Information Extraction

Introduction to Natural Language Processing

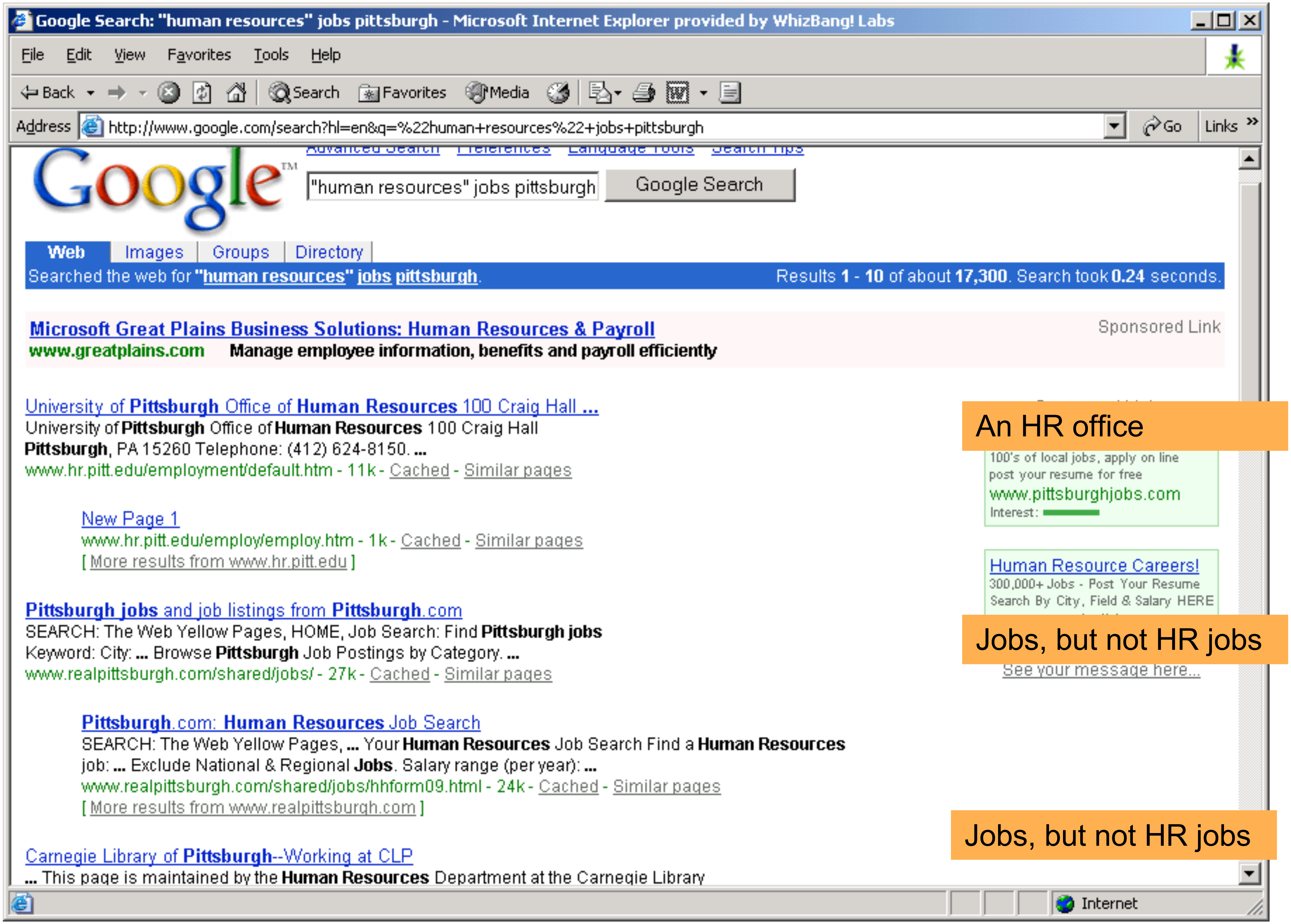
CMPSCI 585, Fall 2007

University of Massachusetts Amherst

Andrew McCallum

Goal:

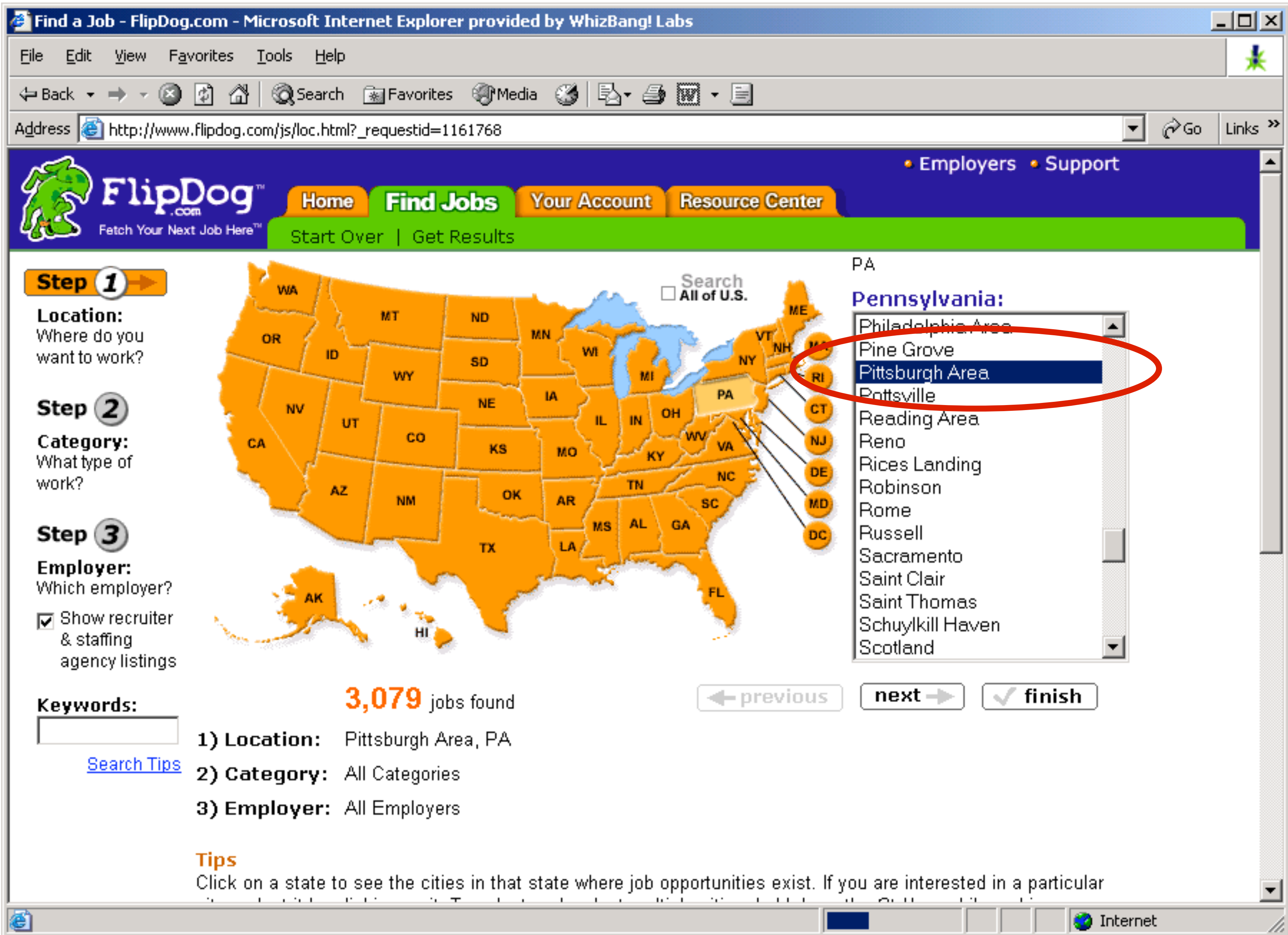
Mine actionable knowledge from unstructured text.

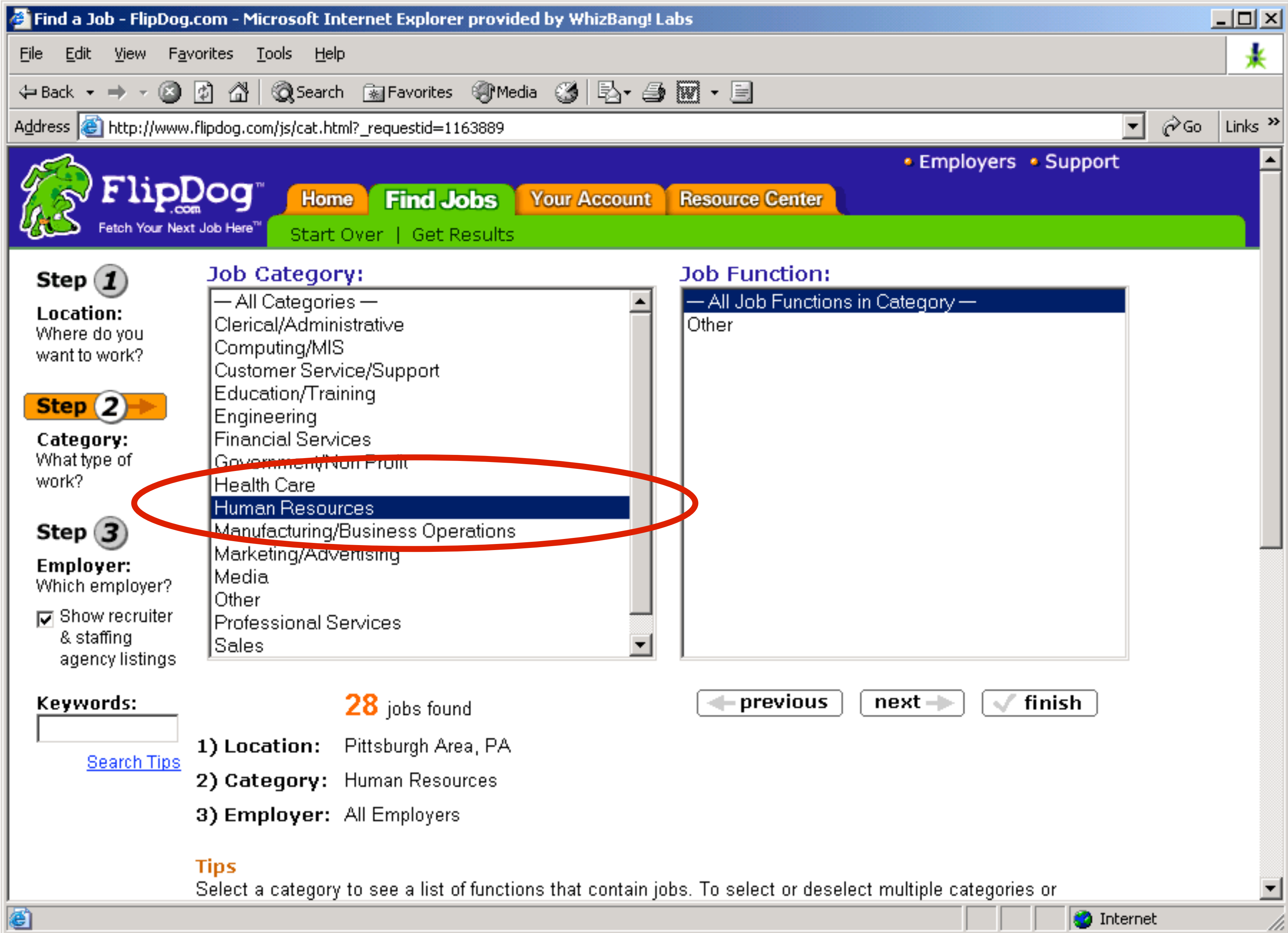


Google Search: "human resources" jobs pittsburgh - Microsoft Internet Explorer provided by WhizBang! Labs
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.google.com/search?hl=en&q=%22human+resources%22+jobs+pittsburgh
Go Links
"human resources" jobs pittsburgh
Google Search
Web Images Groups Directory
Searched the web for "human resources" jobs pittsburgh.
Results 1 - 10 of about 17,300. Search took 0.24 seconds.
Microsoft Great Plains Business Solutions: Human Resources & Payroll
www.greatplains.com Manage employee information, benefits and payroll efficiently
Sponsored Link
University of Pittsburgh Office of Human Resources 100 Craig Hall ...
University of Pittsburgh Office of Human Resources 100 Craig Hall
Pittsburgh, PA 15260 Telephone: (412) 624-8150. ...
www.hr.pitt.edu/employment/default.htm - 11k - Cached - Similar pages
New Page 1
www.hr.pitt.edu/employ/employ.htm - 1k - Cached - Similar pages
[More results from www.hr.pitt.edu]
Pittsburgh jobs and job listings from Pittsburgh.com
SEARCH: The Web Yellow Pages, HOME, Job Search: Find Pittsburgh jobs
Keyword: City: ... Browse Pittsburgh Job Postings by Category. ...
www.realpittsburgh.com/shared/jobs/ - 27k - Cached - Similar pages
Pittsburgh.com: Human Resources Job Search
SEARCH: The Web Yellow Pages, ... Your Human Resources Job Search Find a Human Resources
job: ... Exclude National & Regional Jobs. Salary range (per year): ...
www.realpittsburgh.com/shared/jobs/hhform09.html - 24k - Cached - Similar pages
[More results from www.realpittsburgh.com]
Carnegie Library of Pittsburgh--Working at CLP
... This page is maintained by the Human Resources Department at the Carnegie Library
An HR office
100's of local jobs, apply on line
post your resume for free
www.pittsburghjobs.com
Interest:
Human Resource Careers!
300,000+ Jobs - Post Your Resume
Search By City, Field & Salary HERE
Jobs, but not HR jobs
See your message here...
Jobs, but not HR jobs
Internet

