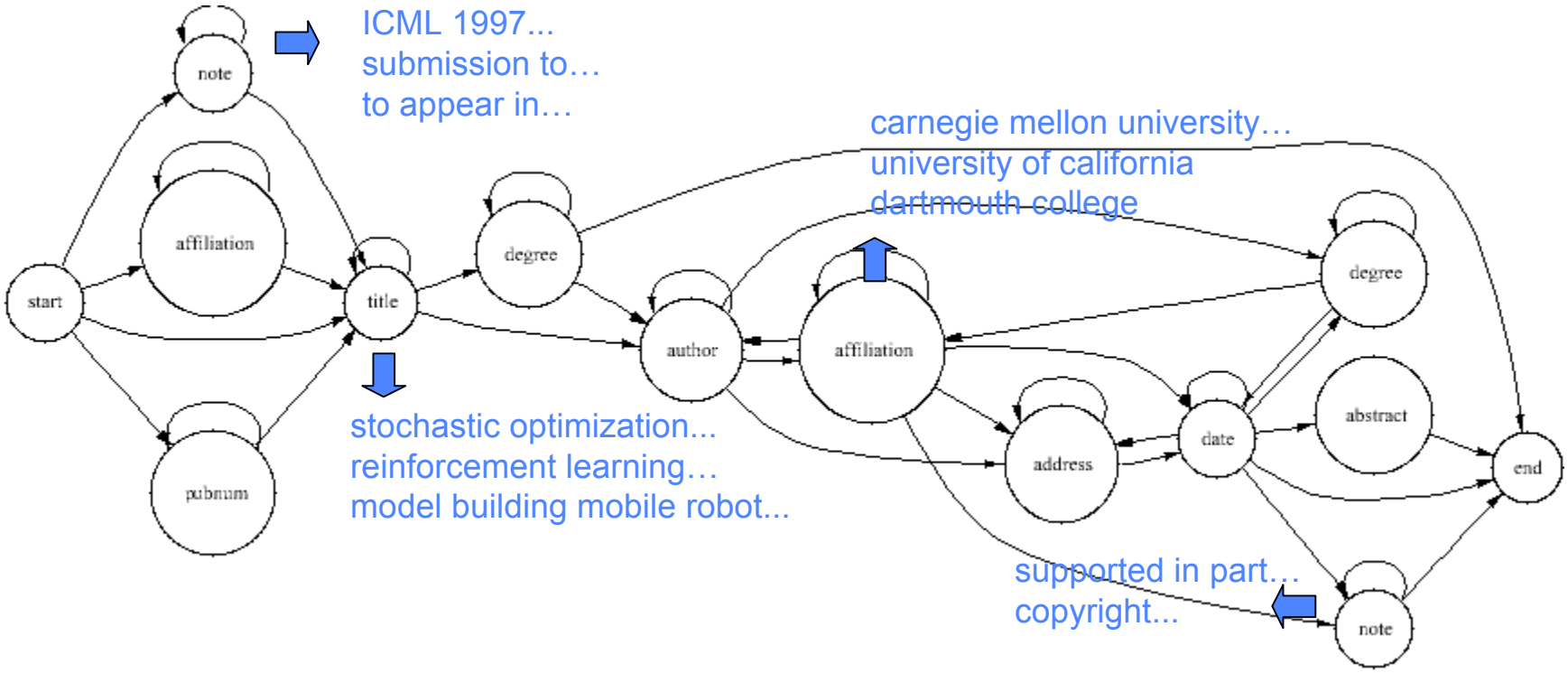


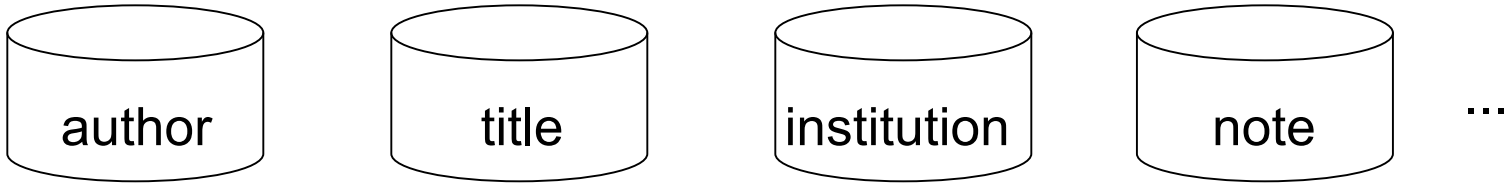


P(D | M) can be calculated with the Forward algorithm

P(M) model prior can be formulated to reflect a preference for smaller models

HMM Emissions





2 million words of BibTeX data from the Web

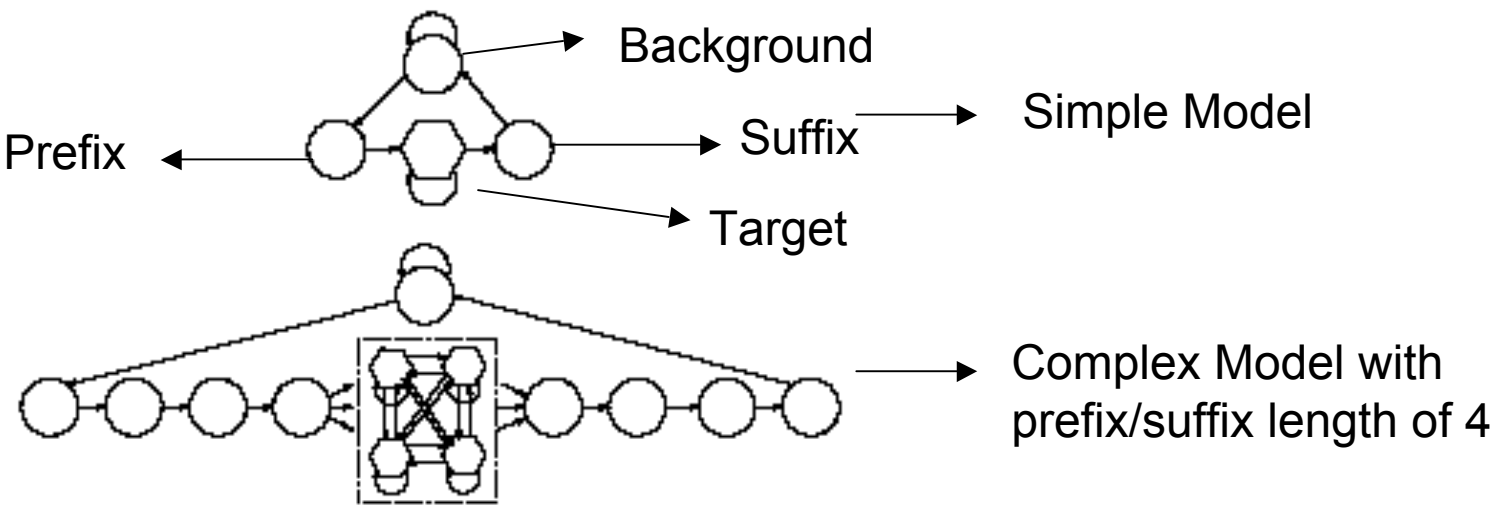
HMM Information Extraction Results

Per-word error rate

	Headers	References
One state/class Labeled data only	0.095	
Model Merging Labeled data only	0.087 *(8% better)*	
One state/class +BibTeX data	0.076 *(20% better)*	
Model Merging +BibTeX	0.071 *(25% better)*	0.066

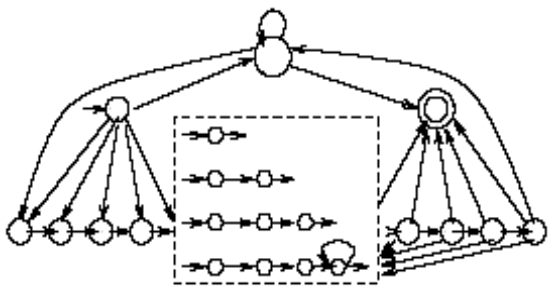
Stochastic Optimization

- Start with a simple model
- Perform hill-climbing in the space of possible structures
- Make several runs and take the average to avoid local optima



State Operations

- Lengthen a prefix
- Split a prefix
- Lengthen a suffix
- Split a suffix
- Lengthen a target string
- Split a target string
- Add a background state



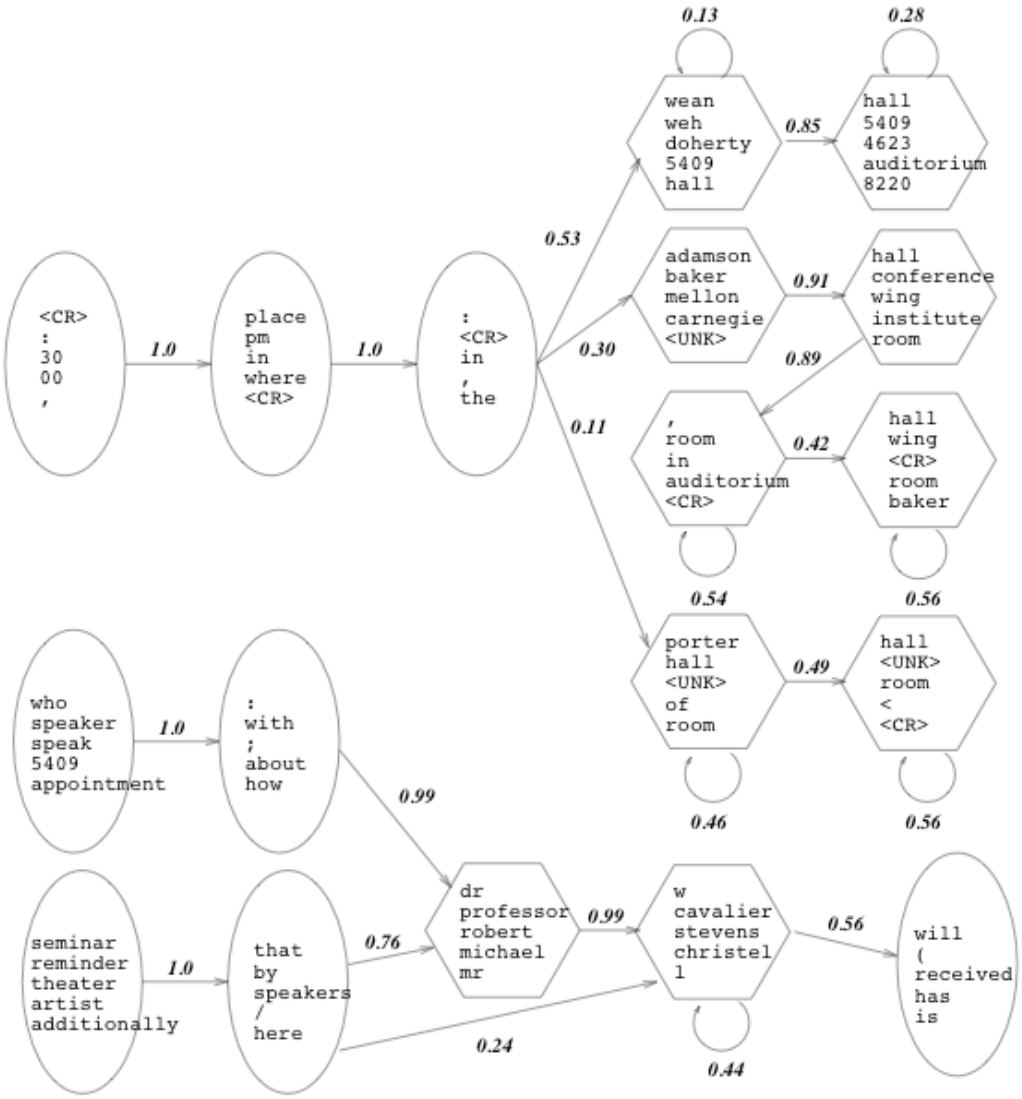
LearnStructure Algorithm

```
procedure LearnStructure(LabeledSet, Ops)
    ValidSet ← 1/3 of LabeledSet
    TrainSet ← LabeledSet − ValidSet
    CurModel ← the simple model
    Keepers ← {CurModel}
    I ← 0
    while I < 20 and CurModel has fewer than 25 states
        Candidates ← {M | M ∈ op(CurModel) ∧ op ∈ Ops}
        for M ∈ Candidates
            score(M) ← average of 3 runs trained on
                TrainSet and scored for F1 on ValidSet
        CurModel ← M ∈ Candidates with highest score
        Keepers ← Keepers ∪ {CurModel}
        I ← I + 1
    for M ∈ Keepers
        score(M) ← average F1 from
            3-fold cross-validation on LabeledSet
    return M ∈ Keepers with highest score
```

