Computer Workstations as Intelligent Agents

Tom M. Mitchell

Center for Automated Learning and Discovery
Carnegie Mellon University

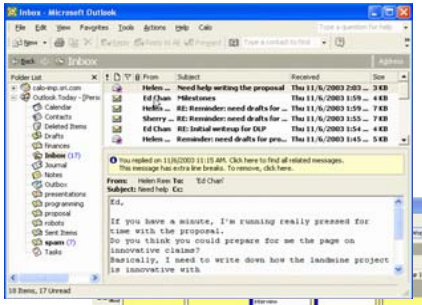
SIGMOD Keynote Talk
June, 2005

Thanks to many collaborators:
Vitor Carvalho, William Cohen, Dinesh Govindaraju,
Yifen Huang, Sophie Wang, and the entire CALO team

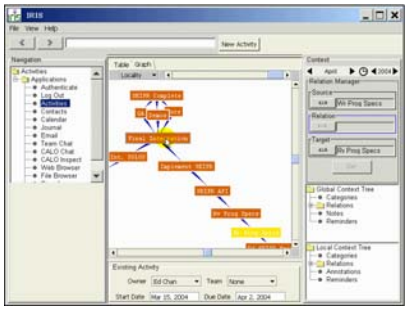
Our workstations are huge semi-structured databases

Why don’t we have useful intelligent agents to operate over them?

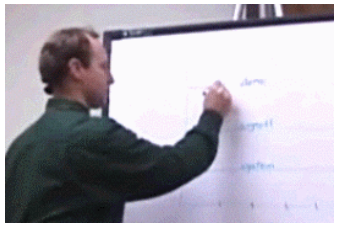
Technical Challenges



- Interpreting all that text
  - and mix of structured/unstructured data
  - .jpg, .wav, .pdf, .dat, .ppt, .txt, .dba, …

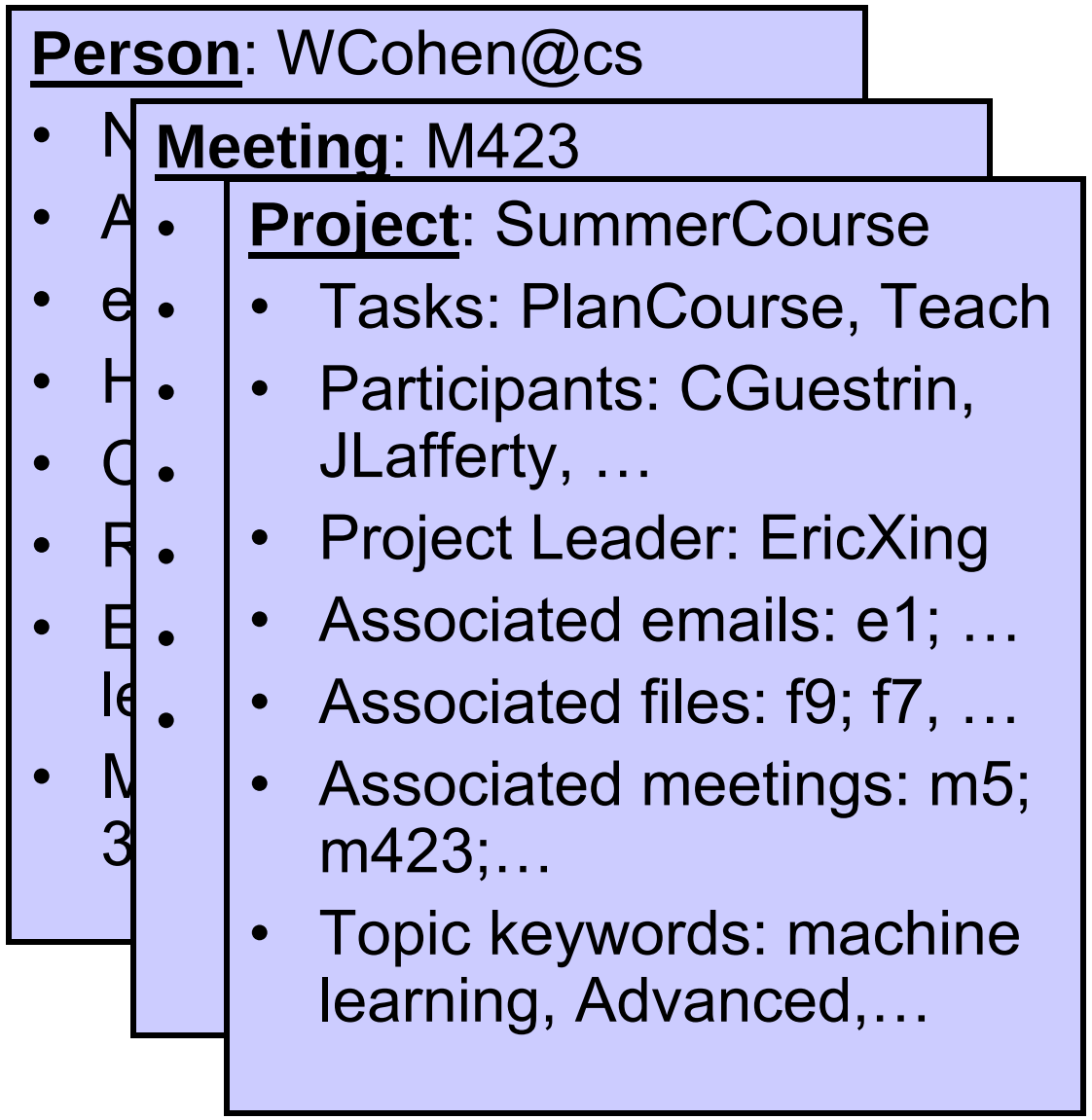


- Learning to customize to user
  - Scheduling, email preferences
  - Work routines
  - What parts of web and world does user (not) know about