Example: A Solution

Extracting Job Openings from the Web


Find a Job - FlipDog.com - Microsoft Internet Explorer provided by WhizBang! Labs
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.flipdog.com/js/loc.html?_requestid=1161768
Go Links
Employers Support
FlipDog.com
Fetch Your Next Job Here
Home Find Jobs Your Account Resource Center
Start Over | Get Results
Step 1
Location: Where do you want to work?
Step 2
Category: What type of work?
Step 3
Employer: Which employer?
Show recruiter & staffing agency listings
Keywords:
Search Tips
Search All of U.S.
WA OR CA ID NV MT WY UT AZ CO NM ND SD NE KS OK TX MN IA MO AR LA WI IL MI IN OH KY TN MS AL GA FL SC NC VA WV PA NY VT NH ME MA RI CT NJ DE MD DC AK HI
PA
Pennsylvania:
Philadelphia Area
Pine Grove
Pittsburgh Area
Pottsville
Reading Area
Reno
Rices Landing
Robinson
Rome
Russell
Sacramento
Saint Clair
Saint Thomas
Schuylkill Haven
Scotland
previous next finish
3,079 jobs found
1) Location: Pittsburgh Area, PA
2) Category: All Categories
3) Employer: All Employers
Tips
Click on a state to see the cities in that state where job opportunities exist. If you are interested in a particular
Internet


Find a Job - FlipDog.com - Microsoft Internet Explorer provided by WhizBang! Labs
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.flipdog.com/js/cat.html?_requestid=1163889
Go Links
Employers Support
FlipDog.com
Fetch Your Next Job Here™
Home Find Jobs Your Account Resource Center
Start Over | Get Results
Step 1
Location:
Where do you want to work?
Step 2
Category:
What type of work?
Step 3
Employer:
Which employer?
Show recruiter & staffing agency listings
Keywords:
Search Tips
Job Category:
— All Categories —
Clerical/Administrative
Computing/MIS
Customer Service/Support
Education/Training
Engineering
Financial Services
Government/Non Profit
Health Care
Human Resources
Manufacturing/Business Operations
Marketing/Advertising
Media
Other
Professional Services
Sales
Job Function:
— All Job Functions in Category —
Other
28 jobs found
previous next finish
1) Location: Pittsburgh Area, PA
2) Category: Human Resources
3) Employer: All Employers
Tips
Select a category to see a list of functions that contain jobs. To select or deselect multiple categories or
Internet


Job Search Results - FlipDog.com - Microsoft Internet Explorer provided by WhizBang! Labs
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.flipdog.com/js/jobsearch-results.html?loc=PA-Pittsburgh+Area&cat=Human+Resources&job=1
Go Links
Employers Support
FlipDog.com
Fetch Your Next Job Here
Home Find Jobs Your Account Resource Center
Return to Results | Modify Search | New Search
FREE! Professional Resume Plus
Great looking professional resumes with the click of a button.
Kennedy-Western University FREE CATALOG CLICK HERE
Apply your previous education and work experience towards your degree. FREE Catalog!
Go to FlipDog.com
1 - 25 of 28 jobs shown below
1 2 Next
Search within results for:
GO! Search tips
Premium Postings
What are Premium Postings?
Partner Consultant at Profiles International.com
Independent Consultant Independent consultant calling on business and industry marketing 21st century Assessment Instruments to help with hiring, developing and managing of human capital needs. We are a business service operating in 38+ countries. This
April 17, 2002
Pittsburgh, PA
Human Resources
Other
Web Directory
What is Web Directory?
Coordinating Interviewer at University of Pittsburgh April 25, 2002 Pittsburgh, PA
Director of Employee Relations at Macromedia April 25, 2002 Pittsburgh, PA
COMPENSATION & BENEFITS MANAGER at Chelsea Building Products, Inc. April 25, 2002 Oakmont, PA
National Fleet Safety Manager at Western Pennsylvania Chapter, American Society of Safety Engineers April 25, 2002 Greensburg, PA
Drafters at Oxford Technology April 25, 2002 Pittsburgh, PA
Career Services Student Counselor at University of Pittsburgh April 25, 2002 Pittsburgh, PA
Internet

Data Mining the Extracted Job Information

IE from Research Papers

[McCallum et al '99]

Reinforcement Learning: A Survey

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

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

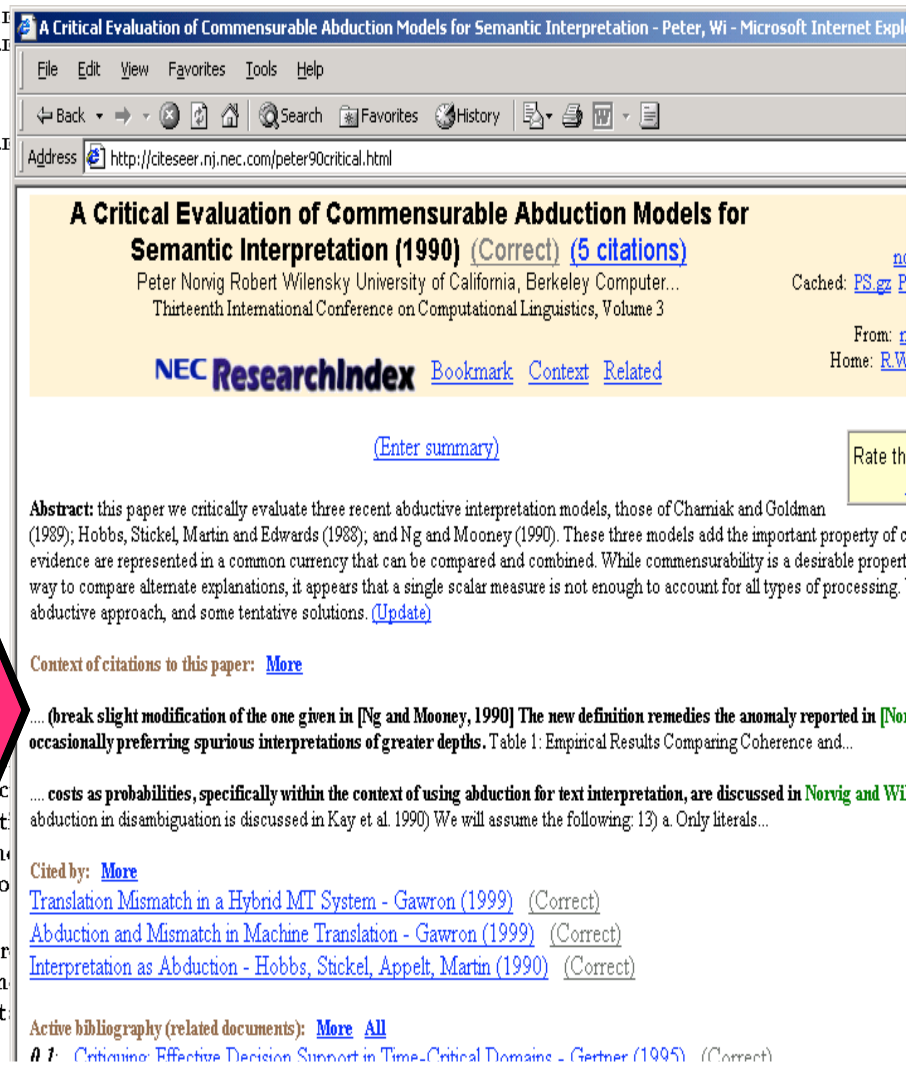
Abstract

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 stat psychology, neuroscience, and computer science. In the last five to ten years, it has attrac rapidly increasing interest in the machine learning and artificial intelligence communit Its promise is beguiling—a way of programming agents by reward and punishment with needing to specify *how* the task is to be achieved. But there are formidable computatio obstacles to fulfilling the promise.

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



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

What is “Information Extraction”

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

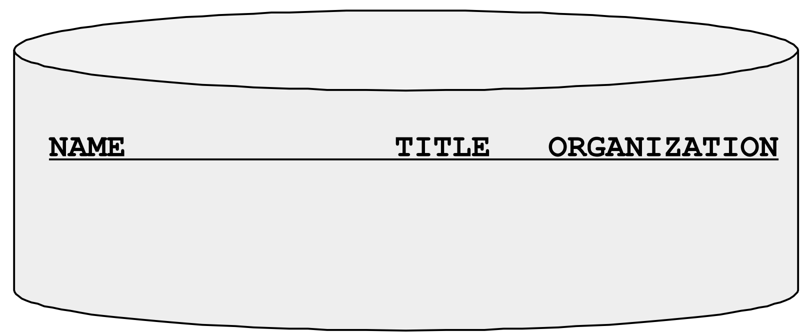
October 14, 2002, 4:00 a.m. PT

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.**

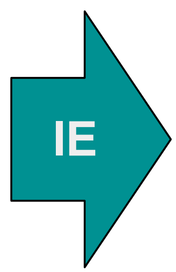
October 14, 2002, 4:00 a.m. PT

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…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

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…

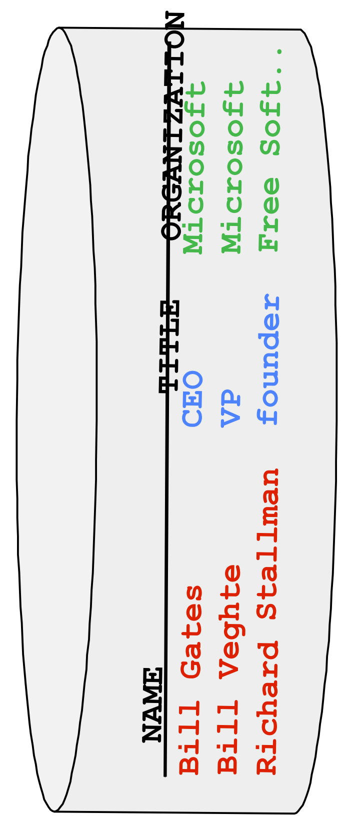
* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

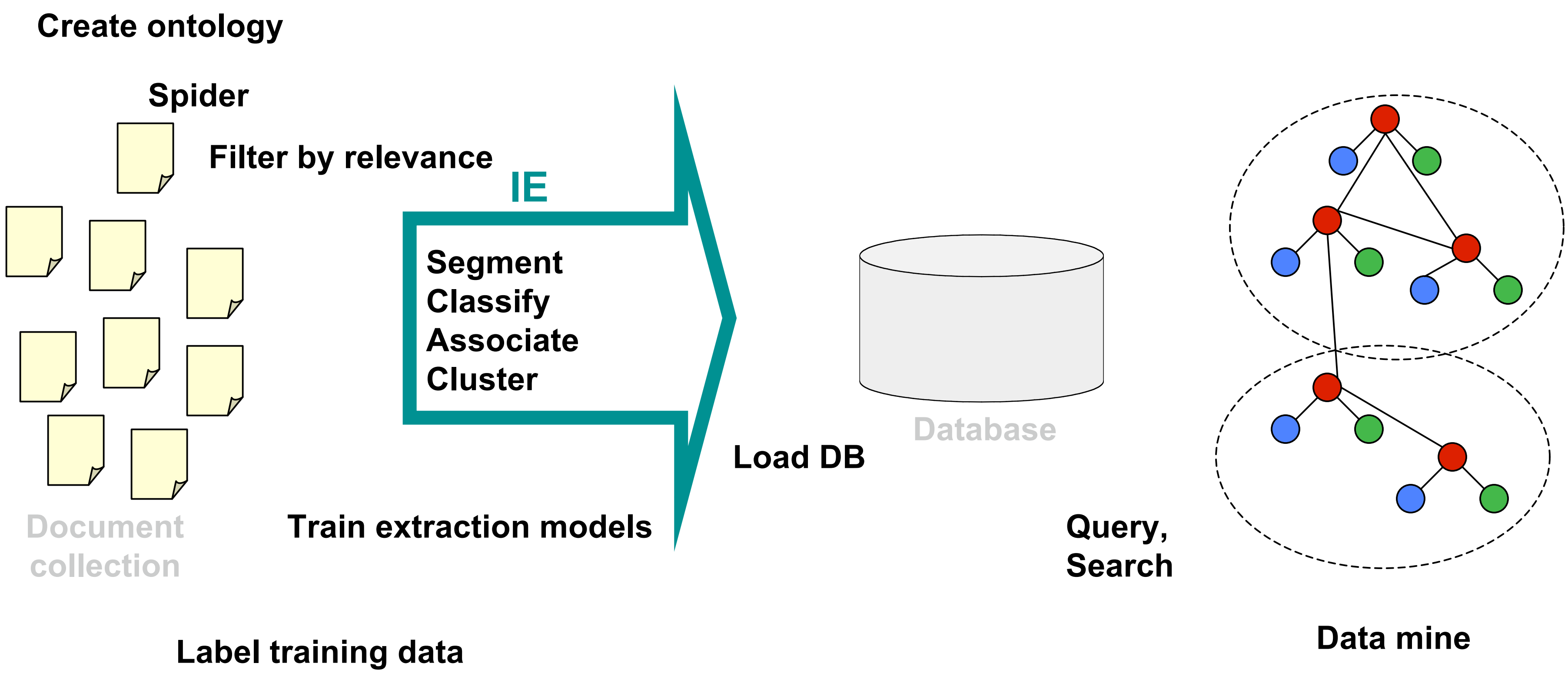
* Microsoft

Bill Veghte
* Microsoft
VP

Richard Stallman
founder
Free Software Foundation



IE in Context



Why Information Extraction (IE)?