Part of Example Learned Structure

Locations

Speakers



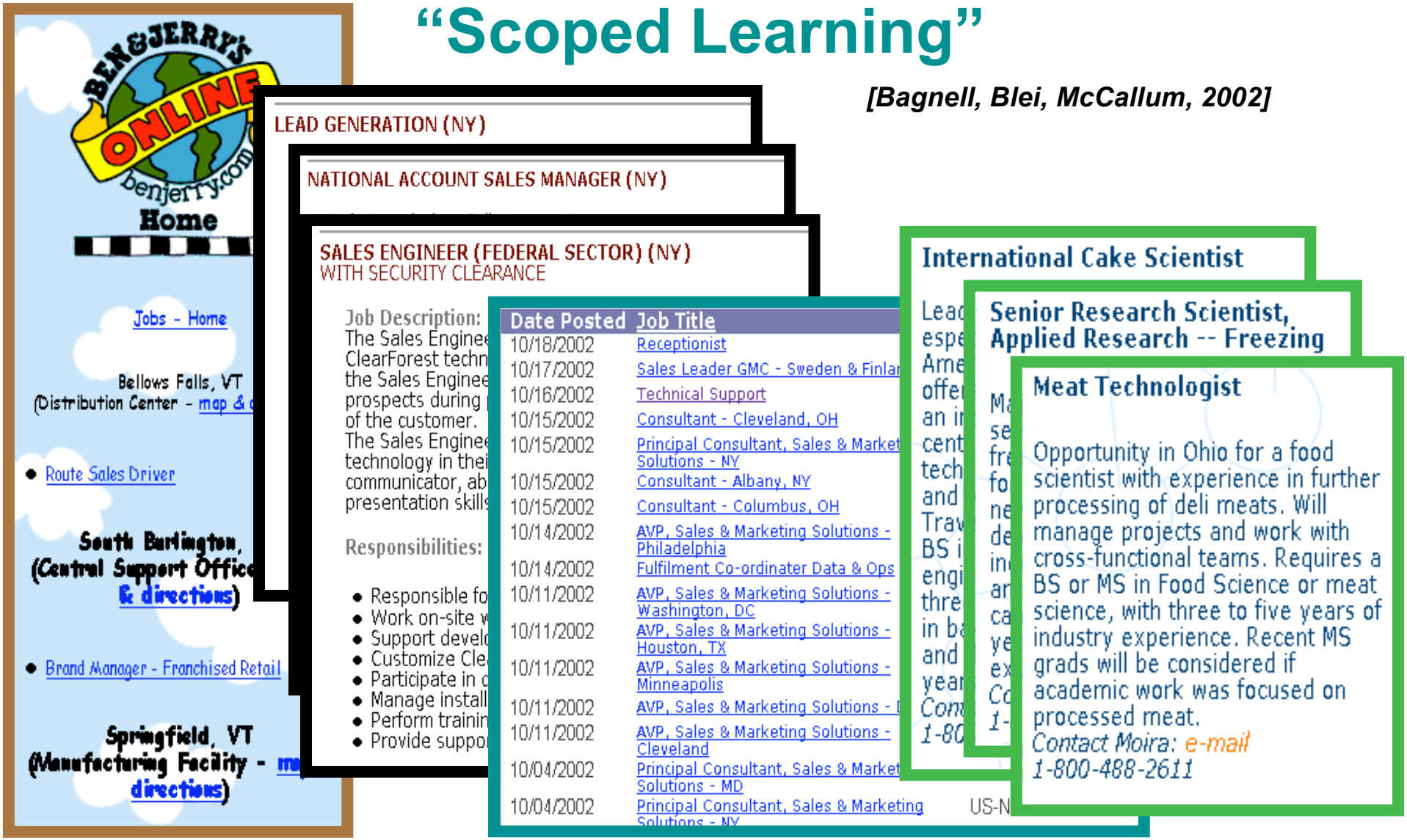
Accuracy of Automatically-Learned Structures

	speaker	*location*	*acquired*	*dlramt*	*title*	*company*	*conf*	*deadline*	***Average***
Grown HMM	76.9	87.5	41.3	54.4	58.3	65.4	27.2	46.5	57.2
vs. SRV	+19.8	+16.0	+1.1	-1.6	—	—	—	—	+8.8
vs. Rapier	+23.9	+14.8	+12.5	+15.1	-11.7	+24.9	—	—	+13.3
vs. Simple HMM	+24.3	+5.6	+14.3	+5.6	+5.7	+11.1	+15.7	+6.7	+11.1
vs. Complex HMM	-2.1	+6.7	+7.5	-0.3	-0.3	+19.1	+0.0	-6.8	+3.0

Table 2: Difference in F1 performance between the HMM using a learned structure and other methods. The + numbers indicate how much better our Grown HMM did than the alternative method.

Learning Formatting Patterns "On the Fly": "Scoped Learning"

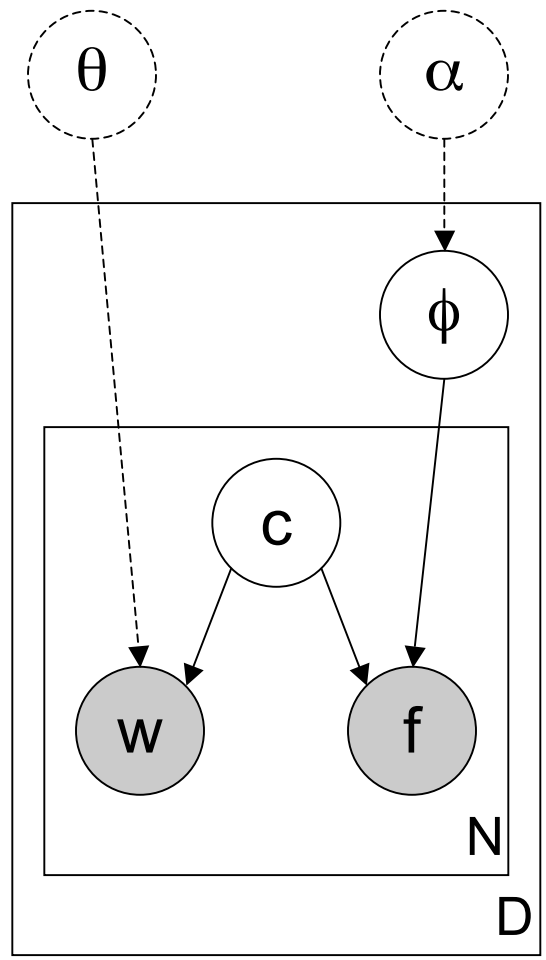
[Bagnell, Blei, McCallum, 2002]



Formatting is regular on each site, but there are too many different sites to wrap. Can we get the best of both worlds?

Scoped Learning Generative Model

1. For each of the D documents:
   a) Generate the multinomial formatting feature parameters ϕ from $p(\phi|\alpha)$
2. For each of the N words in the document:
   a) Generate the *n*th category c_n from $p(c_n)$.
   b) Generate the *n*th word (global feature) from $p(w_n|c_n, \theta)$
   c) Generate the *n*th formatting feature (local feature) from $p(f_n|c_n, \phi)$



$$p(\phi, \mathbf{c}, \mathbf{w}, \mathbf{f}) = p_\alpha(\phi) \prod_{n=1}^{N} p(c_n) p_\theta(w_n|c_n) p(f_n|c_n, \phi)$$

Inference

Given a new web page, we would like to classify each word resulting in **c** = {c_1, c_2,…, c_n}

$$p(\mathbf{c}|\mathbf{w},\mathbf{f}) = \frac{\int \prod_{n=1}^{N} p(w_n|c_n)p(f_n|c_n,\phi)p(c_n)p(\phi)d\phi}{\int \prod_{n=1}^{N} \sum_{c_n} p(w_n|c_n)p(f_n|c_n,\phi)p(c_n)p(\phi)d\phi}$$

This is not feasible to compute because of the integral and sum in the denominator. We experimented with two approximations:

- MAP point estimate of ϕ
- Variational inference

MAP Point Estimate

If we approximate ϕ with a point estimate, $\hat{\phi}$, then the integral disappears and c decouples. We can then label each word with:

$$\hat{c}_n = \arg\max_{c_n} p(w_n|c_n)p(f_n|c_n, \hat{\phi})p(c_n)$$

A natural point estimate is the posterior mode: a maximum likelihood estimate for the local parameters given the document in question:

$$\hat{\phi} = \arg\max_{\phi} p(\phi|\mathbf{f}, \mathbf{w})$$

E-step:

$$p^{(t+1)}(c_n|w_n, f_n; \phi) \propto p^{(t)}(f_n|c_n; \phi)p(w_n|c_n)p(c_n)$$

M-step:

$$\hat{\phi}_{c,f} = p^{(t+1)}(f|c; \phi) \propto \sum_{\{n:c_n=c, f_n=f\}} p^{(t)}(c_n|f_n, w_n)$$

Work at the Y! - Microsoft Internet Explorer provided by WhizBang! Labs

File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\Documents and Settings\mccallum\Desktop\scoped\jobopening-global.html Links

Family Services Director Creative, energetic and enjoy working with people? Seeking director for program development, implementation and administration. Must possess a Bachelor's Degree in Recreation, Family Studies or related field. Strong interpersonal and organizational skills a must. Excellent benefits. Send resumes to Jane Kim, Dir of Camping and Family Services, North Suburban YMCA, 2705 Techny Road, Northbrook, IL 60062.

Massage Therapist - Male The North Suburban YMCA is seeking a certified massage therapist to work part time in our men's program center. Flexible hours, y membership, on-site child care available if needed. Please contact Harlan Stritchko by email or call at 847-272-7250.

Starbucks Server Early day, evening and weekend shifts available for in-house cafe serving the Starbuck's product line. An exciting opportunity and membership is included! Contact Sarah Tucker at 847-272-7250 x.213.

Teacher for ChildCare Center Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early childhood or Education or similar subject. At least one year experience working with 2-5 year olds. Contact Helen at (847) 272-7250 x222 and fax resume to (847) 272-7587.

Art Coordinator Creative? Enjoy working with children? the North suburban Y is looking for an art coordinator for the summer. Call Jane at (847)272-7250 for more information.

Teachers Seeking part-time early childhood teachers for summer or all year. 2-3 mornings per week from 9am-11:15am. Free child care on-site while you work. Free YMCA membership. College degree required in education or related field. Pick up an application at the front desk or call Caryn Shulman, Child Development Coordinator at (847) 272-7250 x232.