- Perception of environment and user's world
  - Seeing/hearing office activity
  - Inferring work groups, friends, strangers, …

What Structured Data Types Should We Extract from Workstations?



Person

Meeting

Project

Task

Negotiation

File

WebPage

Date/Time

Deadline

Organization

Just a few dozen should do…

What Good Is Stuctured Data?

Person: WCohen@cs

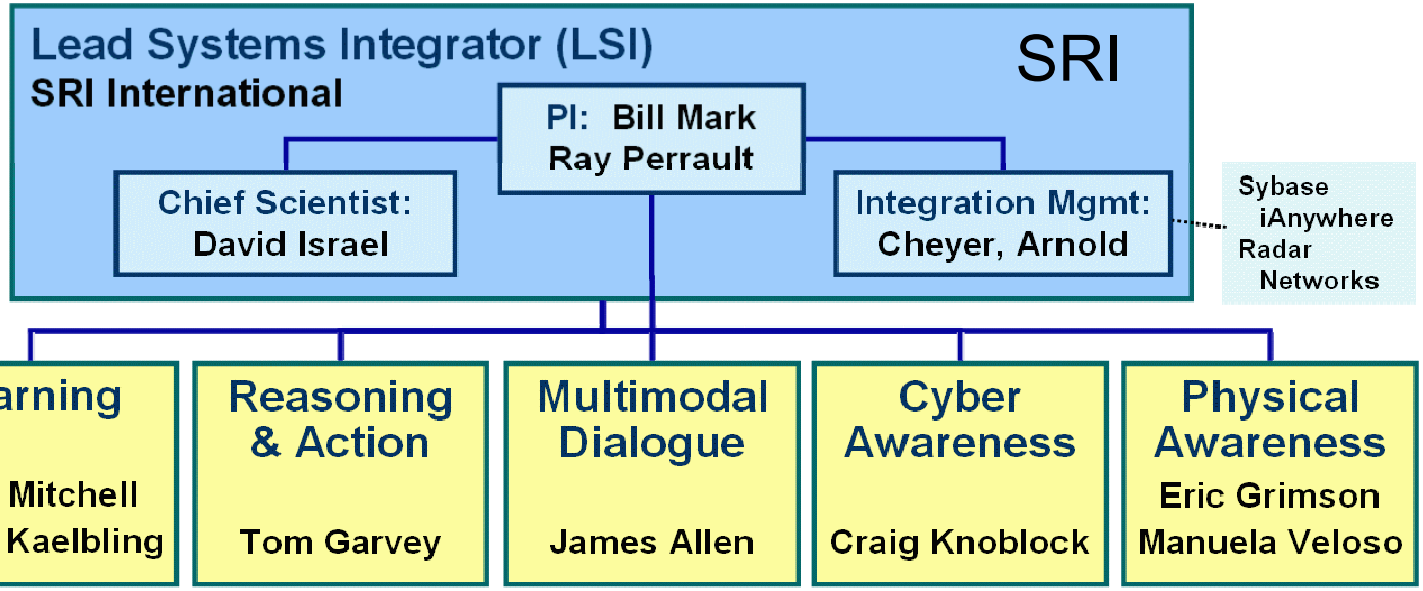
- Name: William Cohen
- A
- e
- H
- C
- F
- E
- I