- Science
  - Grand old dream of AI: Build large KB* and reason with it.
    IE enables the automatic creation of this KB.
  - IE is a complex problem that inspires new advances in machine learning.
- Profit
  - Many companies interested in leveraging data currently "locked in unstructured text on the Web".
  - Not yet a monopolistic winner in this space.
- Fun!
  - Build tools that we researchers like to use ourselves:
    Cora & CiteSeer, MRQE.com, FAQFinder,…
  - See our work get used by the general public.

* KB = "Knowledge Base"

Outline

- Examples of IE and Data Mining
- $\boxed{\text{Landscape of problems and solutions}}$
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

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

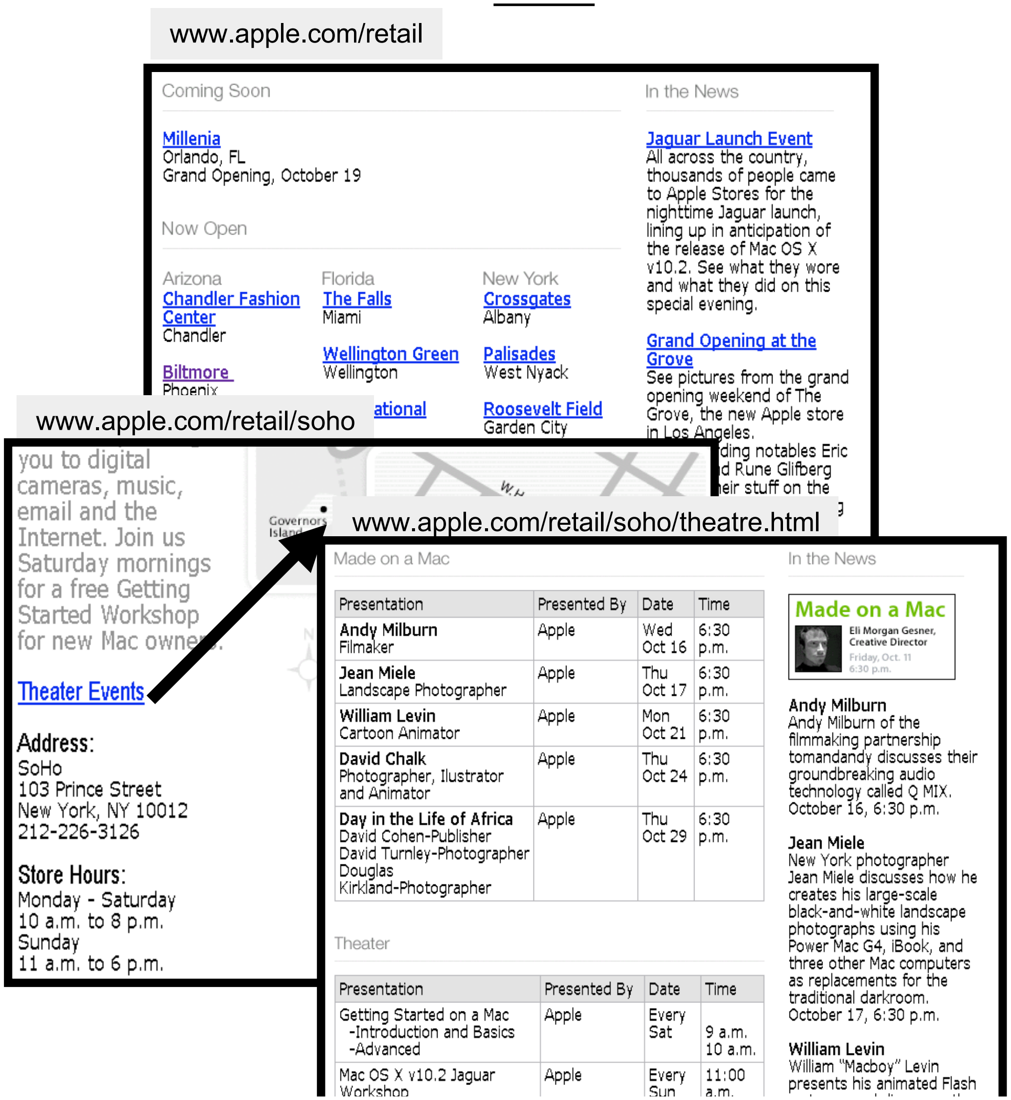
Newswire

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.

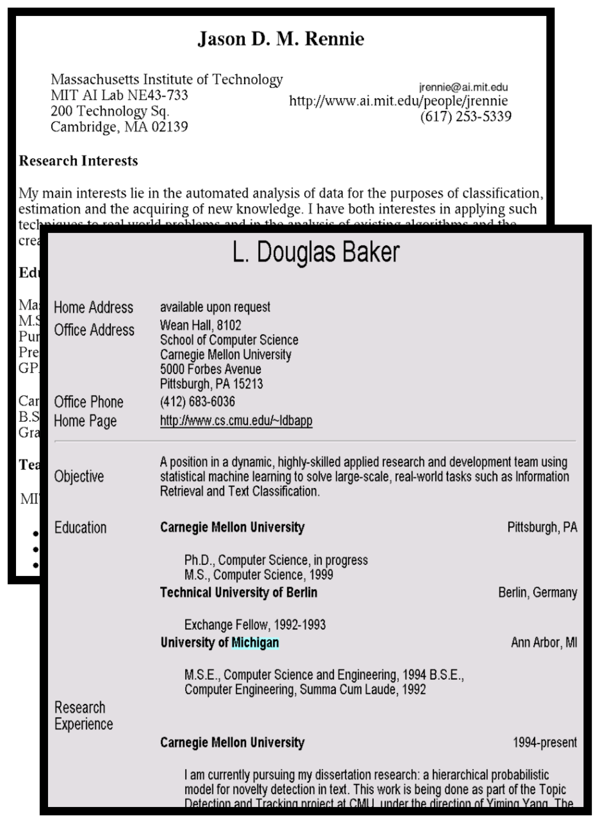
Cohen, Paul R. (413) 545-3638 cohen@cs.umass.edu CS278
Professor.
Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

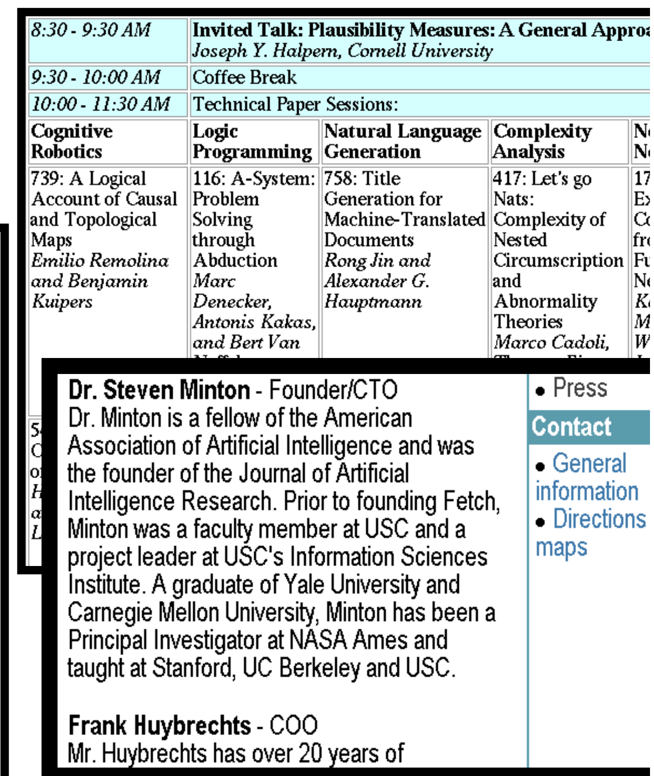
Tables

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	Genre specific	Wide, non-specific
Formatting	Layout	Language
Amazon.com Book Pages	Resumes	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:

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

PRED:

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

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

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

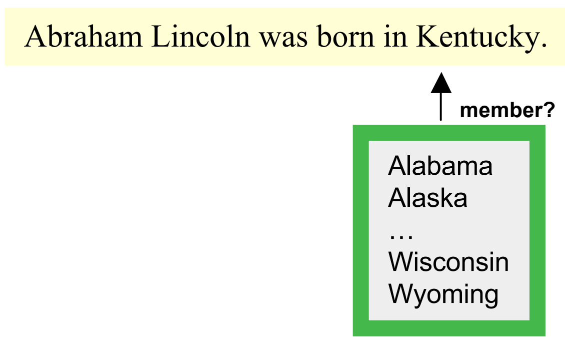
$$\text{F1} = \text{Harmonic mean of Precision \& Recall} = \frac{1}{((1/P) + (1/R)) / 2}$$

State of the Art Performance

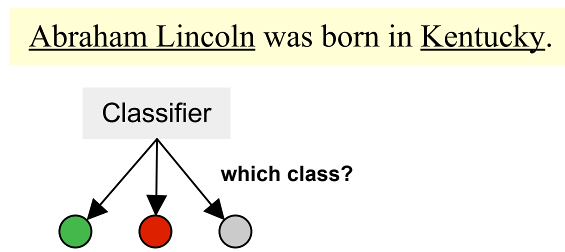
- Named entity recognition
  - Person, Location, Organization, …
  - F1 in high 80's or low- to mid-90's
- Binary relation extraction
  - Contained-in (Location1, Location2)
    Member-of (Person1, Organization1)
  - F1 in 60's or 70's or 80's
- Wrapper induction
  - Extremely accurate performance obtainable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

Landscape of IE Techniques (1/1): Models

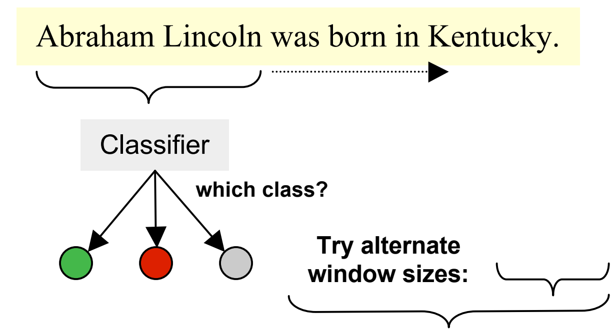
Lexicons



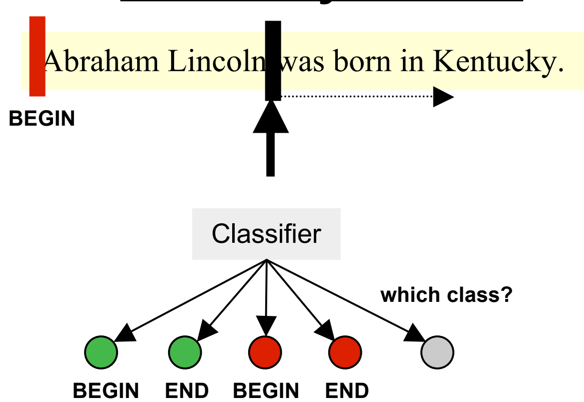
Classify Pre-segmented Candidates



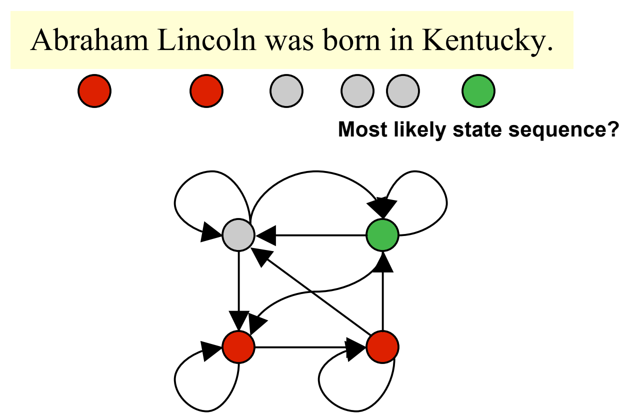
Sliding Window



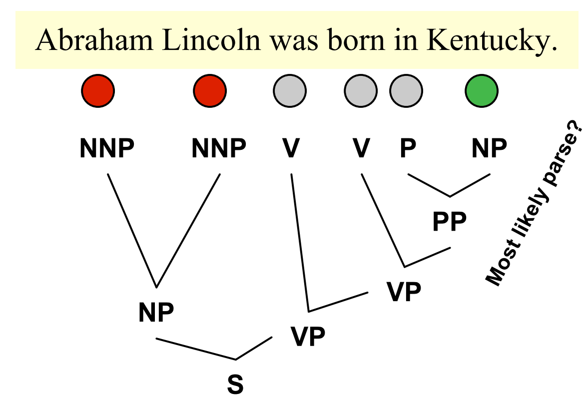
Boundary Models



Finite State Machines



Context Free Grammars



…and beyond

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

Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

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

"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.