Group Exercise Personal Training Interested individuals with proper certification may contact Myleen Signorini at (847) 272-7250 x 217

Cutomer Service Rep **OVERQUALIFIED APPLY HERE!** Hone your skills by working in a friendly environment. The front desk is looking for part time staff to work flexible shifts for early weekday mornings, day and evening shifts. Benefits include YMCA membership and babysitting during your shift. Please contact Sarah Tucker or Cheryl Stewart at (847) 272-7250 x 213.

Lifeguards and Swim Instructors Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and

Done My Computer

Global Extractor: Precision = 46%, Recall = 75%

Work at the Y! - Microsoft Internet Explorer provided by WhizBang! Labs

File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\Documents and Settings\mccallum\Desktop\scoped\jobopening-localglobal.html Links

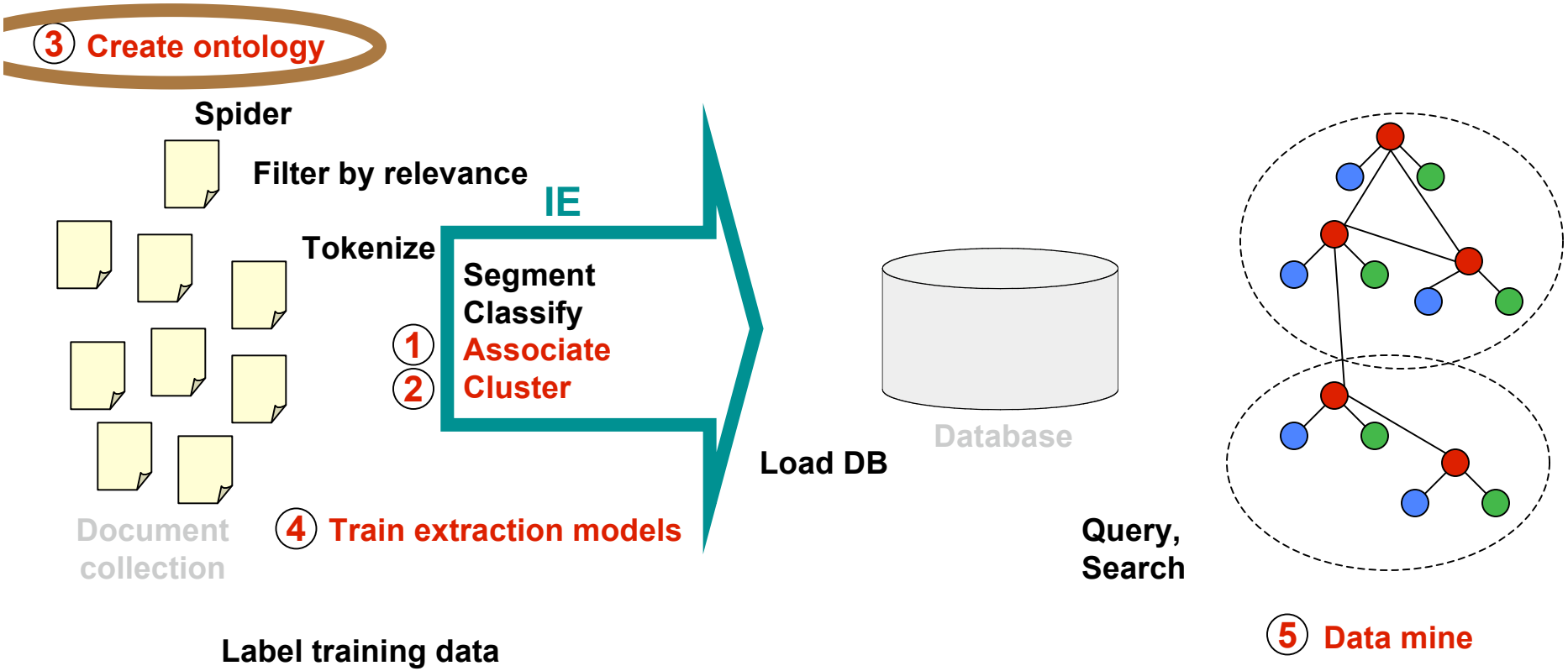
Family Services Director	Creative, energetic and enjoy working with people? Seeking director for program development, implementation and administration. Must possess a Bachelor's Degree in Recreation, Family Studies or related field. Strong interpersonal and organizational skills a must. Excellent benefits. Send resumes to Jane Kim, Dir of Camping and Family Services, North Suburban YMCA, 2705 Techny Road, Northbrook, IL 60062.
Massage Therapist - Male	The North Suburban YMCA is seeking a certified massage therapist to work part time in our men's program center. Flexible hours, y membership, on-site child care available if needed. Please contact Harlan Stritchko by email or call at 847-272-7250.
Starbucks Server	Early day, evening and weekend shifts available for in-house cafe serving the Starbuck's product line. An exciting opportunity and membership is included! Contact Sarah Tucker at 847-272-7250 x.213.
Teacher for ChildCare Center	Part-time 2-6 pm, Monday through Friday. Minimum requirements are 60 college credit hours in Early childhood or Education or similar subject. At least one year experience working with 2-5 year olds. Contact Helen at (847) 272-7250 x222 and fax resume to (847) 272-7587.
Art Coordinator	Creative? Enjoy working with children? the North suburban Y is looking for an art coordinator for the summer. Call Jane at (847)272-7250 for more information.
Teachers	Seeking part-time early childhood teachers for summer or all year. 2-3 mornings per week from 9am-11:15am. Free child care on-site while you work. Free YMCA membership. College degree required in education or related field. Pick up an application at the front desk or call Caryn Shulman, Child Development Coordinator at (847) 272-7250 x232.
Group Exercise Personal Training	Interested individuals with proper certification may contact Myleen Signorini at (847) 272-7250 x 217
Cutomer Service Rep	**OVERQUALIFIED APPLY HERE!** Hone your skills by working in a friendly environment. The front desk is looking for part time staff to work flexible shifts for early weekday mornings, day and evening shifts. Benefits include YMCA membership and babysitting during your shift. Please contact Sarah Tucker or Cheryl Stewart at (847) 272-7250 x 213.
Lifeguards and Swim Instructors	Love to swim? Love kids? Put the two together and make a difference. The North Suburban YMCA is looking for qualified and experienced swim instructors and

Done My Computer

Scoped Learning Extractor: Precision = 58%, Recall = 75% ***Δ Error = -22%***

Broader View

Now touch on some other issues



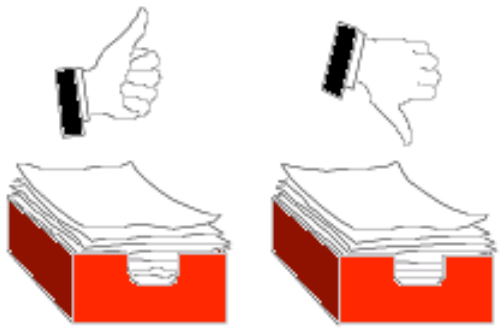
(3) Automatically Inducing an Ontology

[Riloff, '95]

Two inputs:

(1)

preclassified texts



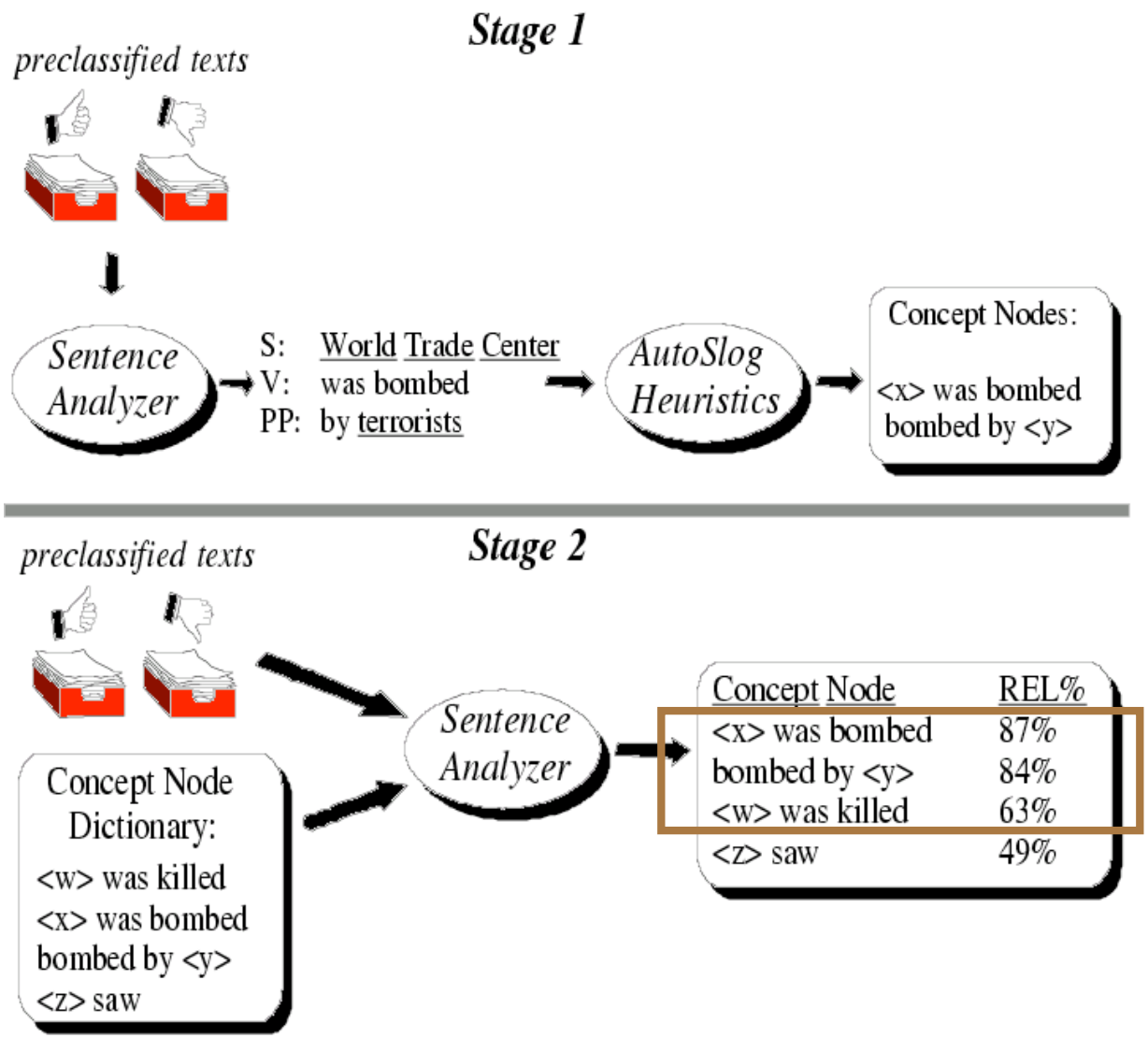
(2)

Heuristic "interesting" meta-patterns.