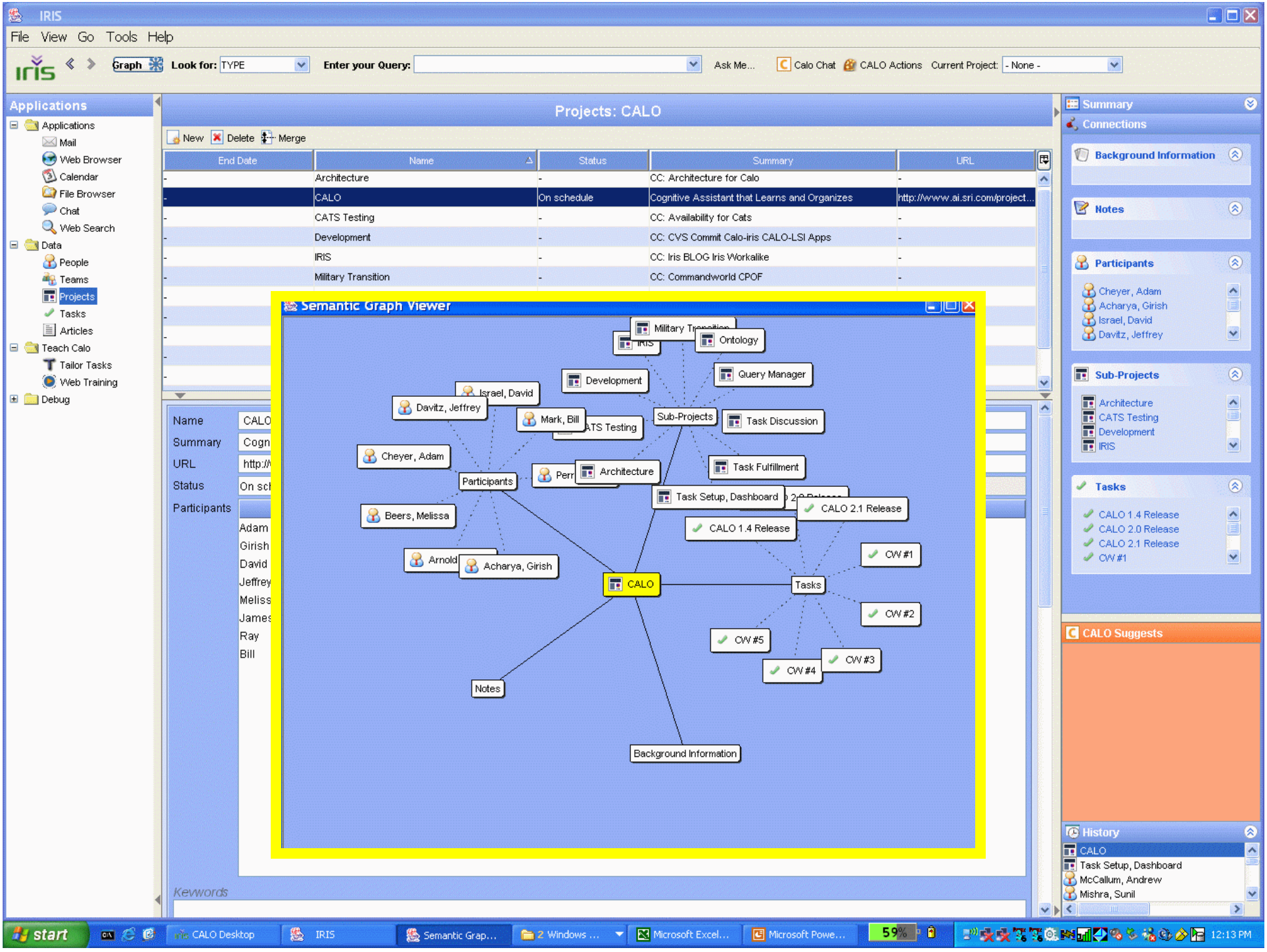
Project: SummerCourse

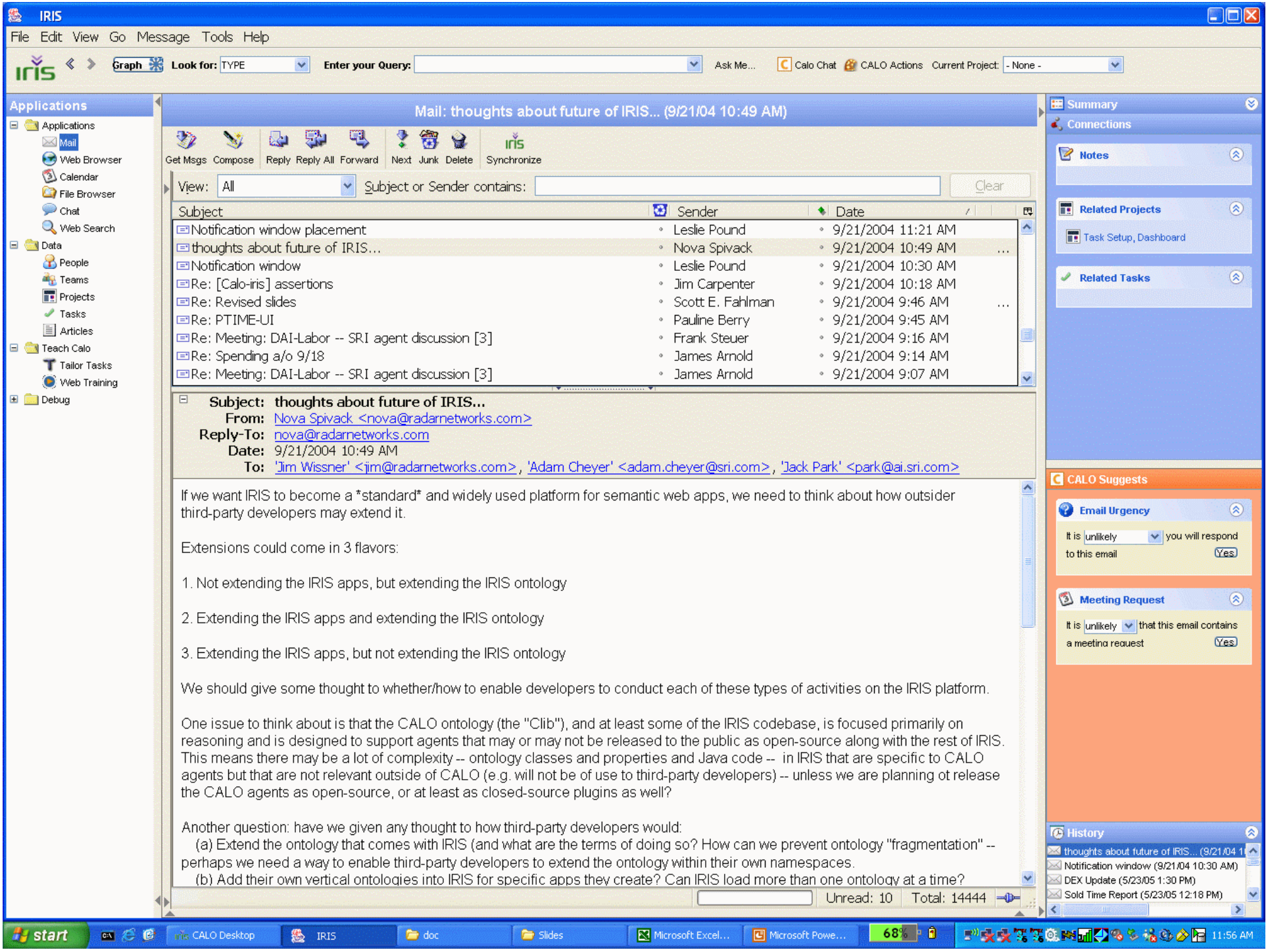
- Type: Course Offering
- Participants: CGuestrin, JLafferty, …
- Project Leader: EricXing
- Associated emails: e1; …
- Associated files: f9; f7, …
- Associated Web pages: w5; w8;…
- Topic keywords: machine learning, Advanced,…