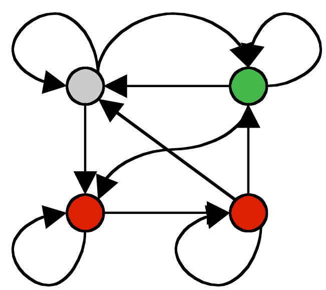
Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

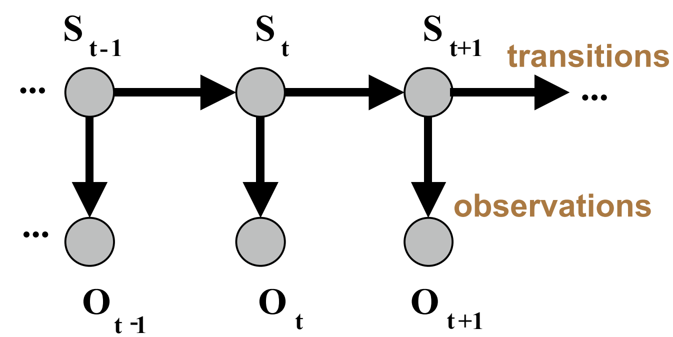
Hidden Markov Models

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

Finite state model



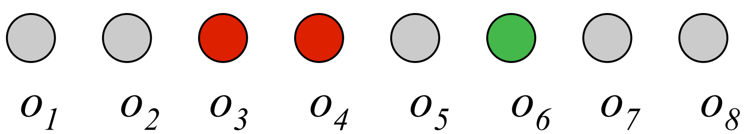
Graphical model



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:

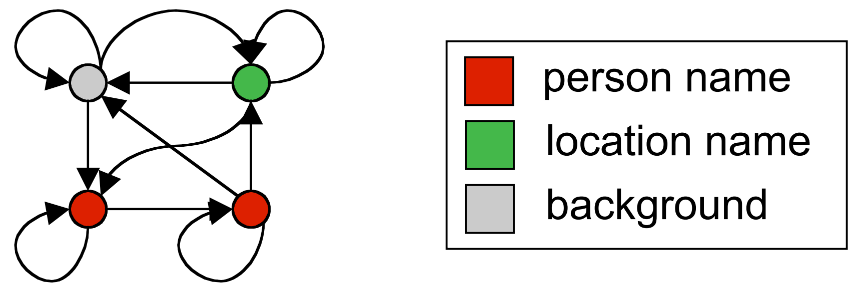
- **Maximize probability of training observations (w/ prior)**

IE with Hidden Markov Models

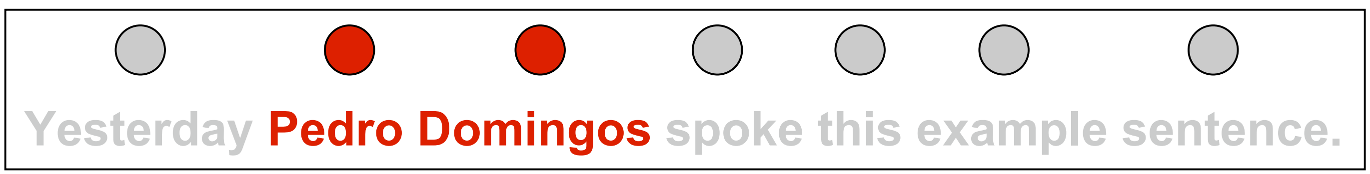
Given a sequence of observations:

Yesterday Pedro Domingos spoke this example sentence.

and a trained HMM:



Find the most likely state sequence: (Viterbi)



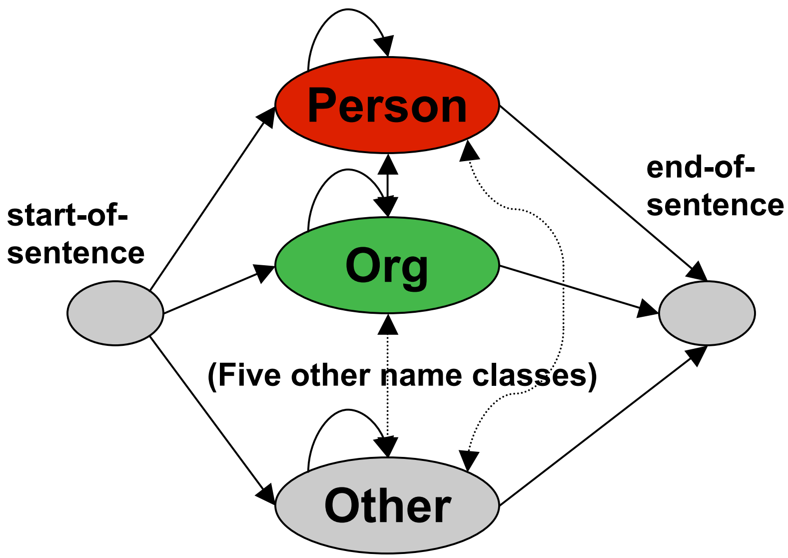
Any words said to be generated by the designated "person name" state extract as a person name:

Person name: Pedro Domingos

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:

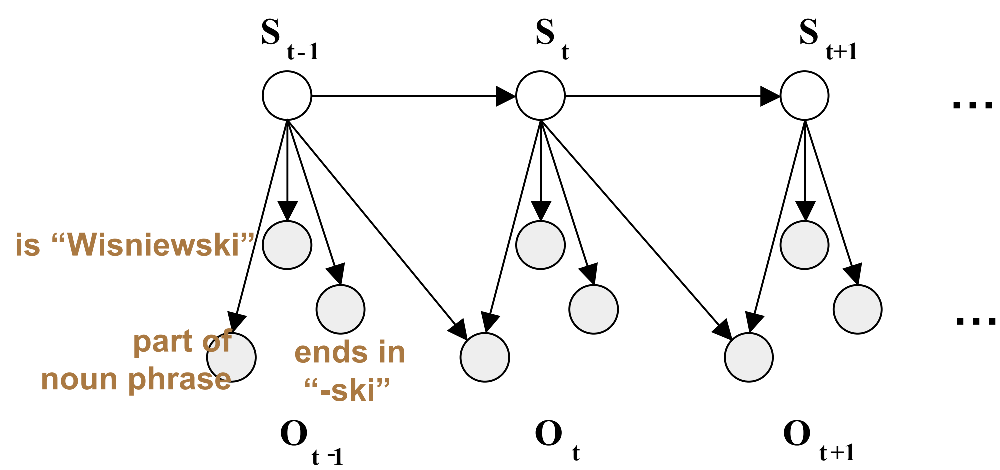
Case	Language	F1 .
Mixed	English	93%
Upper	English	91%
Mixed	Spanish	90%

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

We want More than an Atomic View of Words

Would like richer representation of text: many arbitrary, overlapping features of the words.

- identity of word
- ends in "-ski"
- is capitalized
- is part of a noun phrase
- is in a list of city names
- is under node X in WordNet
- is in bold font
- is indented
- is in hyperlink anchor
- last person name was female
- next two words are "and Associates"



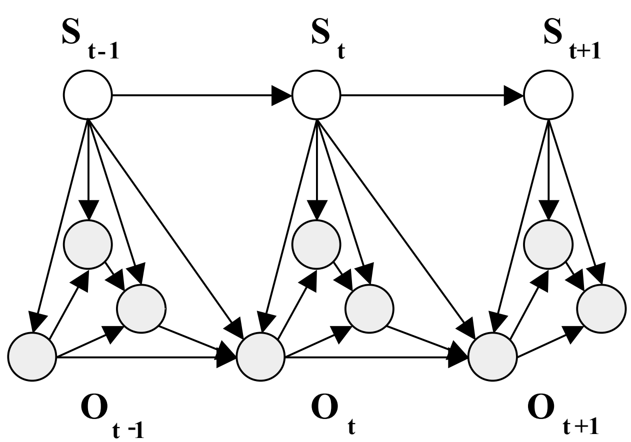
Problems with Richer Representation and a Generative Model

These arbitrary features are not independent.

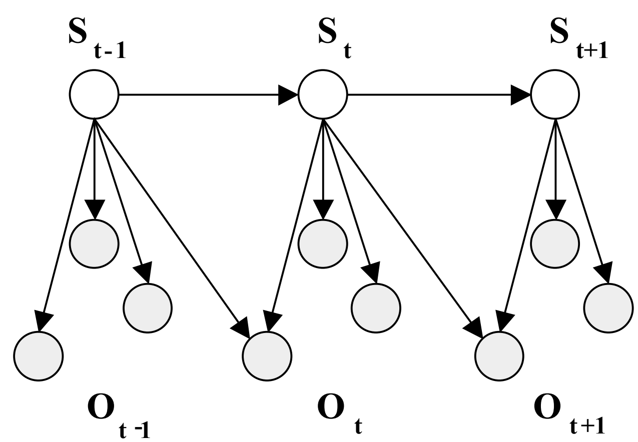
- Multiple levels of granularity (chars, words, phrases)
- Multiple dependent modalities (words, formatting, layout)
- Past & future

Two choices:

Model the dependencies. Each state would have its own Bayes Net. *But we are already starved for training data!*



Ignore the dependencies. This causes “over-counting” of evidence (ala naïve Bayes). *Big problem when combining evidence, as in Viterbi!*



Conditional Sequence Models

- We prefer a model that is trained to maximize a *conditional* probability rather than *joint* probability: **$P(\overline{s}|\overline{o})$ instead of $P(\overline{s},\overline{o})$:**
  - Can examine features, but not responsible for generating them.
  - Don’t have to explicitly model their dependencies.
  - Don’t “waste modeling effort” trying to generate what we are given at test time anyway.

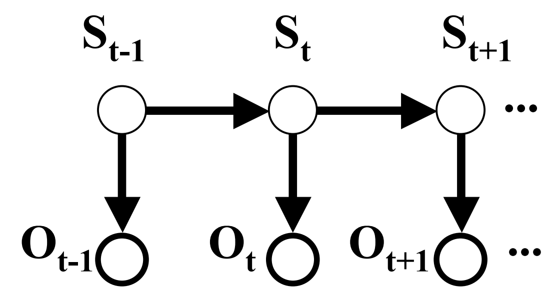
Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

From HMMs to Conditional Random Fields

[Lafferty, McCallum, Pereira 2001]

$\vec{s} = s_1, s_2, \ldots s_n \qquad \vec{o} = o_1, o_2, \ldots o_n$



Joint

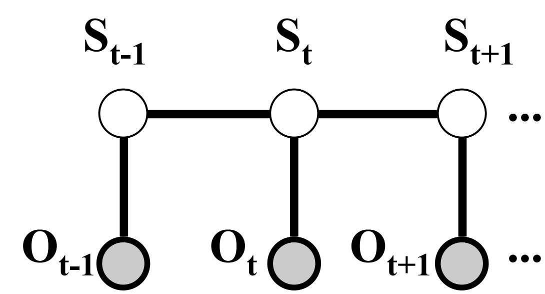
$$P(\vec{s},\vec{o}) = \prod_{t=1}^{|\vec{o}|} P(s_t \mid s_{t-1}) P(o_t \mid s_t)$$

Conditional

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

$$= \frac{1}{Z(\vec{o})} \prod_{t=1}^{|\vec{o}|} \Phi_s(s_t, s_{t-1}) \Phi_o(o_t, s_t)$$

where $\Phi_o(t) = \exp\left(\sum_k \lambda_k f_k(s_t, o_t)\right)$

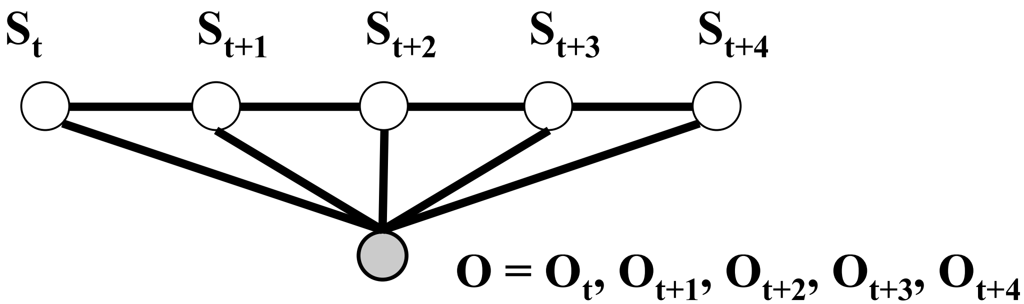


(A super-special case of Conditional Random Fields.)