Linguistic Pattern	Example
1. <subject> active-verb	<perpetrator> bombed
2. <subject> active-verb direct-object[3]	<perpetrator> claimed responsibility
3. <subject> passive-verb	<victim> was murdered
4. <subject> verb infinitive	<perpetrator> attempted to kill
5. <subject> auxiliary noun	<victim> was victim
6. active-verb <direct-object>	bombed <target>
7. passive-verb <direct-object>[4]	killed <victim>
8. infinitive <direct-object>	to kill <victim>
9. verb infinitive <direct-object>	threatened to attack <target>
10. gerund <direct-object>	killing <victim>
11. noun auxiliary <direct-object>	fatality was <victim>
12. noun preposition <noun-phrase>	bomb against <target>
13. active-verb preposition <noun-phrase>	killed with <instrument>
14. passive-verb preposition <noun-phrase>	was aimed at <target>
15. infinitive preposition <noun-phrase>[3]	to fire at <victim>

(3) Automatically Inducing an Ontology

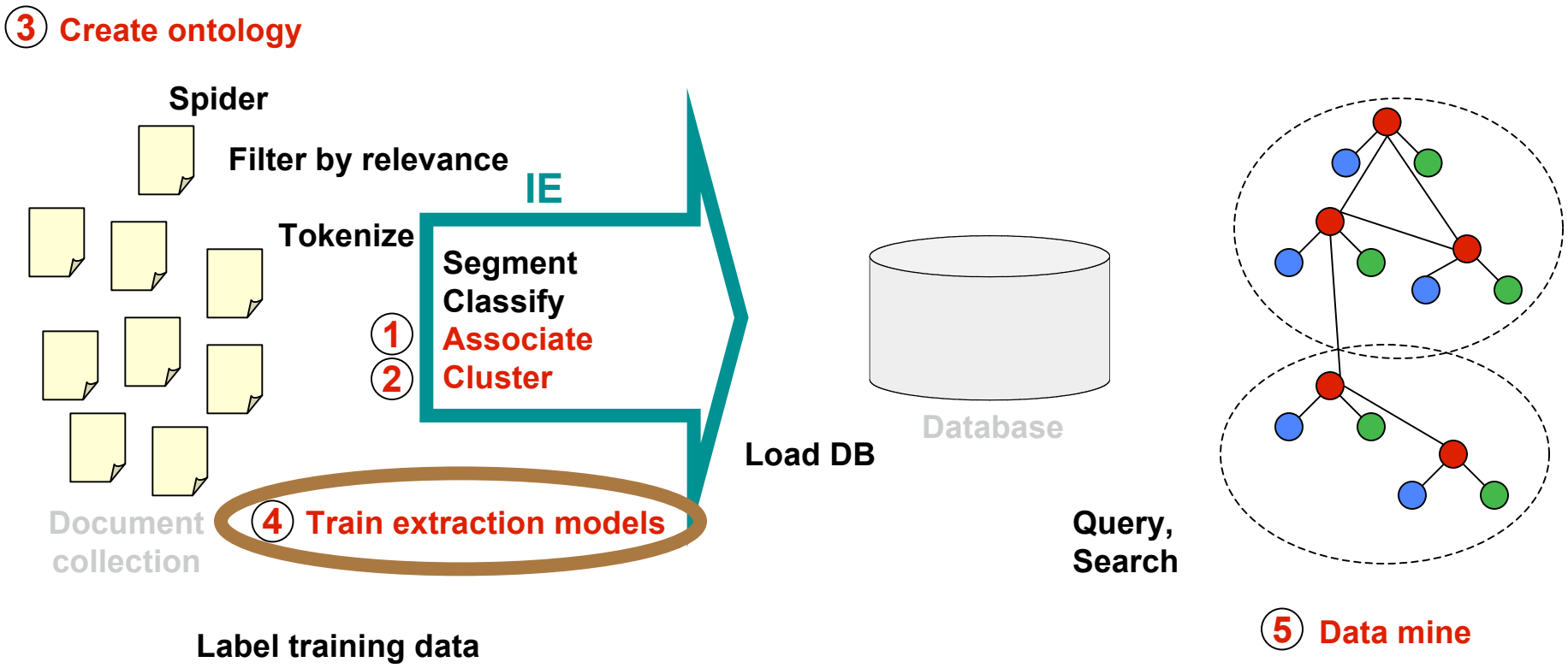
[Riloff, '95]



Subject/Verb/Object patterns that occur more often in the relevant documents than the irrelevant ones.

Broader View

Now touch on some other issues



(4) Training IE Models using Unlabeled Data

[Collins & Singer, 1999]

…says Mr. Cooper, a vice president of …

NNP NNP appositive phrase, head=president

Use two independent sets of features:

Contents: full-string=*Mr._Cooper*, contains(*Mr.*), contains(*Cooper*)
Context: context-type=*appositive*, appositive-head=*president*

1. Start with just seven rules: and ~1M sentences of NYTimes

full-string=New_York	➔ Location
fill-string=California	➔ Location
full-string=U.S.	➔ Location
contains(Mr.)	➔ Person
contains(Incorporated)	➔ Organization
full-string=Microsoft	➔ Organization
full-string=I.B.M.	➔ Organization

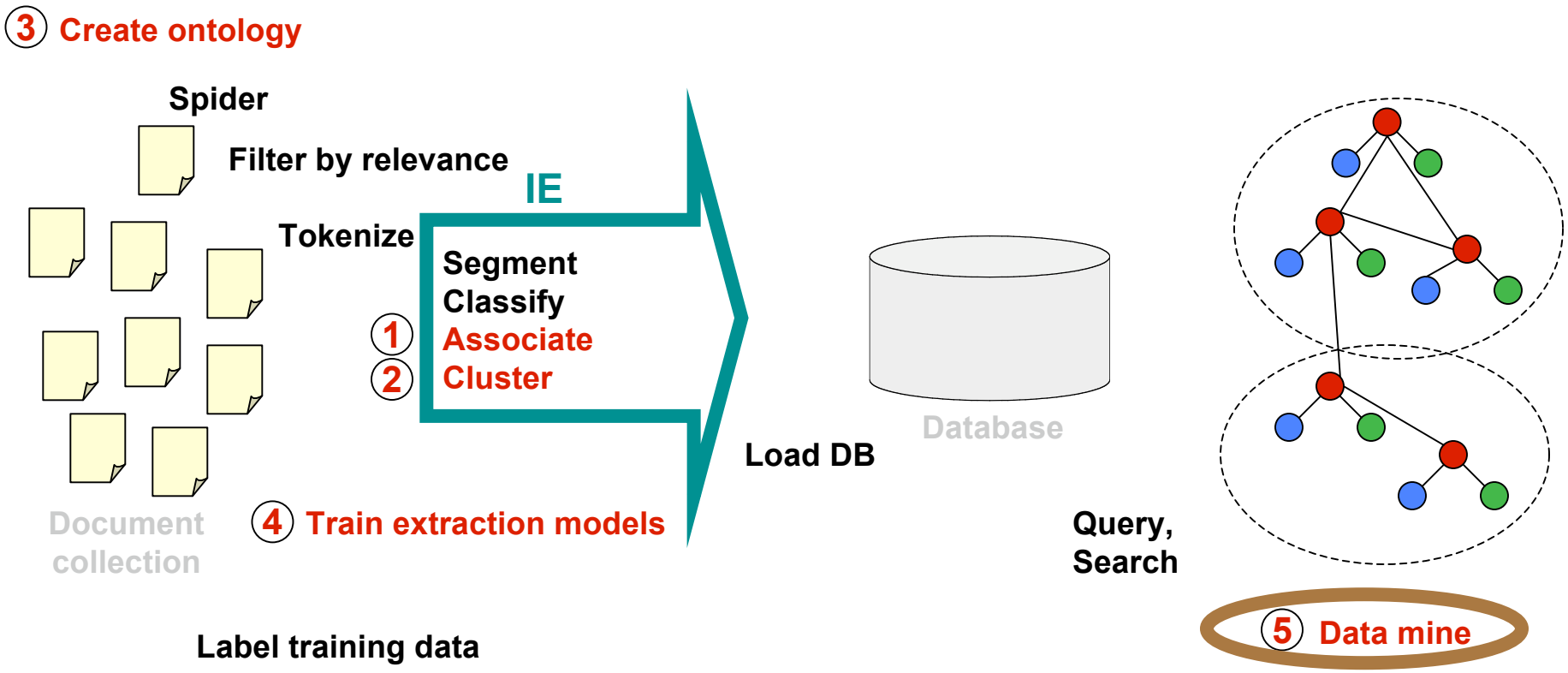
2. Alternately train & label using each feature set.

3. Obtain 83% accuracy at finding *person*, *location*, *organization* & *other* in appositives and prepositional phrases!

See also *[Brin 1998], [Riloff & Jones 1999]*

Broader View

Now touch on some other issues



(5) Data Mining: Working with IE Data

- Some special properties of IE data:
  - It is based on extracted text
  - It is “dirty”, (missing extraneous facts, improperly normalized entity names, etc.
  - May need cleaning before use

- What operations can be done on dirty, unnormalized databases?
  - Query it directly with a language that has “soft joins” across similar, but not identical keys. *[Cohen 1998]*
  - Construct features for learners *[Cohen 2000]*
  - Infer a “best” underlying clean database *[Cohen, Kautz, MacAllester, KDD2000]*

(5) Data Mining: Mutually supportive IE and Data Mining

[Nahm & Mooney, 2000]

Extract a large database
Learn rules to predict the value of each field from the other fields.
Use these rules to increase the accuracy of IE.

Example DB record

Filled Job Template
title: Senior DBMS Consultant
salary: Up to $55K
state: TX
city: Dallas
country: US
language: Powerbuilder, Progress, C, C++, Visual Basic
platform: UNIX, NT
application: SQL Server, Oracle
area: Electronic Commerce, Customer Service
required years of experience: 3
desired years of experience: 5
required degree: BS

Sample Learned Rules

platform:AIX & !application:Sybase & application:DB2
➔application:Lotus Notes

language:C++ & language:C & application:Corba & title=SoftwareEngineer
➔ platform:Windows

language:HTML & platform:WindowsNT & application:ActiveServerPages
➔ area:Database

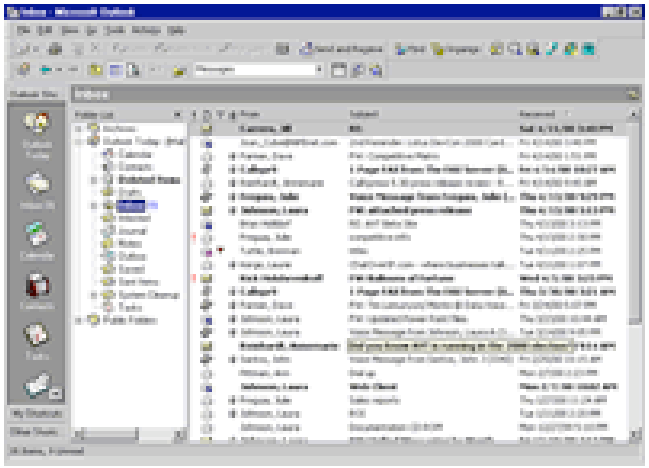
Language:Java & area:ActiveX & area:Graphics
➔ area:Web

Managing and Understanding Connections of People in our Email World

Workplace effectiveness ~ Ability to leverage network of acquaintances

But filling Contacts DB by hand is tedious, and incomplete.