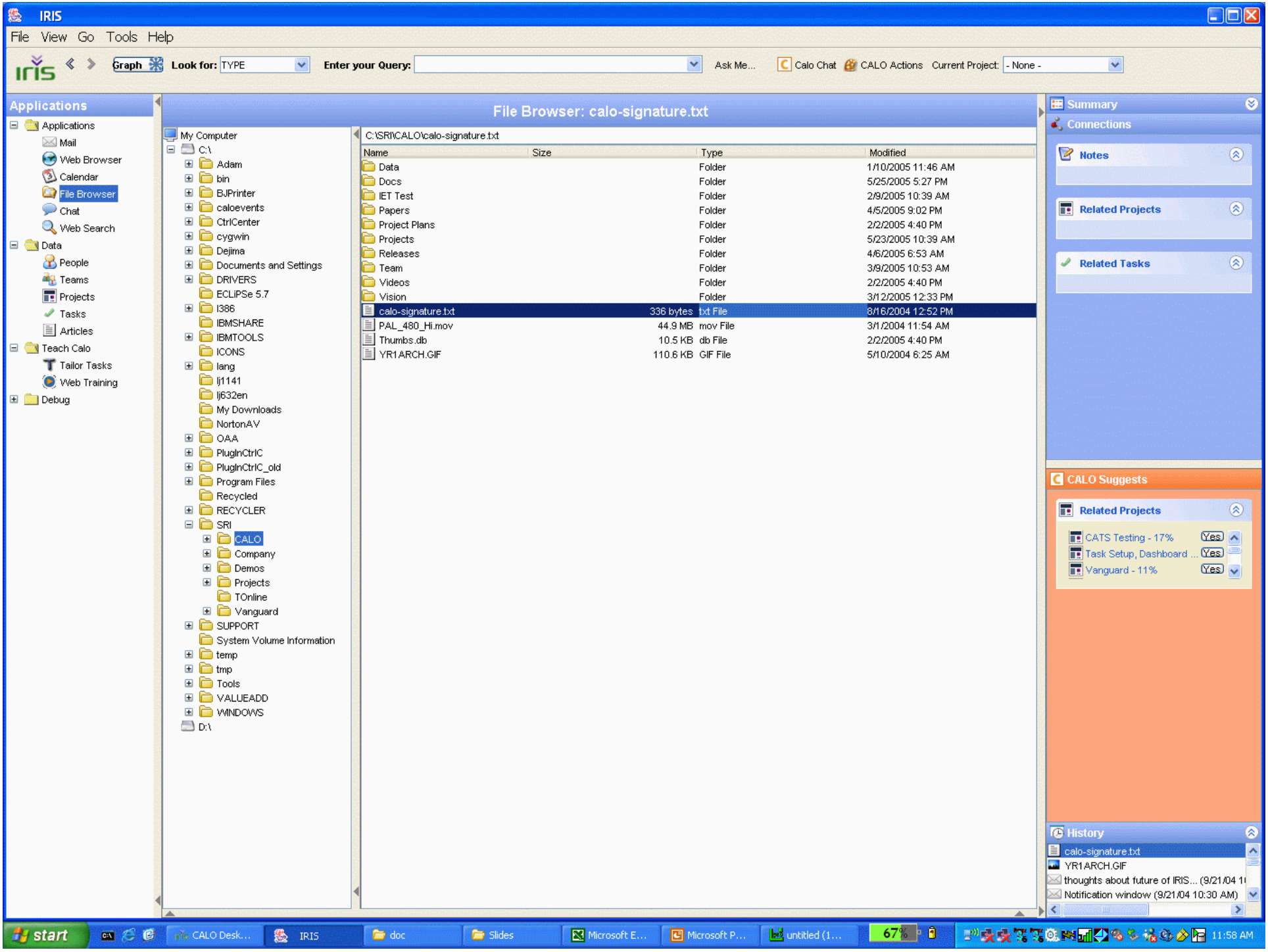
- File incoming information
- Prepare briefing folder before every meeting
- Alert user to time-critical emails
- Automate routine correspondence
- Help negotiate/schedule meetings
- Monitor pending action items per project
- …