Set parameters by maximum likelihood, using optimization method on δ*L*.

Linear Chain Conditional Random Fields

[Lafferty, McCallum, Pereira 2001]



Markov on *s*, conditional dependency on *o.*

$$P(\vec{s} \mid \vec{o}) \propto \frac{1}{Z_{\vec{o}}} \prod_{t=1}^{|\vec{o}|} \exp\left(\sum_j \lambda_j f_j(s_t, s_{t-1}, \vec{o}, t) \right)$$

Hammersley-Clifford-Besag theorem stipulates that the CRF has this form—an exponential function of the cliques in the graph.

Assuming that the dependency structure of the states is tree-shaped (linear chain is a trivial tree), inference can be done by dynamic programming in time $O(|o|\ |S|^2)$—just like HMMs.

CRFs vs. HMMs

- More general and expressive modeling technique
- Comparable computational efficiency
- Features may be arbitrary functions of *any* or *all* observations
- Parameters need not fully specify generation of observations; require less training data
- Easy to incorporate domain knowledge
- State means only “state of process”, vs
“state of process” and “observational history I’m keeping”

Training CRFs

$$\text{Maximize log - likelihood of parameters given training data :}$$

$$L(\{\lambda_k\} \mid \{\langle \vec{o}, \vec{s} \rangle^{(i)}\})$$

$$\text{Log - likelihood gradient :}$$

$$\frac{\partial L}{\partial \lambda_k} = \sum_i C_k(\vec{s}^{(i)}, \vec{o}^{(i)}) - \sum_i \sum_{\vec{s}} P_{\{\lambda_k\}}(\vec{s} \mid \vec{o}^{(i)})\ C_k(\vec{s}, \vec{o}^{(i)}) - \lambda_k^{\ 2}$$

$$C_k(\vec{s}, \vec{o}) = \sum_t f_k(\vec{o}, t, s_{t-1}, s_t)$$

Feature count using correct labels **-** **Feature count using predicted labels** **-** **Smoothing penalty**

Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

Table Extraction from Government Reports

Cash receipts from marketings of milk during 1995 at $19.9 billion dollars, was slightly below 1994. Producer returns averaged $12.93 per hundredweight, $0.19 per hundredweight below 1994. Marketings totaled 154 billion pounds, 1 percent above 1994. Marketings include whole milk sold to plants and dealers as well as milk sold directly to consumers.

An estimated 1.56 billion pounds of milk were used on farms where produced, 8 percent less than 1994. Calves were fed 78 percent of this milk with the remainder consumed in producer households.

Milk Cows and Production of Milk and Milkfat:
United States, 1993-95

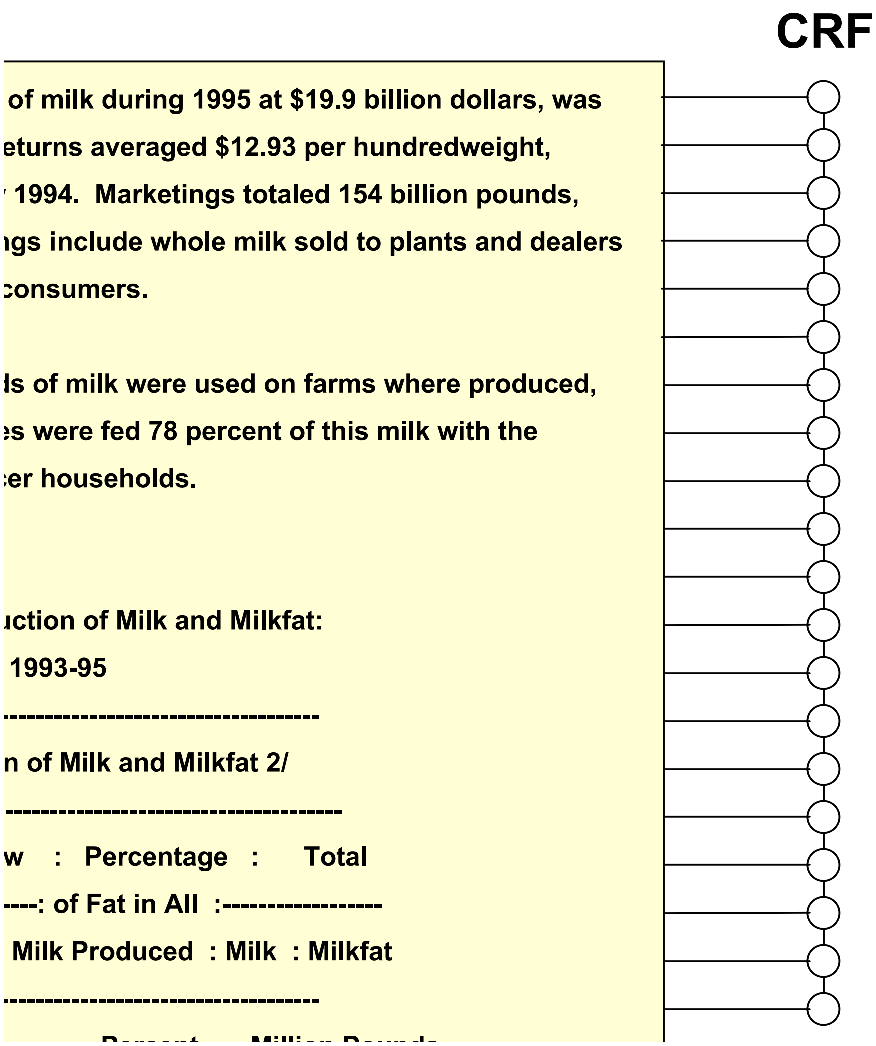
Year	Number of Milk Cows 1/	Production of Milk and Milkfat 2/				
		Per Milk Cow		Percentage of Fat in All Milk Produced	Total	
		Milk	Milkfat		Milk	Milkfat
	1,000 Head	--- Pounds ---		Percent	Million Pounds	
1993	9,589	15,704	575	3.66	150,582	5,514.4
1994	9,500	16,175	592	3.66	153,664	5,623.7
1995	9,461	16,451	602	3.66	155,644	5,694.3

1/ Average number during year, excluding heifers not yet fresh.
2/ Excludes milk sucked by calves.

Table Extraction from Government Reports

[Pinto, McCallum, Wei, Croft, 2003 SIGIR]

100+ documents from www.fedstats.gov



Labels:

- Non-Table
- Table Title
- Table Header
- Table Data Row
- Table Section Data Row
- Table Footnote
- ... *(12 in all)*

Features:

- Percentage of digit chars
- Percentage of alpha chars
- Indented
- Contains 5+ consecutive spaces
- Whitespace in this line aligns with prev.
- ...
- Conjunctions of all previous features, time offset: {0,0}, {-1,0}, {0,1}, {1,2}.

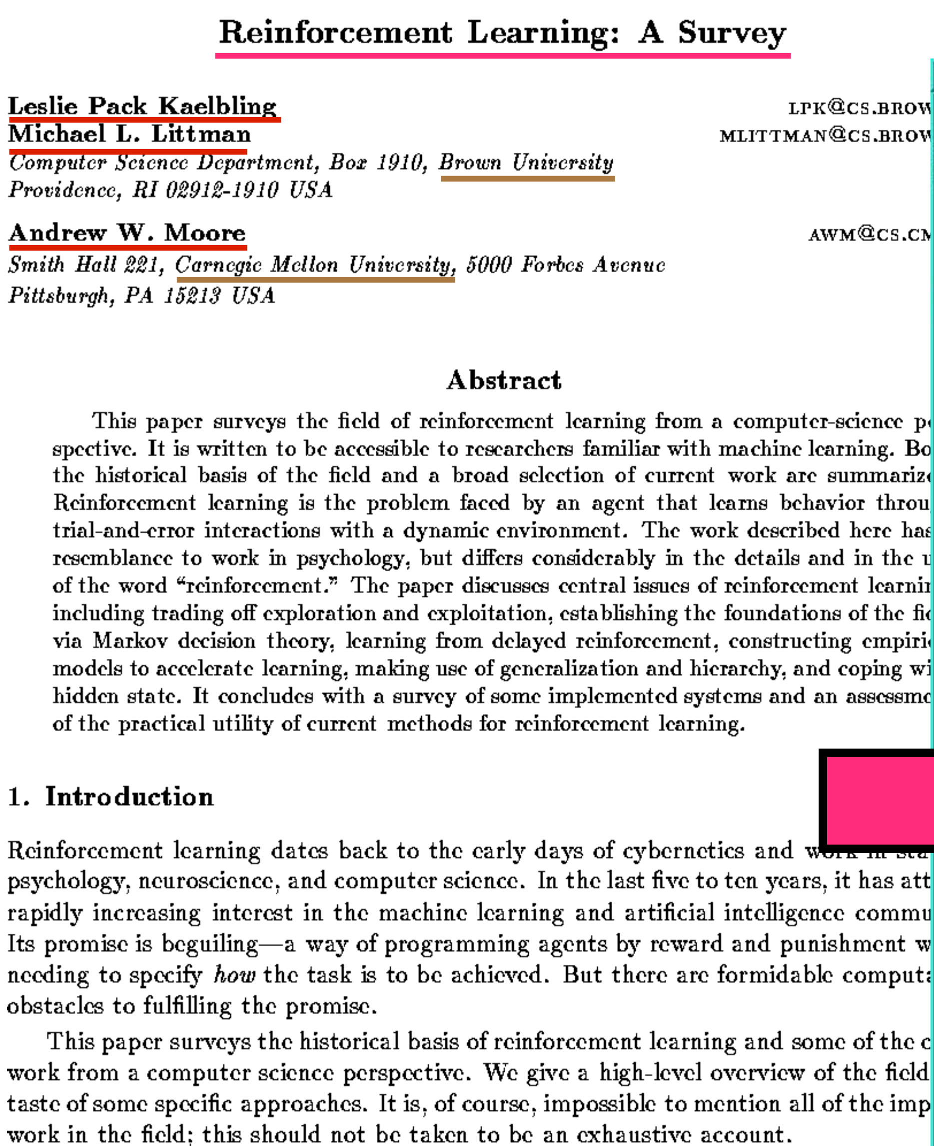
Table Extraction Experimental Results

[Pinto, McCallum, Wei, Croft, 2003 SIGIR]

	Line labels, percent correct	Table segments, F1
HMM	65 %	64 %
Stateless MaxEnt	85 %	-
CRF	95 %	92 %

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 pe 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 fie 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 assessme 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.

Title, author, institution and abstract are automatically extracted, and are often, but not always correct.