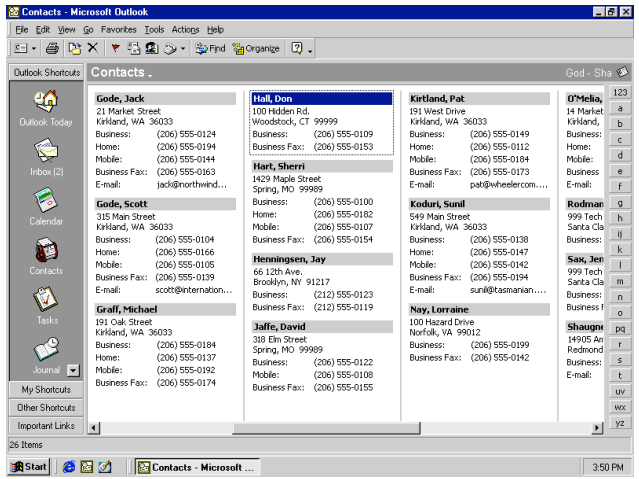
Email Inbox



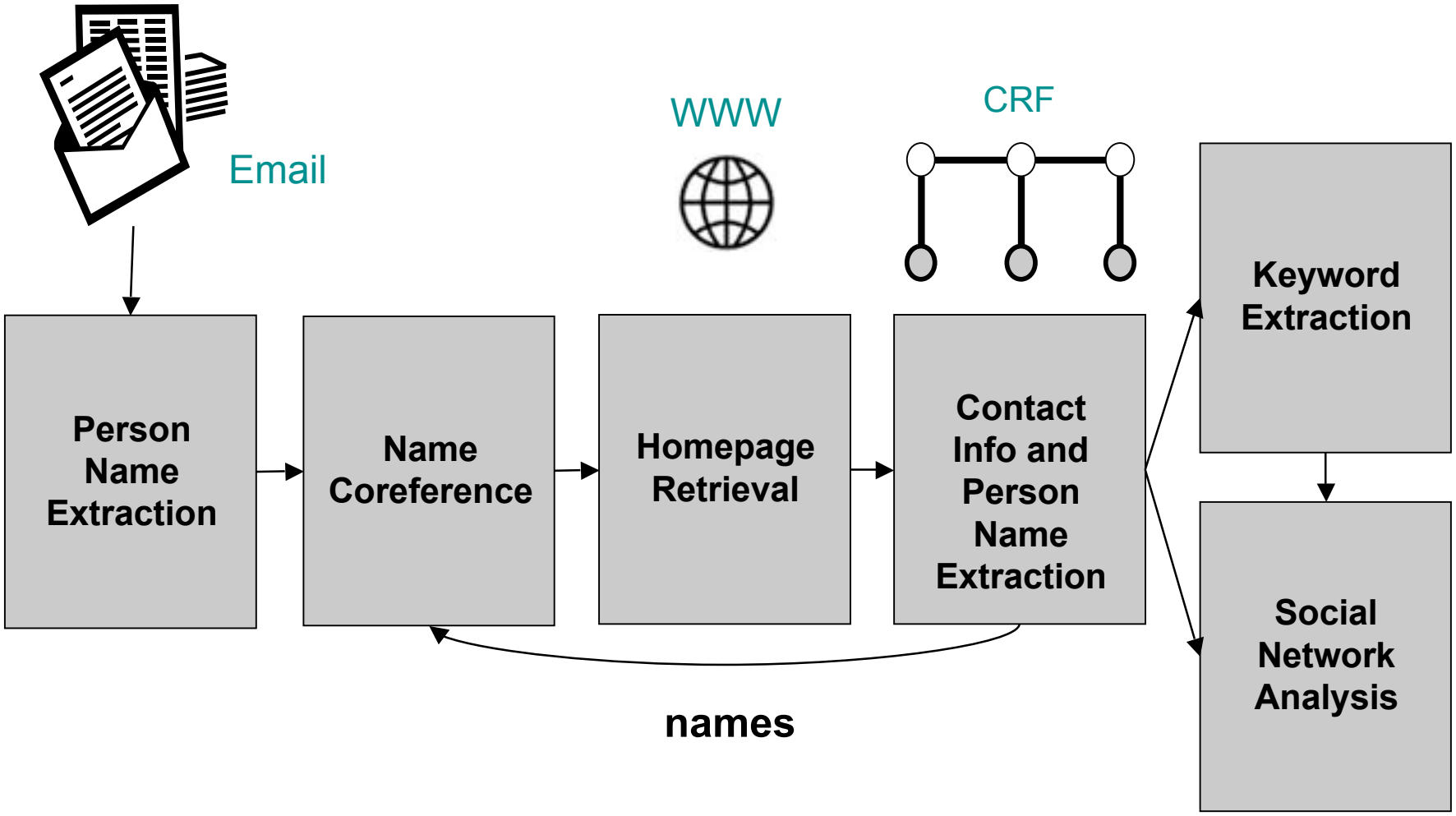
Automatically

WWW

Contacts DB



System Overview



An Example

To: "Andrew McCallum" mccallum@cs.umass.edu

Subject ...

Google Web Images Groups News Froogle New! more »

"andrew mccallum" site:www.cs.umass.edu Search

Web Results 1 - 10 of about 97 from www.cs.umass.edu for "a

Andrew McCallum's Home Page
Andrew McCallum Associate Professor Department of Computer Science University of Massachusetts Amherst 140 Governors Drive Amherst, MA 01003 voice: (413) 545 ...
www.cs.umass.edu/~mccallum/ - 6k - Cached - Similar pages

Search for new people

Andrew McCallum's Home Page
www.cs.umass.edu/~mccallum/
people - research music daily

Andrew McCallum
Associate Professor
Department of Computer Science
University of Massachusetts
140 Governors Drive
Amherst, MA 01003

voice: (413) 545-1323
fax: (413) 545-1789
mccallum@cs.umass.edu

Andrew McCallum's Students and other Collaborators
http://www.cs.umass.edu/~mccallum/collaborators.html
people - research music daily

Students

- Charles Sutton, *(Ph.D. 4th-year)*
- Wei Li, *(Ph.D. 4th-year)*
- Ben Wellner, *(Ph.D. 2nd-year)*
- Aron Culotta, *(Ph.D. 2nd-year)*

The main goal of my research is to dramatically increase our ability to mine actionable knowledge from unstructured text. I am especially interested in **information extraction** from the Web, understanding the connections between people and between organizations, expert finding, **social network analysis**, and mining the scientific literature &

First Name:	Andrew
Middle Name:	Kachites
Last Name:	McCallum
JobTitle:	Associate Professor
Company:	University of Massachusetts
Street Address:	140 Governor's Dr.
City:	Amherst
State:	MA
Zip:	01003
Company Phone:	(413) 545-1323
Links:	Fernando Pereira, Sam Roweis,…
Key Words:	Information extraction, social network,…

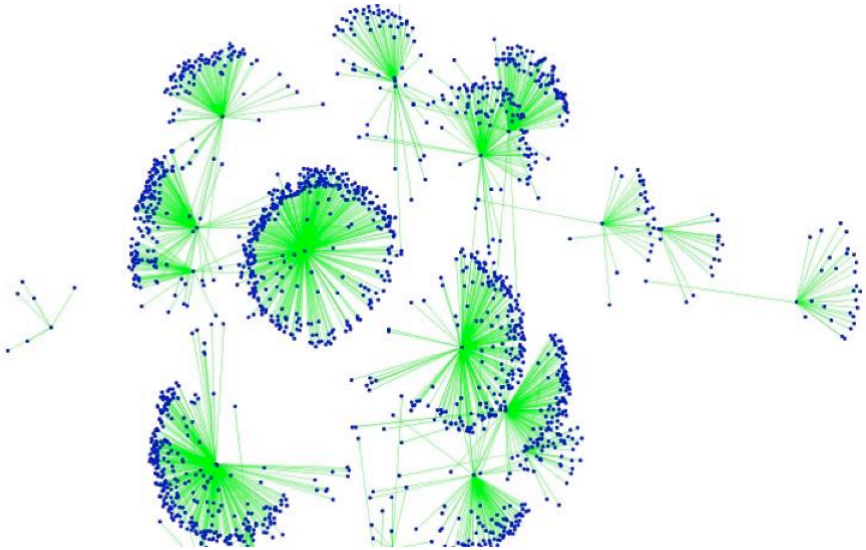
Summary of Results

Example keywords extracted

Person	Keywords
William Cohen	Logic programming Text categorization Data integration Rule learning
Daphne Koller	Bayesian networks Relational models Probabilistic models Hidden variables
Deborah McGuiness	Semantic web Description logics Knowledge representation Ontologies
Tom Mitchell	Machine learning Cognitive states Learning apprentice Artificial intelligence

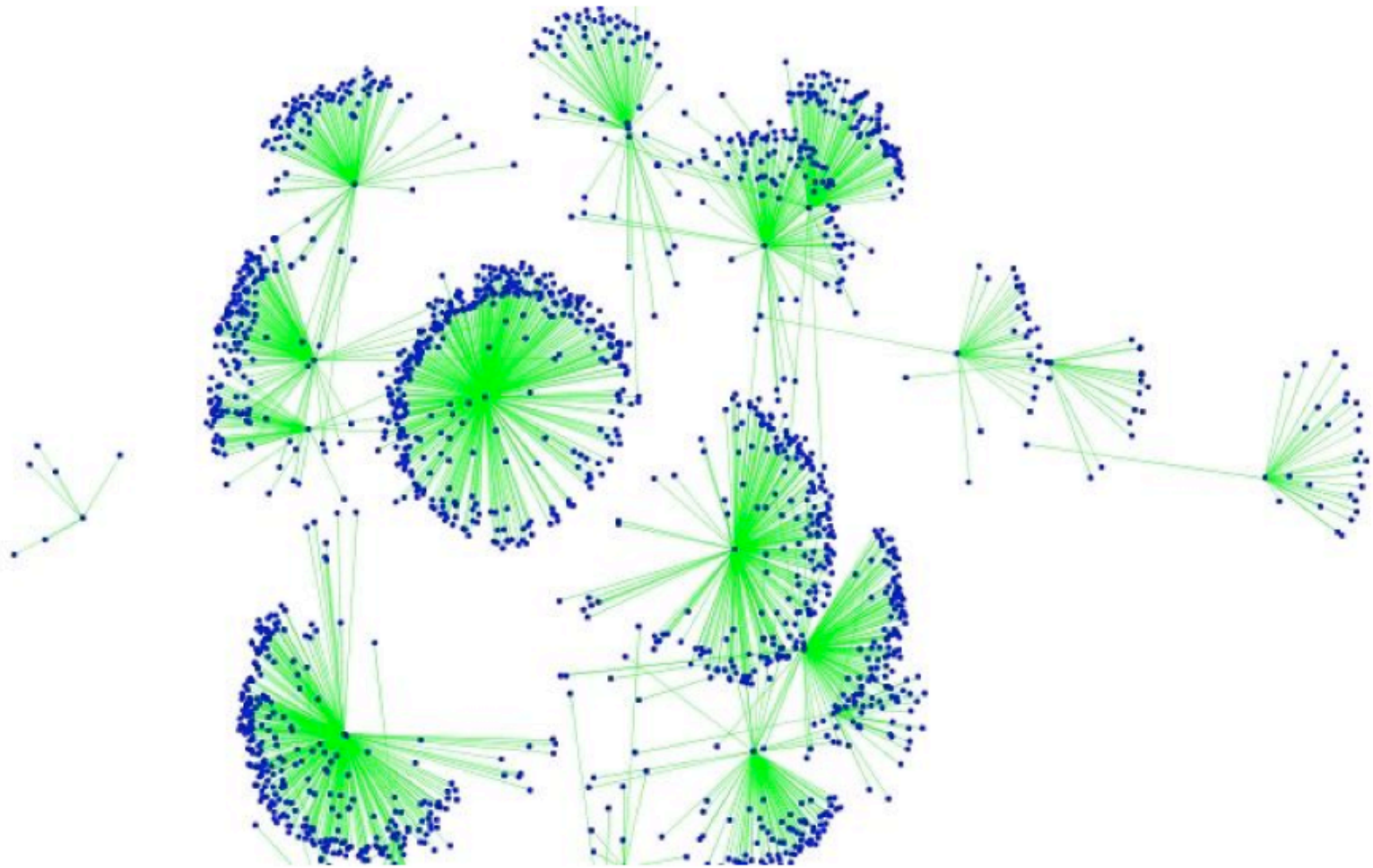
Contact info and name extraction performance (25 fields)