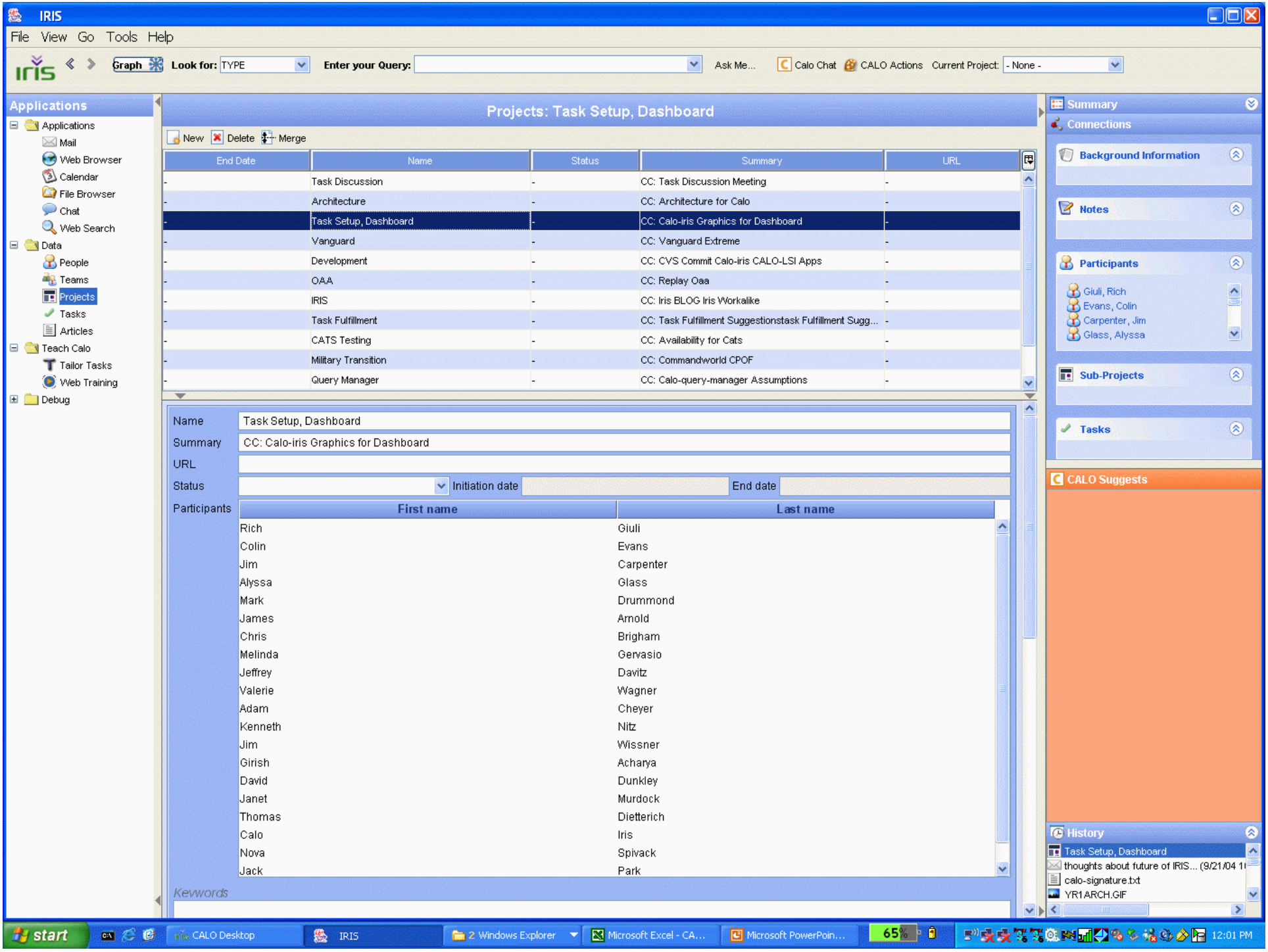
CALO Project Team

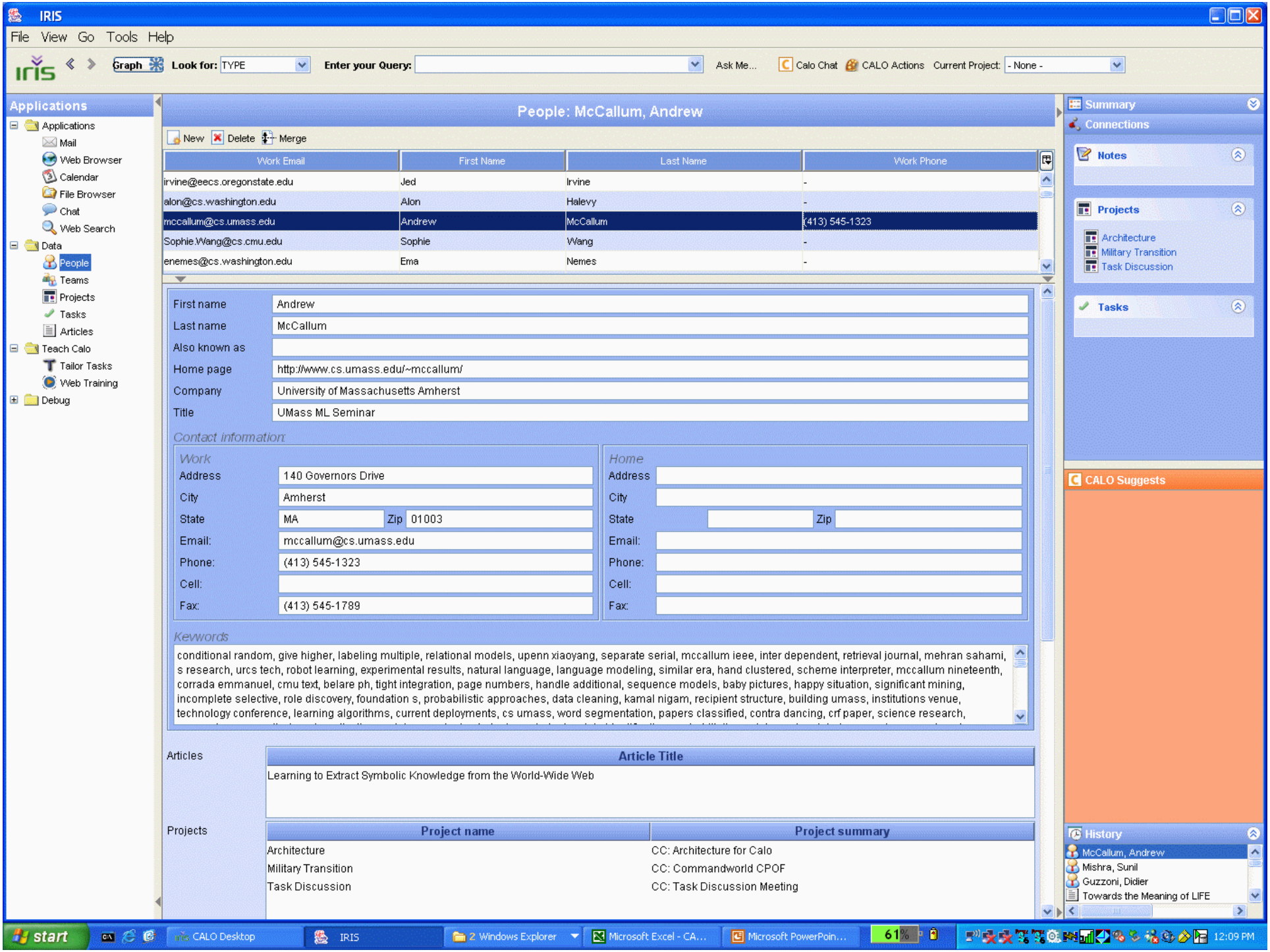


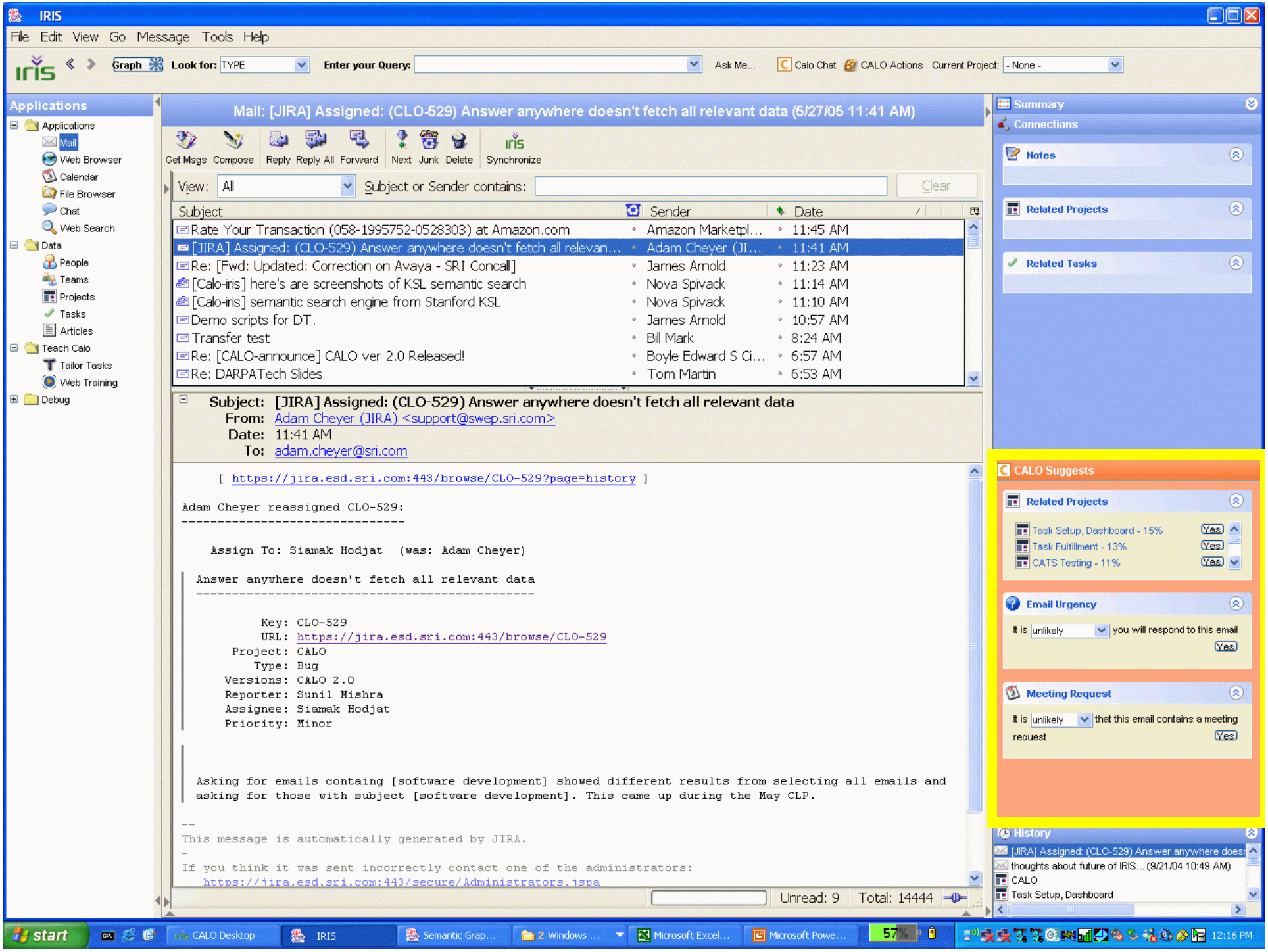

IRIS
File View Go Tools Help
Graph
Look for: TYPE
Enter your Query:
Ask Me...
Calo Chat
CALO Actions
Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
Projects: CALO
New
Delete
Merge
End Date
Name
Status
Summary
URL
Architecture
CC: Architecture for Calo
CALO
On schedule
Cognitive Assistant that Learns and Organizes
http://www.ai.sri.com/project...