Number of hits found: 64

1. A Machine Learning Architecture for Optimizing Web Search Engines
Justin Boyan, Dayne Freitag, and Thorsten Joachims
Abstract: Indexing systems for the World Wide Web, such as Lycos and Alta Vista, play an essential role in m useful and usable. These systems are based on Information Retrieval methods for indexing plain text document heuristics for adjusting their document rankings based on the special HTML structure of Web documents. In th describe a wide range of such heuristicslincluding a novel one inspired by reinforcement learning techniques fo rewards through a graphlwhich can be used to affect a search engine's rankings. We then demonstrate a syste combine these heuristics automatically, based on feedback collected unintrusively from users, resulting in muc rankings.
Postscript Referring Page Details BibTeX Entry Word Matches: boyan, search engines Score: 1

2. Value Function Based Production Scheduling
Jeff G. Schneider Justin A. Boyan Andrew W. Moore
Abstract: Production scheduling, the problem of sequentially configuring a factory to meet forecasted demands problem throughout the manufacturing industry. The requirement of maintaining product inventories in the face demand and stochastic factory output makes standard scheduling models, such as job-shop, inadequate. Curre algorithms, such as simulated annealing and constraint propagation, must employ ad-hoc methods such as freq cope with uncertainty. In this paper, we describe a Markov Decision Process (MDP) formulation of production captures stochasticity in both production and demands. The solution to this MDP is a value function which can generate optimal scheduling decisions online. A simple example illustrates the theoretical superiority of this ap replanning-based methods. We then describe an industrial application and two reinforcement learning methods approximate value function on this domain. Our results demonstrate that in both deterministic and noisy scenar approximation is an effective technique.
Postscript Referring Page Details BibTeX Entry Word Matches: boyan Score: 0.6094

3. Least-Squares Temporal Difference Learning
Justin A. Boyan
Abstract: Submitted to NIPS-98 TD() is a popular family of algorithms for approximate policy evaluation in lar works by incrementally updating the value function after each observed transition. It has two major drawbacks: inefficient use of data, and it requires the user to manually tune a stepsize schedule for good performance. For value function approximations and = 0, the Least-Squares TD (LSTD) algorithm of Bradtke and Barto [5] elimi parameters and improves data efficiency. This paper extends Bradtke and Barto's work in three significant way presents a simpler derivation of the LSTD algorithm. Second, it generalizes from = 0 to arbitrary values of ; at t the resulting algorithm is shown to be a practical formulation of supervised linear regression. Third, it presents

IE from Research Papers

	Field-level F1
Hidden Markov Models (HMMs) *[Seymore, McCallum, Rosenfeld, 1999]*	**75.6**
Support Vector Machines (SVMs) *[Han, Giles, et al, 2003]*	**89.7**
Conditional Random Fields (CRFs) *[Peng, McCallum, 2004]*	**93.9**

Δ error 40%

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

Automatically Induced Features

[McCallum & Li, 2003, CoNLL]

Index	*Feature*
0	inside-noun-phrase (o_{t-1})
5	stopword (o_t)
20	capitalized (o_{t+1})
75	word=the (o_t)
100	in-person-lexicon (o_{t-1})
200	word=in (o_{t+2})
500	word=Republic (o_{t+1})
711	word=RBI (o_t) & header=BASEBALL
1027	header=CRICKET (o_t) & in-English-county-lexicon (o_t)
1298	company-suffix-word ($firstmention_{t+2}$)
4040	location (o_t) & POS=NNP (o_t) & capitalized (o_t) & stopword (o_{t-1})
4945	moderately-rare-first-name (o_{t-1}) & very-common-last-name (o_t)
4474	word=the (o_{t-2}) & word=of (o_t)

Named Entity Extraction Results

[McCallum & Li, 2003, CoNLL]

Method	F1
HMMs BBN's Identifinder	73%
CRFs w/out Feature Induction	83%
CRFs with Feature Induction based on LikelihoodGain	90%

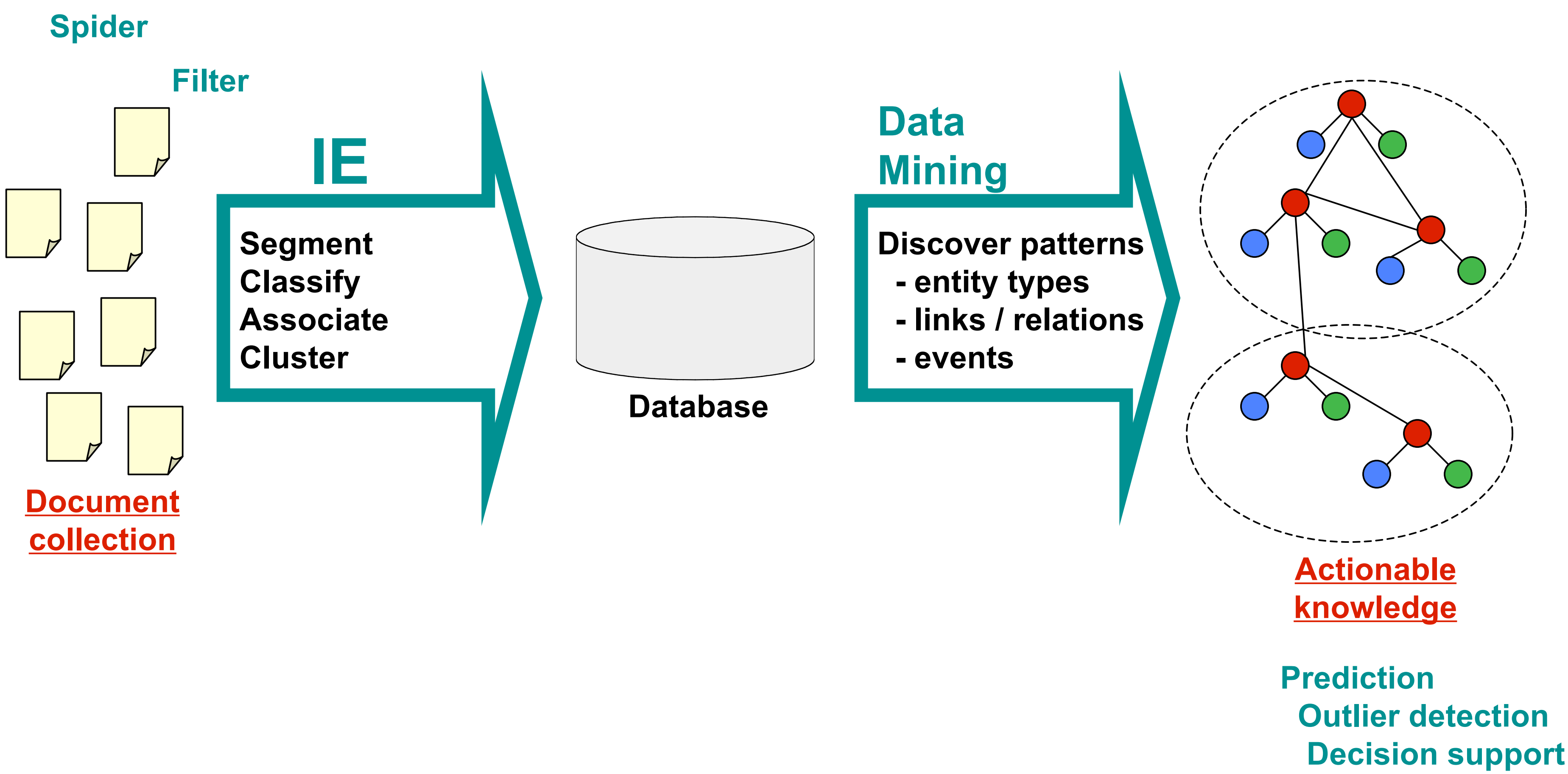
Related Work

- CRFs are widely used for information extraction ...including more complex structures, like trees:
  - [Zhu, Nie, Zhang, Wen, ICML 2007] Dynamic Hierarchical Markov Random Fields and their Application to Web Data Extraction
  - [Viola & Narasimhan]: Learning to Extract Information from Semi-structured Text using a Discriminative Context Free Grammar
  - [Jousse et al 2006]: Conditional Random Fields for XML Trees

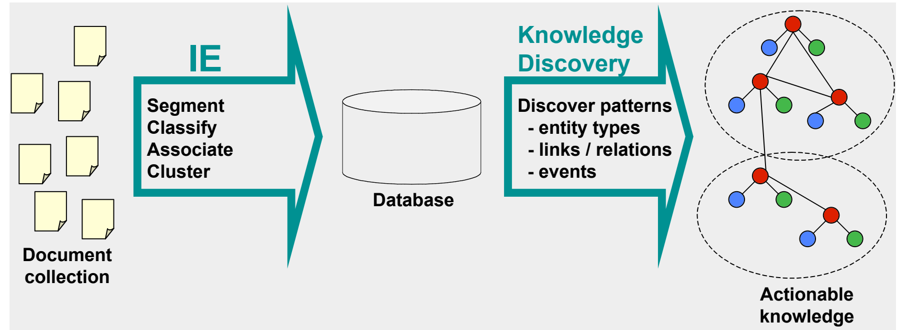
Outline

- Examples of IE and Data Mining
- Landscape of problems and solutions
- Techniques for Segmentation and Classification
  - Sliding Window and Boundary Detection
  - IE with Hidden Markov Models
  - Introduction to Conditional Random Fields (CRFs)
  - Examples of IE with CRFs
- IE + Data Mining

From Text to Actionable Knowledge



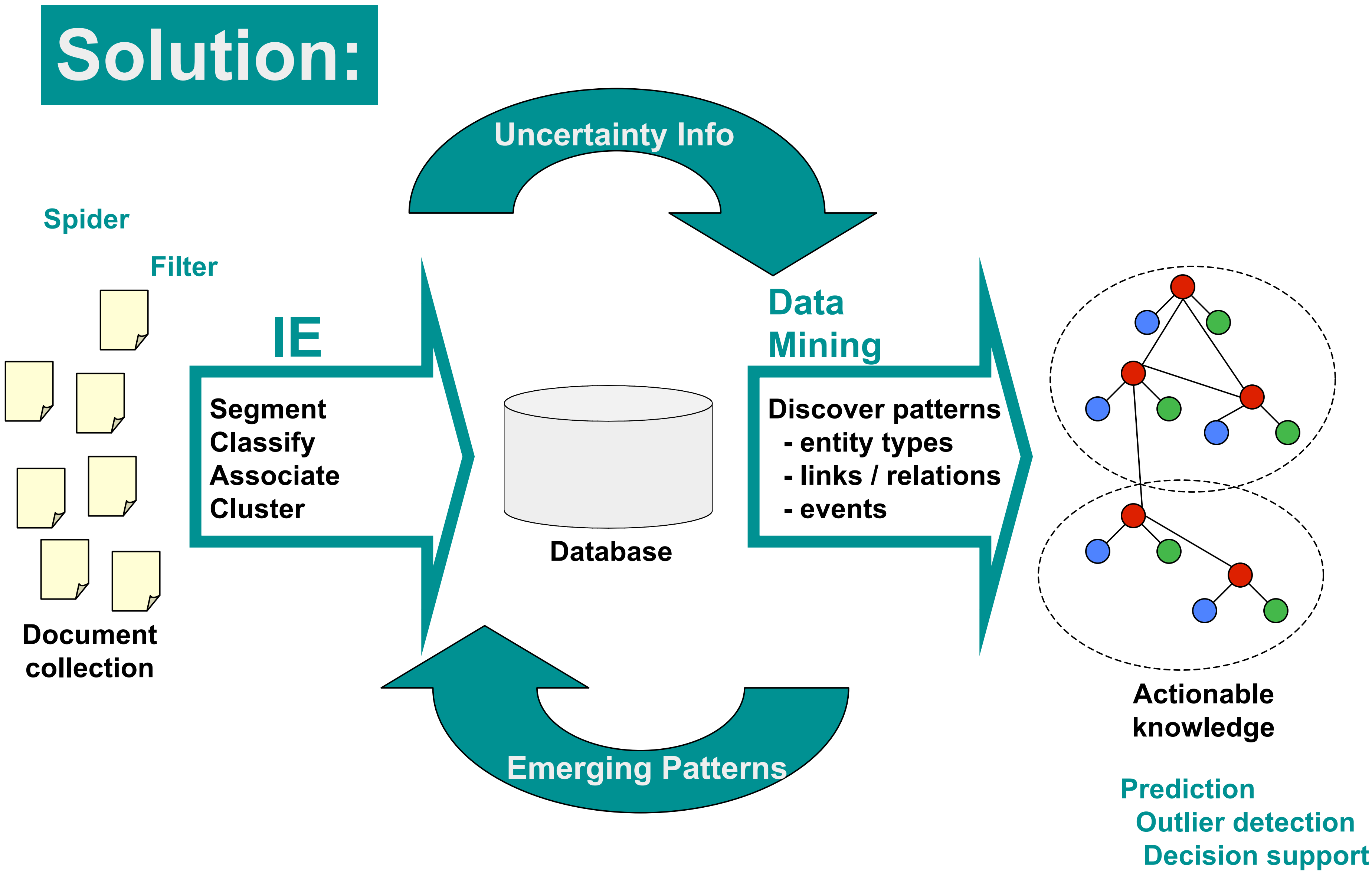
Problem:



Combined in serial juxtaposition, IE and DM are unaware of each others' weaknesses and opportunities.

1) DM begins from a populated DB, unaware of where the data came from, or its inherent errors and uncertainties.
2) IE is unaware of emerging patterns and regularities in the DB.