	Token Acc	Field Prec	Field Recall	Field F1
CRF	94.50	85.73	76.33	80.76



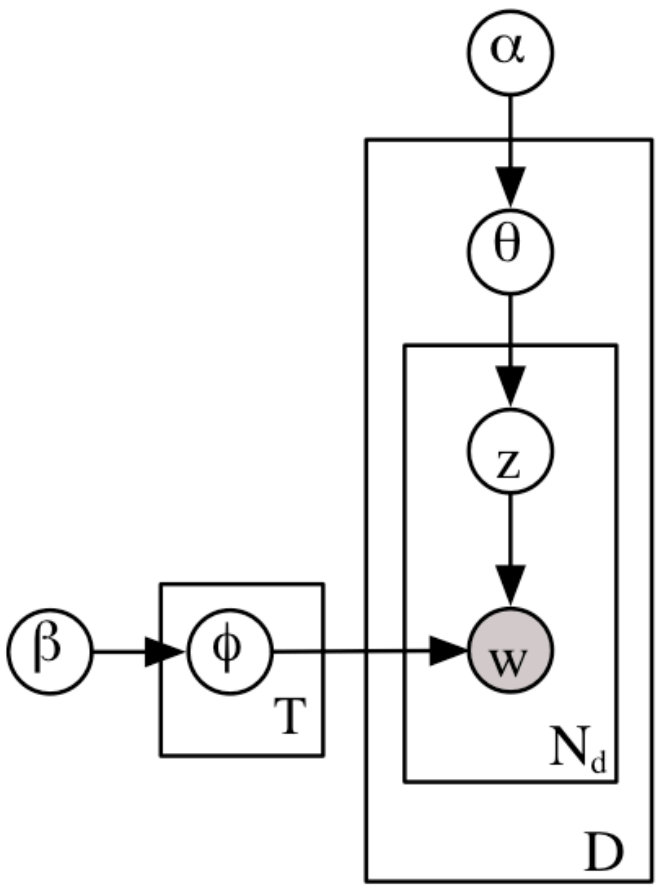
1. ***Expert Finding:***
   When solving some task, find friends-of-friends with relevant expertise.
   Avoid “stove-piping” in large org’s by automatically suggesting collaborators.
   Given a task, automatically suggest the right team for the job. (Hiring aid!)
2. ***Social Network Analysis:***
   Understand the social structure of your organization.
   Suggest structural changes for improved efficiency.

Social Network in an Email Dataset



Clustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]



Generative Process:

Example:

For each document:

Sample a distribution over topics, θ

70% Iraq war
30% US election

For each word in doc

Sample a topic, *z*

Iraq war

Sample a word from the topic, *w*

"bombing"

Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	**FIELD**	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	**FIELD**	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	**FIELD**	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	**FIELD**
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

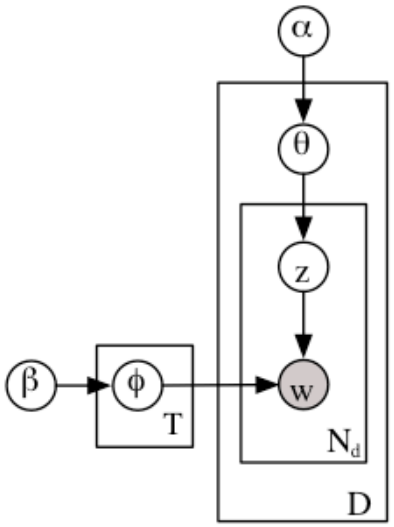
From LDA to Author-Recipient-Topic

(ART)

Latent Dirichlet Allocation

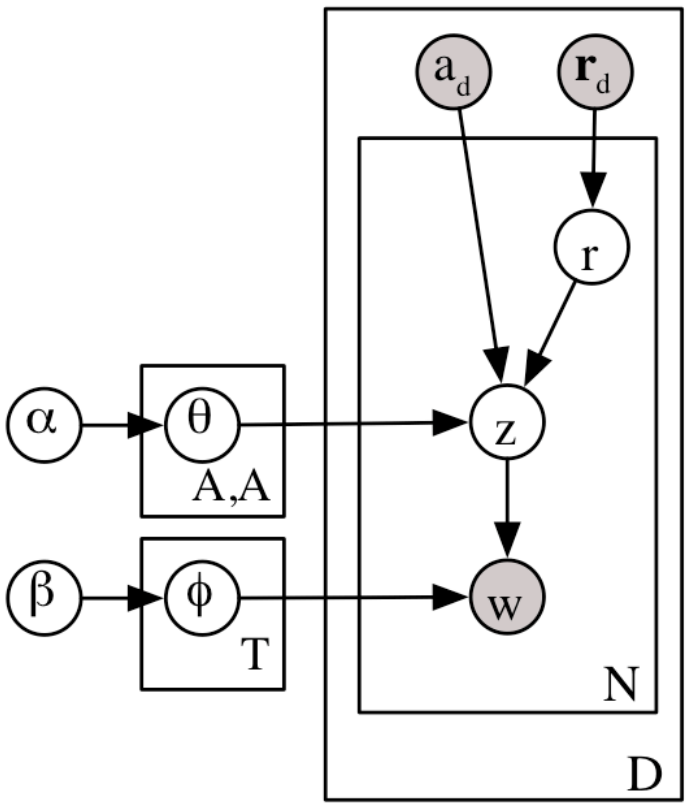
(LDA)

[Blei, Ng, Jordan, 2003]



Inference and Estimation

$$p(\theta, \phi, \mathbf{x}_d, \mathbf{z}_d, \mathbf{w}_d | \alpha, \beta, a_d, \mathbf{r}_d) = p(\theta|\alpha)p(\phi|\beta) \prod_{n=1}^{N_d} p(x_{dn}|\mathbf{r}_d)p(z_{dn}|\theta_{a_d, x_{dn}})p(w_{dn}|\phi_{z_{dn}})$$



Gibbs Sampling:

- Easy to implement
- Reasonably fast

$$P(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \mathbf{w}) \quad \propto \quad \frac{n_{z_i}^{w_v} + \beta_v}{\sum_v n_{z_i}^{w_v} + \beta_v} \frac{n_{x_i}^{z_i} + \alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'} + \alpha_{z'}}$$

$$P(\,\mathrm{r}_i \,|\, \mathbf{z}, \mathbf{x}_{-i}, \mathbf{w}) \quad \propto \quad \frac{n_{x_i}^{z_i} + \alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'} + \alpha_{z'}}$$

Enron Email Corpus

- 250k email messages
- 23k people

```
Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)
From: debra.perlingiere@enron.com
To: steve.hooser@enron.com
Subject: Enron/TransAltaContract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an
electronic copy of our final draft?  Are you OK with this?  If
so, the only version I have is the original draft without
revisions.

DP

Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
dperlin@enron.com
```

Topics, and prominent senders / receivers discovered by ART

Topic names, by hand

Topic 5 "Legal Contracts"		Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
section	0.0299	attached	0.0742	day	0.0419	game	0.0170
party	0.0265	agreement	0.0493	friday	0.0418	draft	0.0156
language	0.0226	review	0.0340	morning	0.0369	week	0.0135
contract	0.0203	questions	0.0257	monday	0.0282	team	0.0135
date	0.0155	draft	0.0245	office	0.0282	eric	0.0130
enron	0.0151	letter	0.0239	wednesday	0.0267	make	0.0125
parties	0.0149	comments	0.0207	tuesday	0.0261	free	0.0107
notice	0.0126	copy	0.0165	time	0.0218	year	0.0106
days	0.0112	revised	0.0161	good	0.0214	pick	0.0097
include	0.0111	document	0.0156	thursday	0.0191	phillip	0.0095
M.Hain J.Steffes	0.0549	G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
J.Dasovich R.Shapiro	0.0377	G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
D.Hyvl K.Ward	0.0362	B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

Topics, and prominent senders / receivers discovered by ART

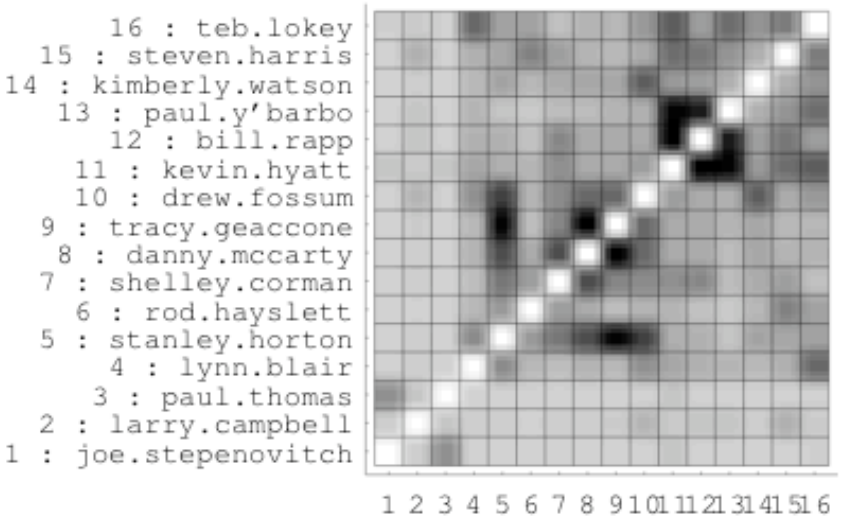
Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338	R.Haylett	0.1432
L.Kitchen		J.Steffes		R.Shapiro		T.Geaccone	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440	T.Geaccone	0.0737
J.Lavorato		R.Shapiro		J.Steffes		R.Haylett	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394	R.Haylett	0.0420
S.White		E.Sager		R.Sanders		D.Fossum	

Beck = "Chief Operations Officer"

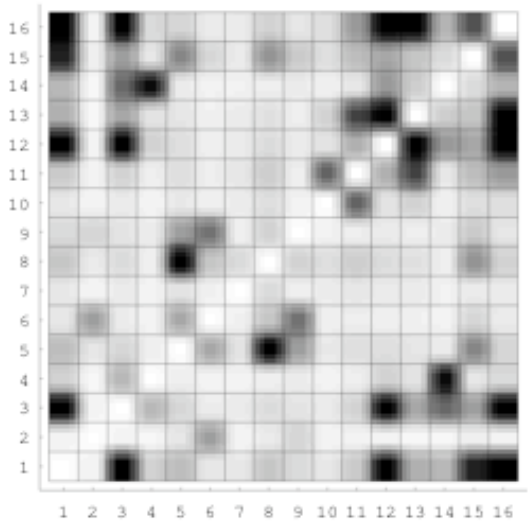
Dasovich = "Government Relations Executive"
Shapiro = "Vice President of Regulatory Affairs"
Steffes = "Vice President of Government Affairs"