CATS Testing
CC: Availability for Cats
Development
CC: CVS Commit Calo-iris CALO-LSI Apps
IRIS
CC: Iris BLOG Iris Workalike
Military Transition
CC: Commandworld CPOF
Name
Summary
URL
Status
Participants
Keywords
Semantic Graph Viewer
Military Transition
IRIS
Ontology
Development
Query Manager
Israel, David
Davitz, Jeffrey
Mark, Bill
Sub-Projects
Task Discussion
Cheyer, Adam
Architecture
Task Fulfillment
Participants
Task Setup, Dashboard
CALO 2.1 Release
Beers, Melissa
CALO 1.4 Release
CW #1
Arnold
Acharya, Girish
CALO
Tasks
CW #2
CW #5
CW #3
CW #4
Notes
Background Information
Summary
Connections
Background Information
Notes
Participants
Cheyer, Adam
Acharya, Girish
Israel, David
Davitz, Jeffrey
Sub-Projects
Architecture
CATS Testing
Development
IRIS
Tasks
CALO 1.4 Release
CALO 2.0 Release
CALO 2.1 Release
CW #1
CALO Suggests
History
CALO
Task Setup, Dashboard
McCallum, Andrew
Mishra, Sunil
start
CALO Desktop
IRIS
Semantic Grap...
2 Windows ...
Microsoft Excel...
Microsoft Powe...
59%
12:13 PM