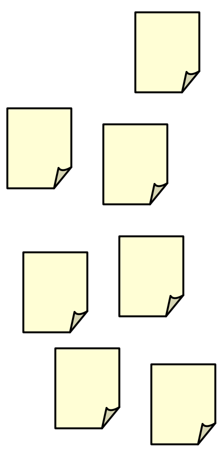
The accuracy of both suffers, and significant mining of complex text sources is beyond reach.


Solution:
Uncertainty Info
Spider
Filter
IE
Segment
Classify
Associate
Cluster
Database
Data Mining
Discover patterns
- entity types
- links / relations
- events
Document collection
Actionable knowledge
Emerging Patterns
Prediction
Outlier detection
Decision support

Solution:

Spider

Filter



Document collection

Unified Model

IE

Segment
Classify
Associate
Cluster

Probabilistic Model

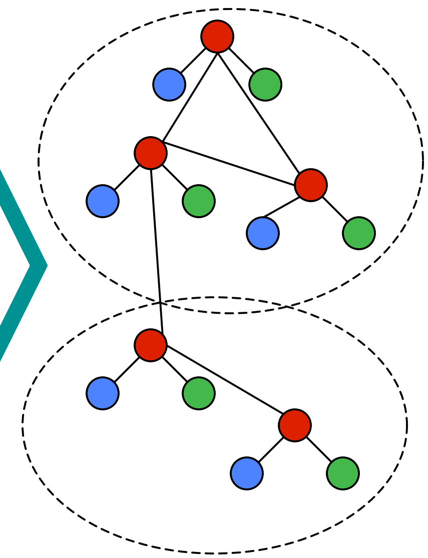
Data Mining

Discover patterns
- entity types
- links / relations
- events

Discriminatively-trained undirected graphical models

Conditional Random Fields
[Lafferty, McCallum, Pereira]

Conditional PRMs
[Koller…], [Jensen…], [Geetor…], [Domingos…]



Actionable knowledge

Prediction
Outlier detection
Decision support

Complex Inference and Learning

Just what we researchers like to sink our teeth into!

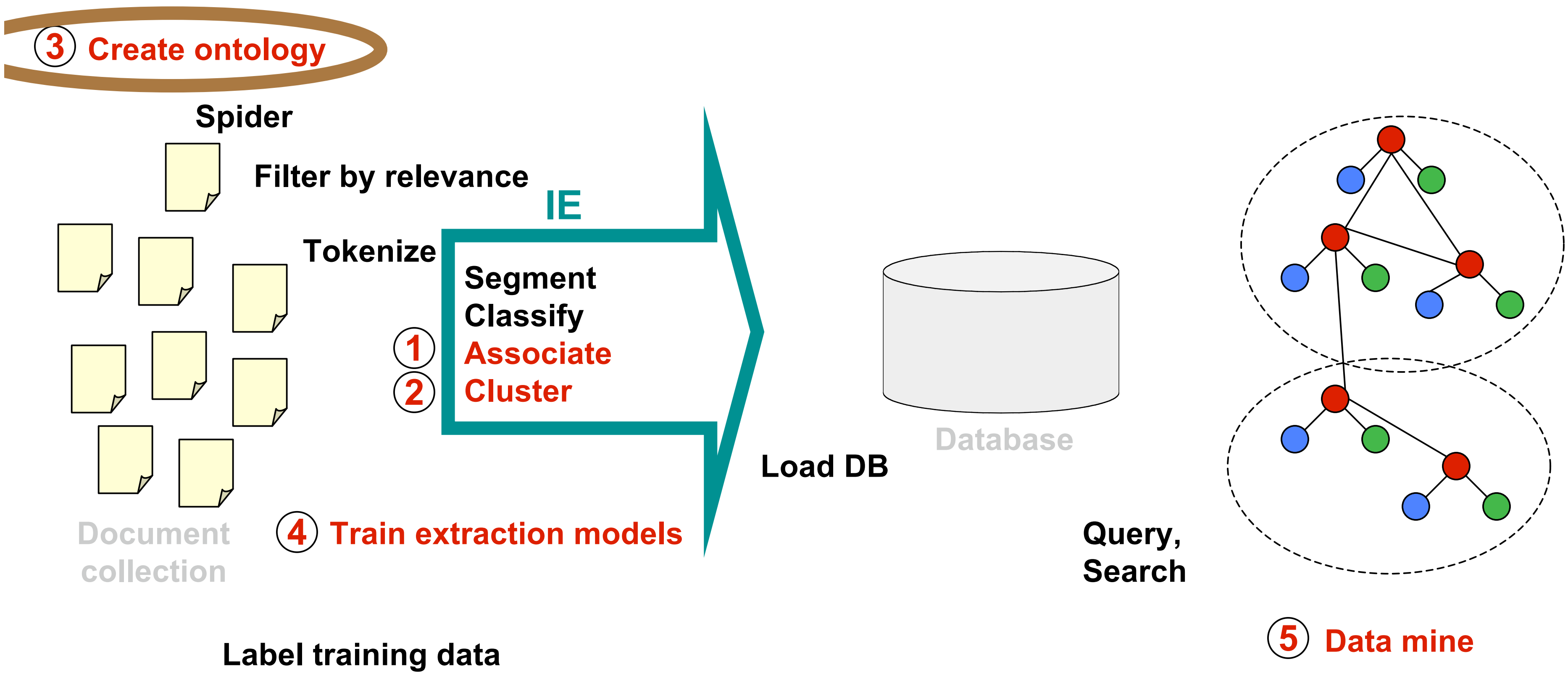
Scientific Questions

- What model structures will capture salient dependencies?
- Will joint inference actually improve accuracy?

- How to do ***inference*** in these large graphical models?
- How to do ***parameter estimation*** efficiently in these models, which are built from multiple large components?
- How to do ***structure discovery*** in these models?

Broader View

Now touch on some other issues

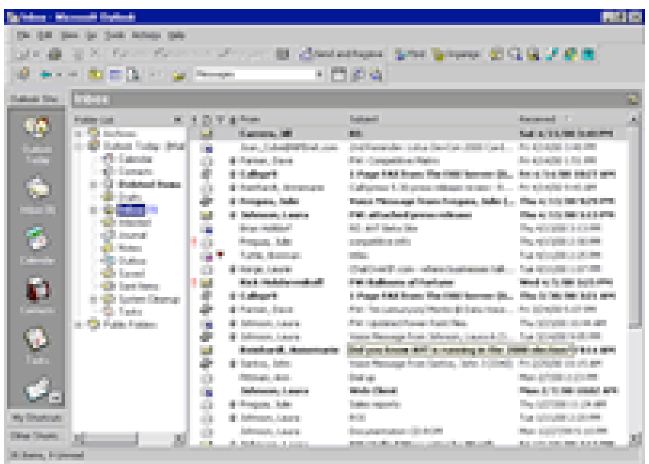


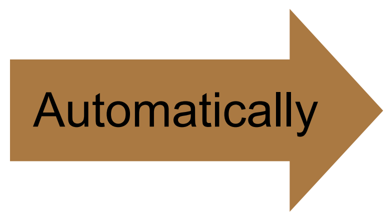
Managing and Understanding Connections of People in our Email World

Workplace effectiveness ~ Ability to leverage network of acquaintances

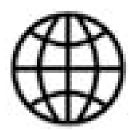
But filling Contacts DB by hand is tedious, and incomplete.