Comparing Role Discovery

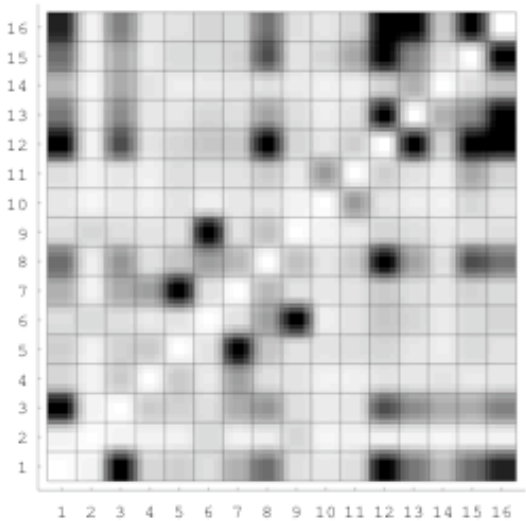
Traditional SNA



ART



Author-Topic



connection strength (A,B) =

distribution over recipients

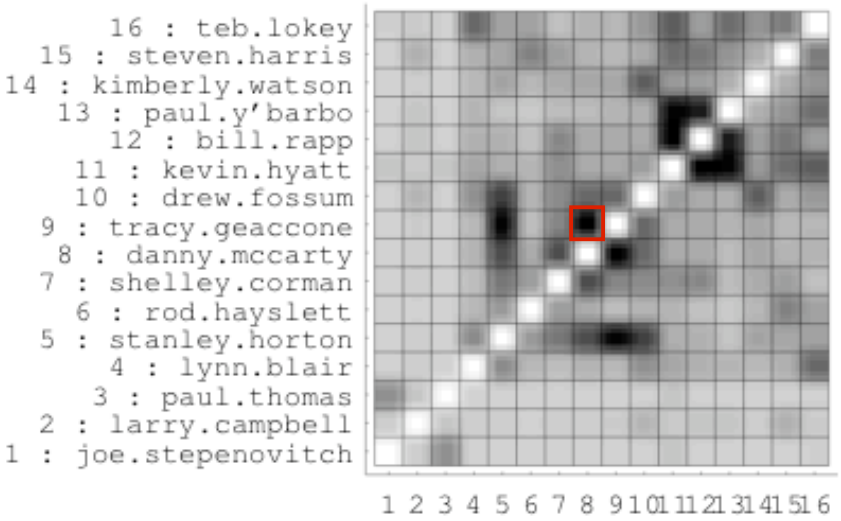
distribution over authored topics

distribution over authored topics

Comparing Role Discovery

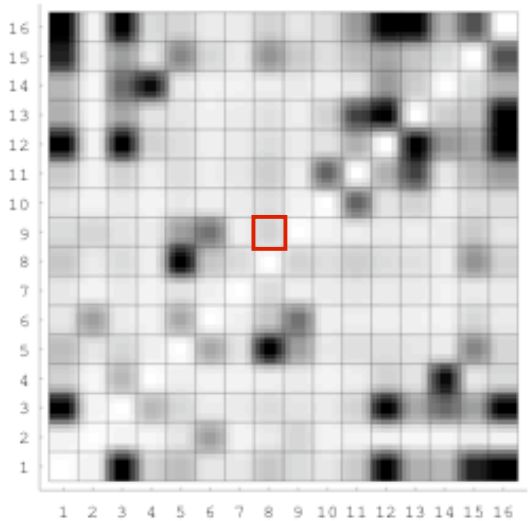
Tracy Geaconne ⇔ Dan McCarty

Traditional SNA



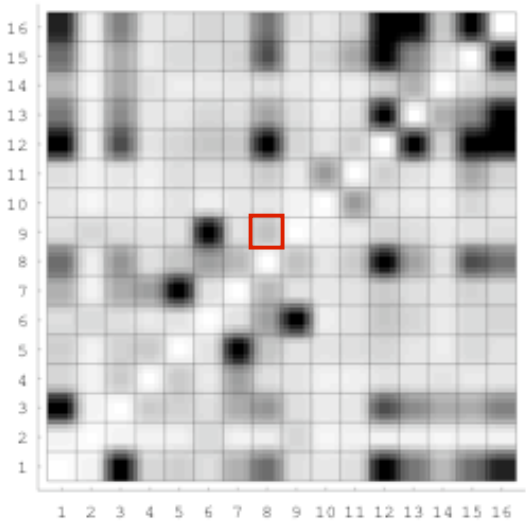
Similar roles

ART



Different roles

Author-Topic



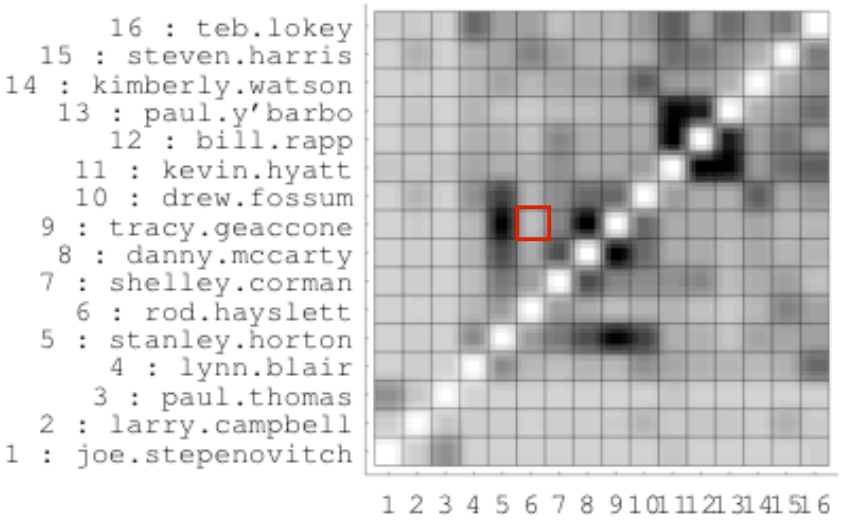
Different roles

Geaconne = "Secretary"
McCarty = "Vice President"

Comparing Role Discovery

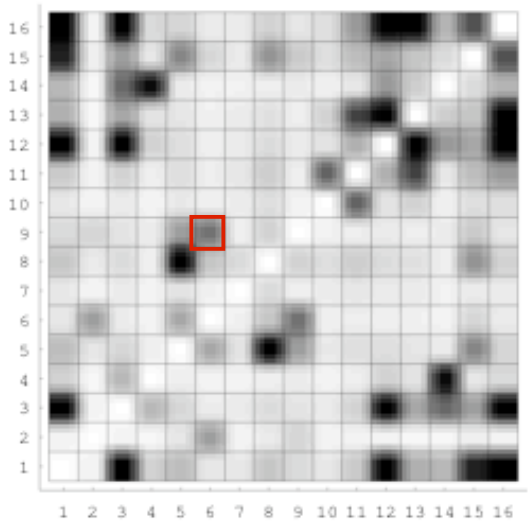
Tracy Geaconne ⇔ Rod Hayslett

Traditional SNA



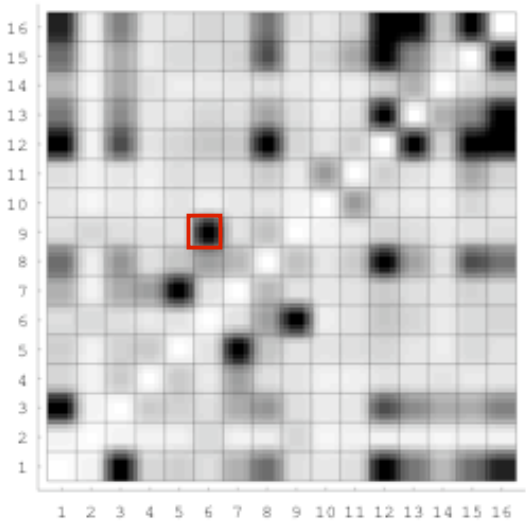
Different roles

ART



Not very similar

Author-Topic



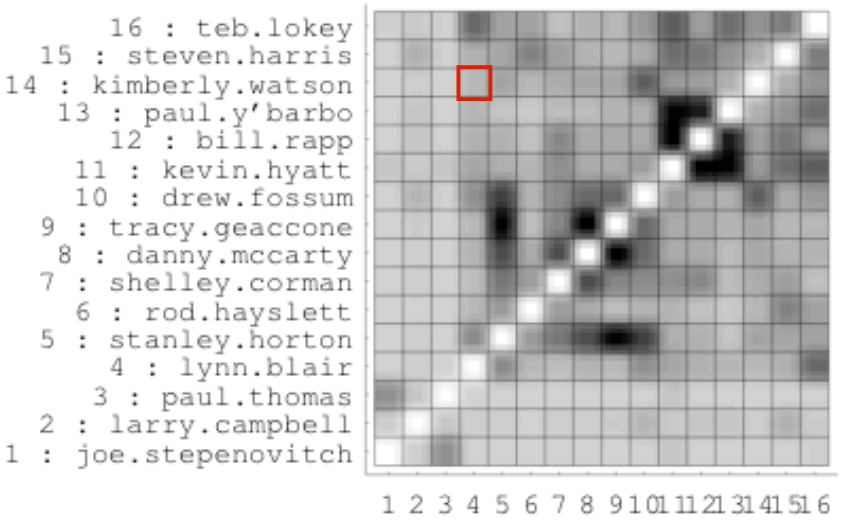
Very similar

Geaconne = "Secretary"
Hayslett = "Vice President & CTO"

Comparing Role Discovery

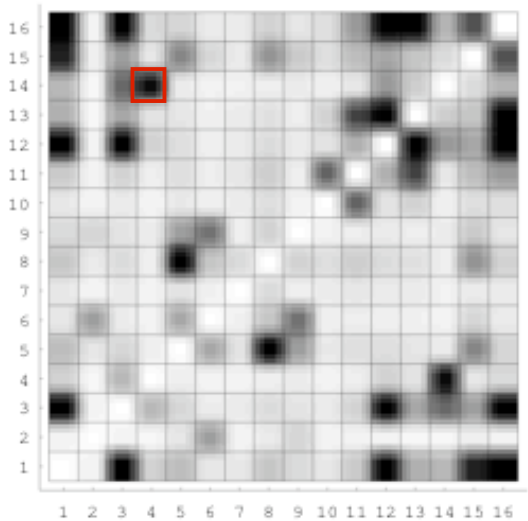
Lynn Blair ⇔ Kimberly Watson

Traditional SNA



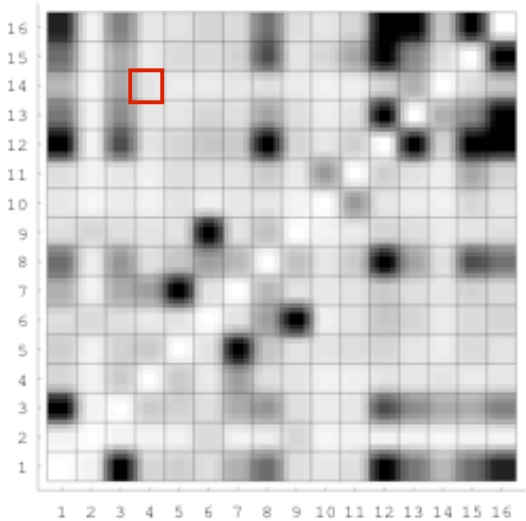
Different roles

ART



Very similar

Author-Topic



Very different

Blair = "Gas pipeline logistics"
Watson = "Pipeline facilities planning"