Email Inbox

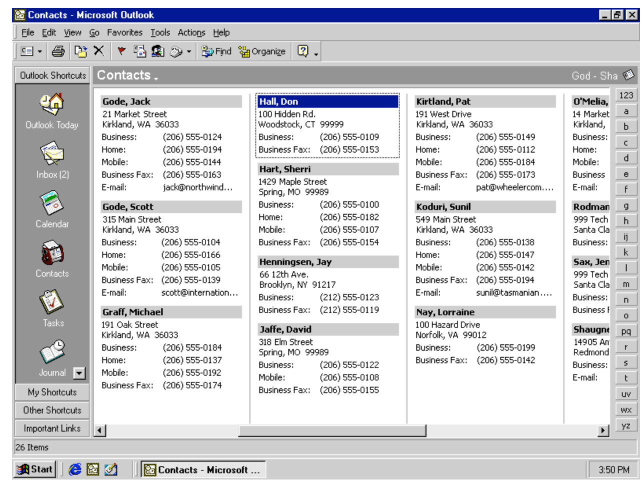




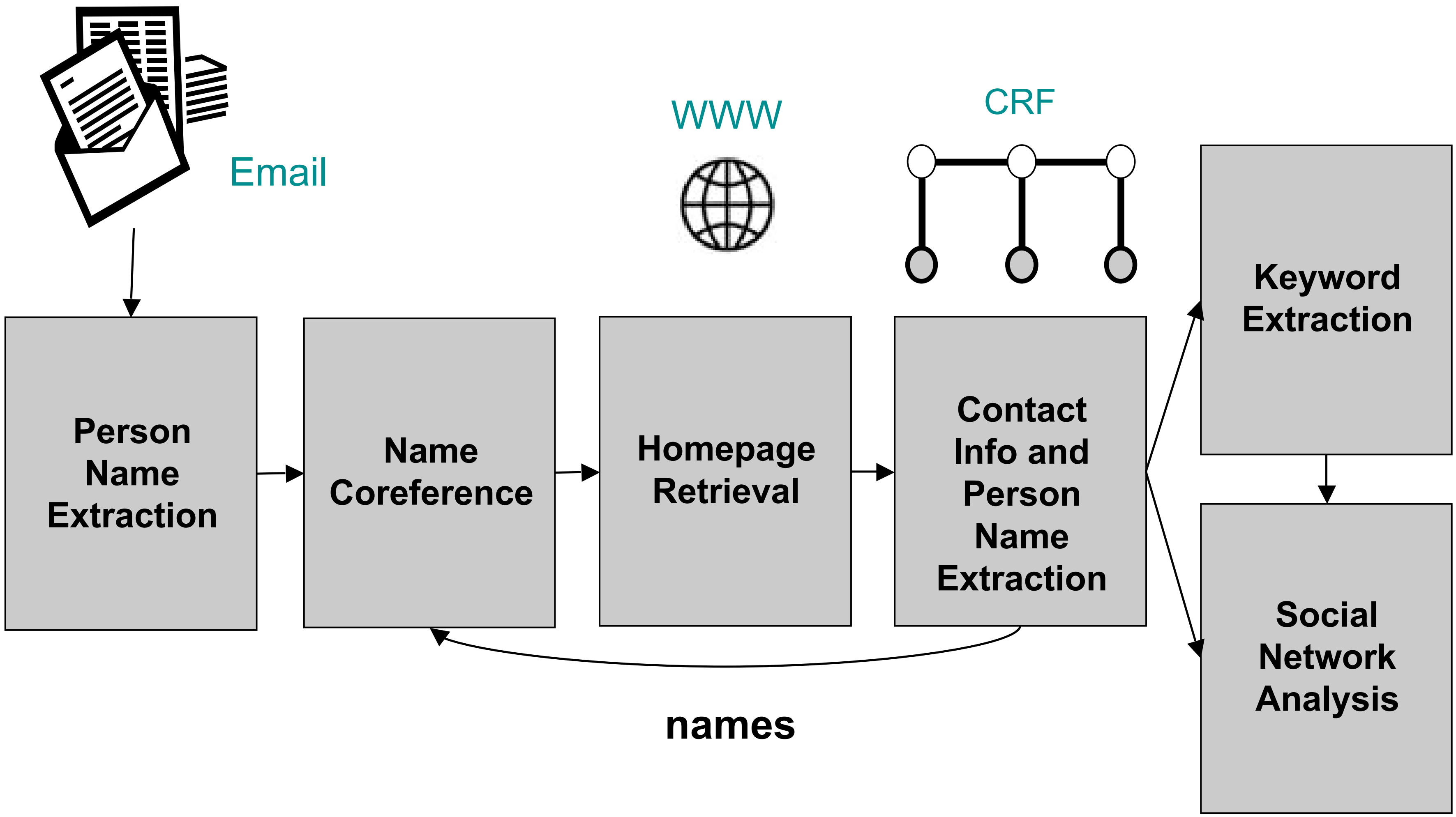
WWW



Contacts DB



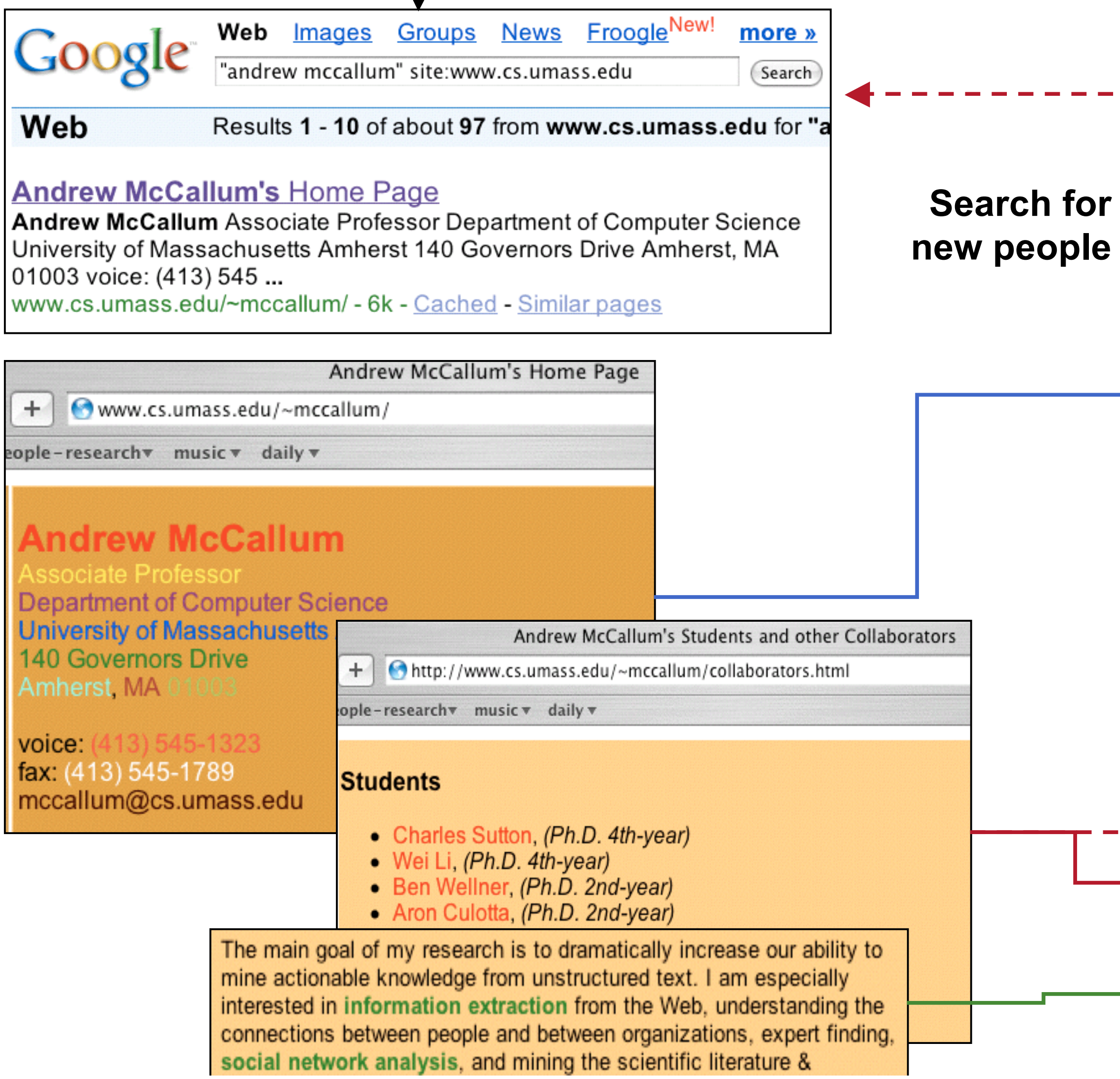
System Overview



An Example

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

Subject ...



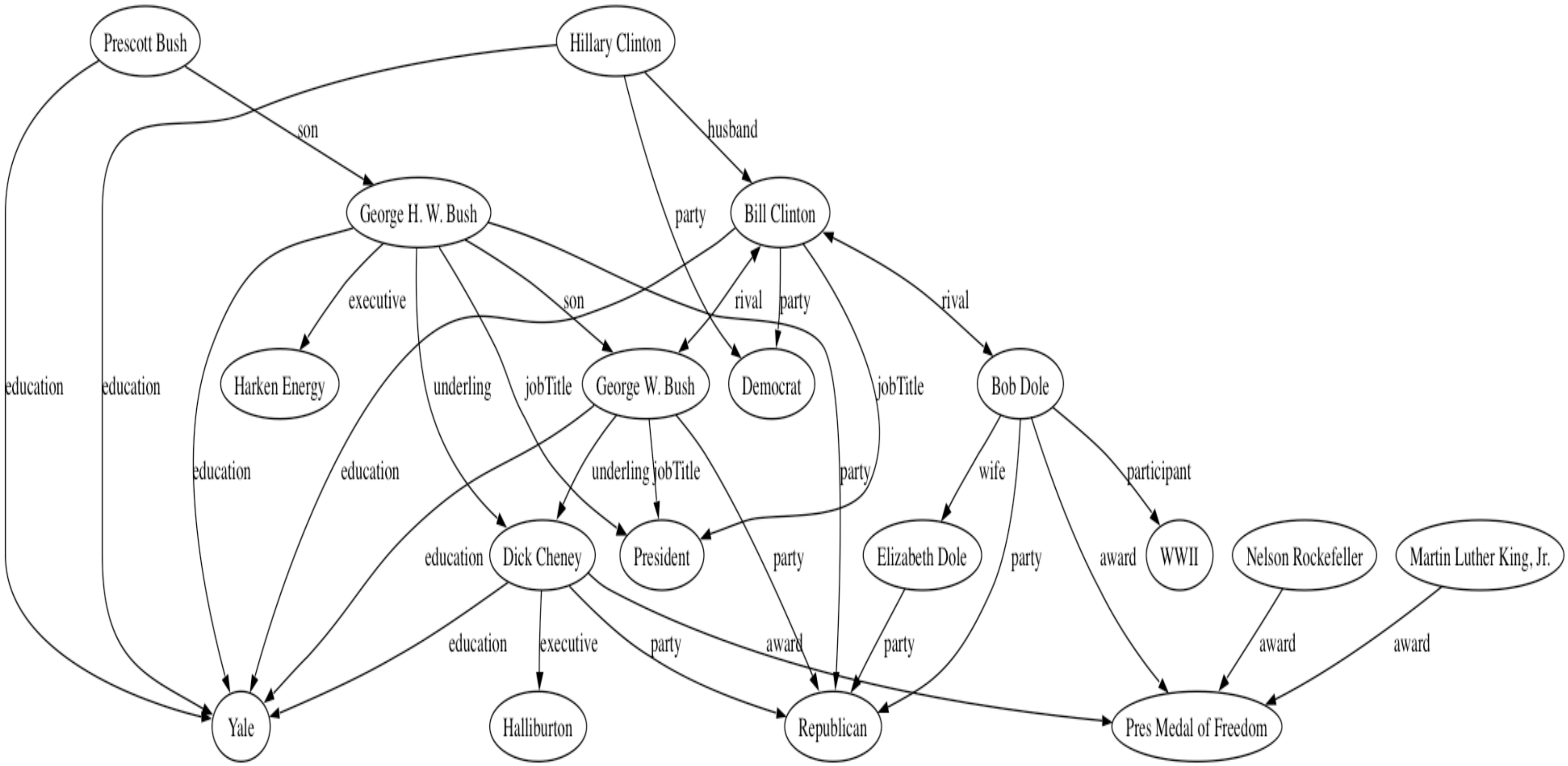
First Name:	Andrew
Middle Name:	Kachites
Last Name:	McCallum
JobTitle:	Associate Professor
Company:	University of Massachusetts
Street Address:	140 Governor's Dr.
City:	Amherst
State:	MA
Zip:	01003
Company Phone:	(413) 545-1323
Links:	Fernando Pereira, Sam Roweis,…
Key Words:	Information extraction,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

Relation Extraction - Data

- 270 Wikipedia articles
- 1000 paragraphs
- 4700 relations

- 52 relation types
  – JobTitle, BirthDay, Friend, Sister, Husband, Employer, Cousin, Competition, Education, …

- Targeted for density of relations
  – Bush/Kennedy/Manning/Coppola families and friends



Prescott Bush
Hillary Clinton
son
husband
George H. W. Bush
party
Bill Clinton
executive
son
rival
party
rival
education
education
Harken Energy
underling
jobTitle
George W. Bush
Democrat
jobTitle
Bob Dole
education
education
underling
jobTitle
party
wife
participant
education
Dick Cheney
President
party
Elizabeth Dole
party
award
WWII
Nelson Rockefeller
Martin Luther King, Jr.
education
executive
party
award
party
award
award
Yale
Halliburton
Republican
Pres Medal of Freedom

George W. Bush

…his father George H. W. Bush…

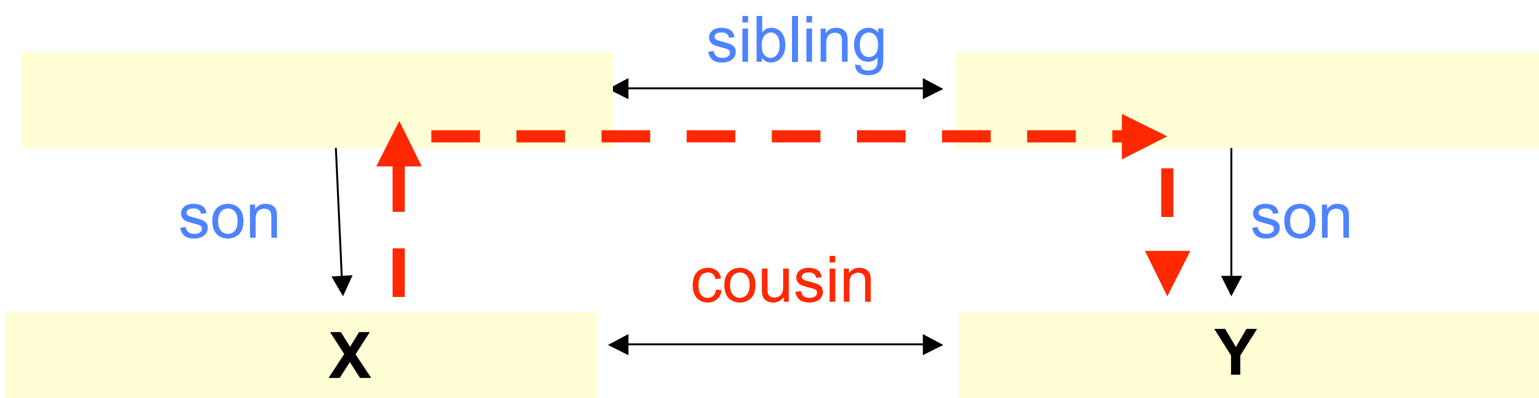
George H. W. Bush

…his sister Nancy Ellis Bush…

Nancy Ellis Bush

…her son John Prescott Ellis…

Cousin = Father's Sister's Son

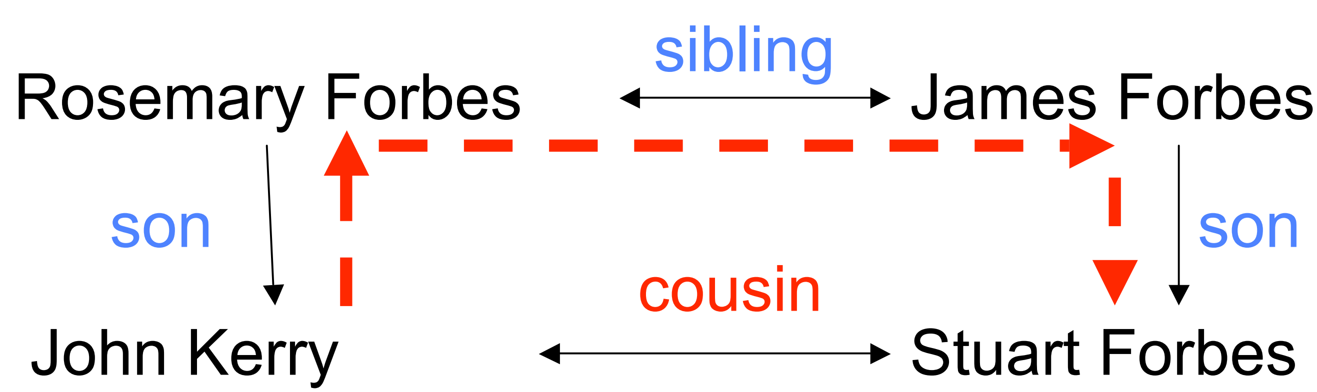


John Kerry likely a cousin

…celebrated with Stuart Forbes…

Name	Son
Rosemary Forbes	John Kerry
James Forbes	Stuart Forbes

Name	Sibling
Rosemary Forbes	James Forbes



Examples of Discovered Relational Features

- Mother: Father→Wife
- Cousin: Mother→Husband→Nephew
- Friend: Education→Student
- Education: Father→Education
- Boss: Boss→Son
- MemberOf: Grandfather→MemberOf
- Competition: PoliticalParty→Member→Competition