McCallum Email Corpus 2004

- January - October 2004
- 23k email messages
- 825 people

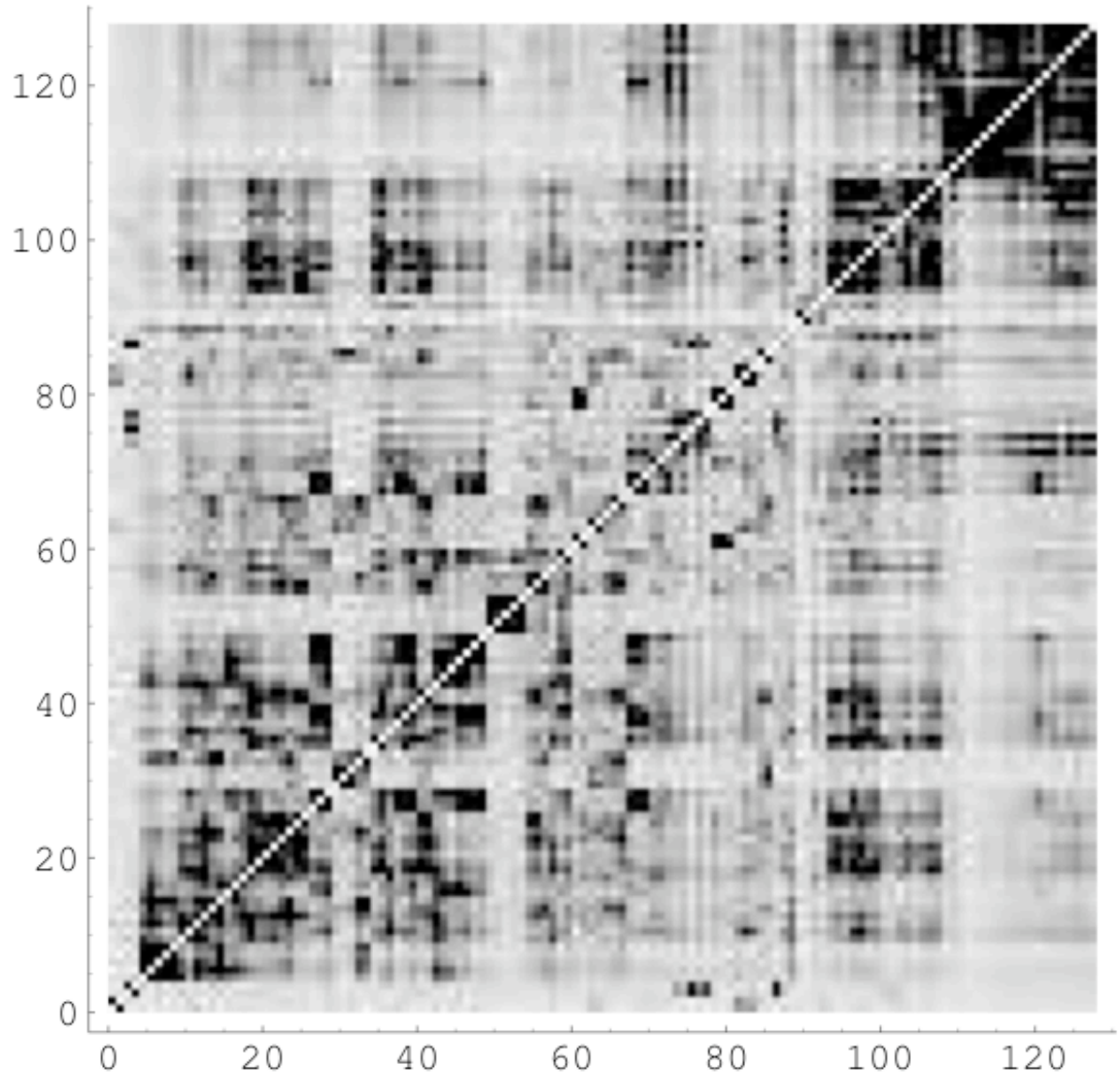
```
From: kate@cs.umass.edu
Subject: NIPS and ....
Date: June 14, 2004 2:27:41 PM EDT
To: mccallum@cs.umass.edu

There is pertinent stuff on the first yellow folder that is
completed either travel or other things, so please sign that
first folder anyway. Then, here is the reminder of the things
I'm still waiting for:

NIPS registration receipt.
CALO registration receipt.

Thanks,
Kate
```

McCallum Email Blockstructure



Four most prominent topics in discussions with ____?

Topic 5 "Grant Proposals"		Topic 31 "Meeting Setup"		Topic 38 "ML Models"		Topic 41 "Friendly Discourse"	
proposal	0.0397	today	0.0512	model	0.0479	great	0.0516
data	0.0310	tomorrow	0.0454	models	0.0444	good	0.0393
budget	0.0289	time	0.0413	inference	0.0191	don	0.0223
work	0.0245	ll	0.0391	conditional	0.0181	sounds	0.0219
year	0.0238	meeting	0.0339	methods	0.0144	work	0.0196
glenn	0.0225	week	0.0255	number	0.0136	wishes	0.0182
nsf	0.0209	talk	0.0246	sequence	0.0126	talk	0.0175
project	0.0188	meet	0.0233	learning	0.0126	interesting	0.0168
sets	0.0157	morning	0.0228	graphical	0.0121	time	0.0162
support	0.0156	monday	0.0208	random	0.0121	hear	0.0132

Topic 5 "Grant Proposals"		Topic 31 "Meeting Setup"		Topic 38 "ML Models"		Topic 41 "Friendly Discourse"	
proposal	0.0397	today	0.0512	model	0.0479	great	0.0516
data	0.0310	tomorrow	0.0454	models	0.0444	good	0.0393
budget	0.0289	time	0.0413	inference	0.0191	don	0.0223
work	0.0245	ll	0.0391	conditional	0.0181	sounds	0.0219
year	0.0238	meeting	0.0339	methods	0.0144	work	0.0196
glenn	0.0225	week	0.0255	number	0.0136	wishes	0.0182
nsf	0.0209	talk	0.0246	sequence	0.0126	talk	0.0175
project	0.0188	meet	0.0233	learning	0.0126	interesting	0.0168
sets	0.0157	morning	0.0228	graphical	0.0121	time	0.0162
support	0.0156	monday	0.0208	random	0.0121	hear	0.0132
smyth mccallum	0.1290	ronb mccallum	0.0339	casutton mccallum	0.0498	mccallum culotta	0.0558
mccallum stowell	0.0746	wellner mccallum	0.0314	icml04-webadmin icml04-chairs	0.0366	mccallum casutton	0.0530
mccallum lafferty	0.0739	casutton mccallum	0.0217	mccallum casutton	0.0343	mccallum ronb	0.0274
mccallum smyth	0.0532	mccallum casutton	0.0200	nips04workflow mccallum	0.0322	mccallum saunders	0.0255
pereira lafferty	0.0339	mccallum wellner	0.0200	weinman mccallum	0.0250	mccallum pereira	0.0181

Two most prominent topics in discussions with ____?

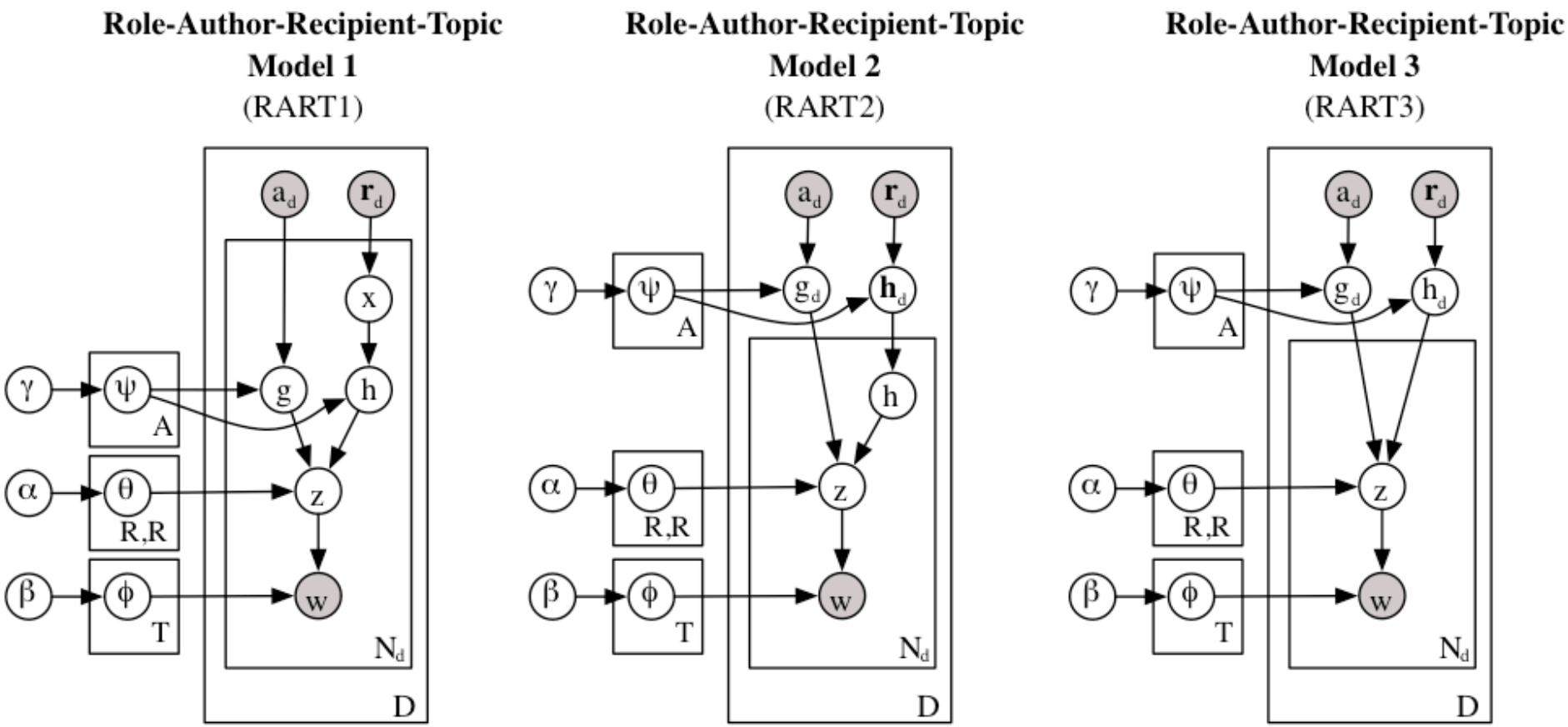
Words	Prob
love	0.030514
house	0.015402
	0.013659
time	0.012351
great	0.011334
hope	0.011043
dinner	0.00959
saturday	0.009154
left	0.009154
ll	0.009009
	0.008282
visit	0.008137
evening	0.008137
stay	0.007847
bring	0.007701
weekend	0.007411
road	0.00712
sunday	0.006829
kids	0.006539
flight	0.006539

Words	Prob
today	0.051152
tomorrow	0.045393
time	0.041289
ll	0.039145
meeting	0.033877
week	0.025484
talk	0.024626
meet	0.023279
morning	0.022789
monday	0.020767
back	0.019358
call	0.016418
free	0.015621
home	0.013967
won	0.013783
day	0.01311
hope	0.012987
leave	0.012987
office	0.012742
tuesday	0.012558

Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coref error)
aepshtey smucker	Both students in McCallum's class
coe laurie	Both UMass admin assistants
mcollins tom.mitchell	Both ML researchers on SRI project
mcollins gervasio	Both ML researchers on SRI project
davitz freeman	Both ML researchers on SRI project
mahadeva pal	Both ML researchers, discussing hiring
kate laurie	Both UMass admin assistants
ang joshuago	Both on org committee for a conference

Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	Spouse is unrelated to visiting researcher
donna yugu	Spouse is unrelated grad student
aepshtey smucker	Both students in McCallum's class
rasmith smucker	Both students in McCallum's class
editor elm	Journal editor and its Production Editor

Role-Author-Recipient-Topic Models



Results with RART: People in “Role #3” in Academic Email

- **`olc`** lead Linux sysadmin
- **`gauthier`** sysadmin for CIIR group
- **`irsystem`** mailing list CIIR sysadmins
- **`system`** mailing list for dept. sysadmins
- **`allan`** Prof., chair of “computing committee”
- **`valerie`** second Linux sysadmin
- **`tech`** mailing list for dept. hardware
- **`steve`** head of dept. I.T. support

Roles for `allan` (James Allan)

- Role #3 I.T. support
- Role #2 Natural Language researcher

Roles for `pereira` (Fernando Pereira)

- Role #2 Natural Language researcher
- Role #4 SRI CALO project participant
- Role #6 Grant proposal writer
- Role #10 Grant proposal coordinator
- Role #8 Guests at McCallum's house