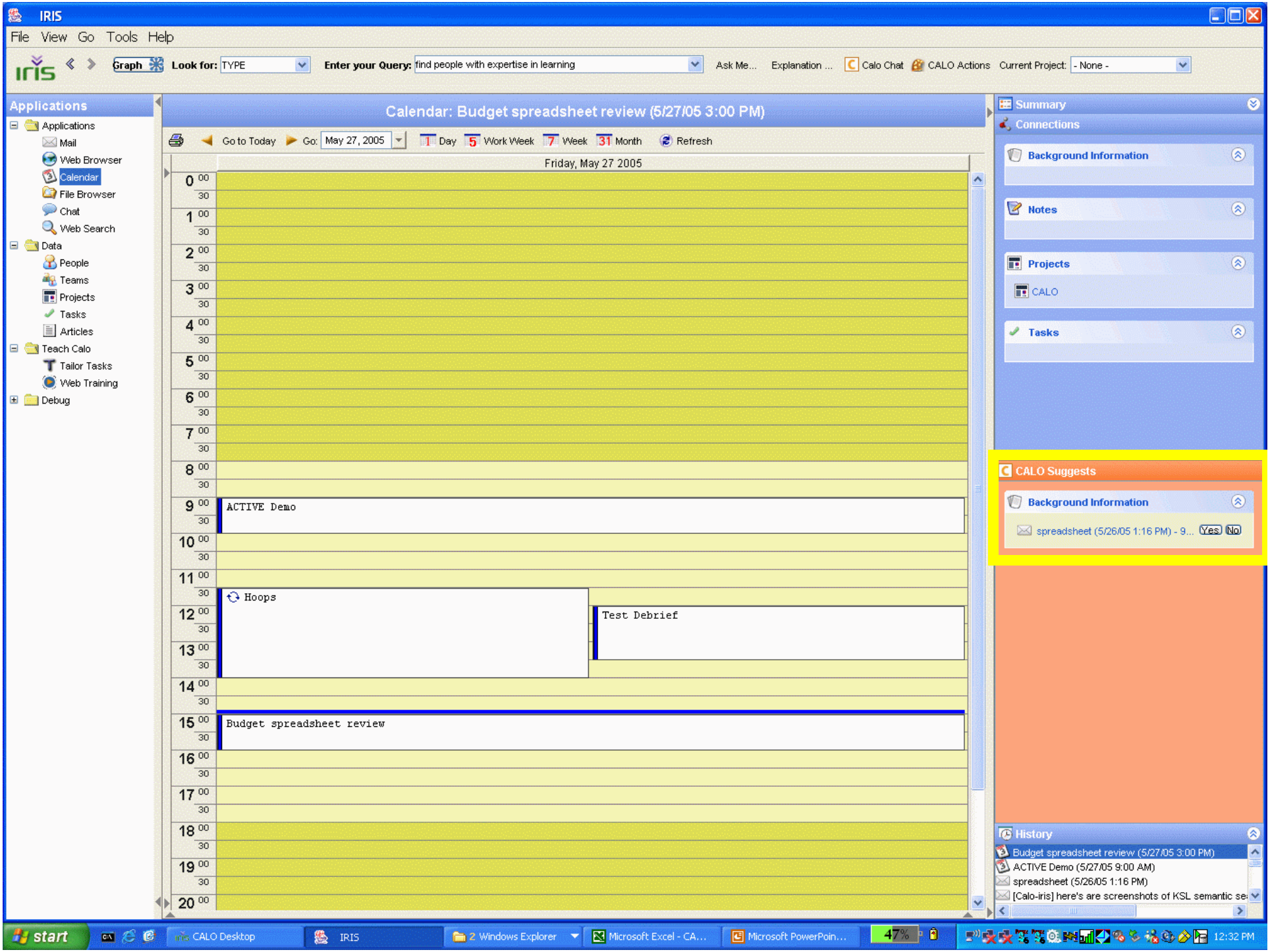

IRIS
File Edit View Go Message Tools Help
Graph Look for: TYPE Enter your Query: Ask Me... Calo Chat CALO Actions Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
Mail: thoughts about future of IRIS... (9/21/04 10:49 AM)
Get Msgs Compose Reply Reply All Forward Next Junk Delete Synchronize
View: All Subject or Sender contains: Clear
Subject Sender Date
Notification window placement Leslie Pound 9/21/2004 11:21 AM
thoughts about future of IRIS... Nova Spivack 9/21/2004 10:49 AM
Notification window Leslie Pound 9/21/2004 10:30 AM
Re: [Calo-iris] assertions Jim Carpenter 9/21/2004 10:18 AM
Re: Revised slides Scott E. Fahlman 9/21/2004 9:46 AM
Re: PTIME-UI Pauline Berry 9/21/2004 9:45 AM
Re: Meeting: DAI-Labor -- SRI agent discussion [3] Frank Steuer 9/21/2004 9:16 AM
Re: Spending a/o 9/18 James Arnold 9/21/2004 9:14 AM
Re: Meeting: DAI-Labor -- SRI agent discussion [3] James Arnold 9/21/2004 9:07 AM
Subject: thoughts about future of IRIS...
From: Nova Spivack <nova@radarnetworks.com>
Reply-To: nova@radarnetworks.com
Date: 9/21/2004 10:49 AM
To: 'Jim Wissner' <jim@radarnetworks.com>, 'Adam Cheyer' <adam.cheyer@sri.com>, 'Jack Park' <park@ai.sri.com>
If we want IRIS to become a *standard* and widely used platform for semantic web apps, we need to think about how outsider third-party developers may extend it.
Extensions could come in 3 flavors:
1. Not extending the IRIS apps, but extending the IRIS ontology
2. Extending the IRIS apps and extending the IRIS ontology
3. Extending the IRIS apps, but not extending the IRIS ontology
We should give some thought to whether/how to enable developers to conduct each of these types of activities on the IRIS platform.
One issue to think about is that the CALO ontology (the "Clib"), and at least some of the IRIS codebase, is focused primarily on reasoning and is designed to support agents that may or may not be released to the public as open-source along with the rest of IRIS. This means there may be a lot of complexity -- ontology classes and properties and Java code -- in IRIS that are specific to CALO agents but that are not relevant outside of CALO (e.g. will not be of use to third-party developers) -- unless we are planning ot release the CALO agents as open-source, or at least as closed-source plugins as well?
Another question: have we given any thought to how third-party developers would:
(a) Extend the ontology that comes with IRIS (and what are the terms of doing so? How can we prevent ontology "fragmentation" -- perhaps we need a way to enable third-party developers to extend the ontology within their own namespaces.
(b) Add their own vertical ontologies into IRIS for specific apps they create? Can IRIS load more than one ontology at a time?
Unread: 10 Total: 14444
Summary
Connections
Notes
Related Projects
Task Setup, Dashboard
Related Tasks
CALO Suggests
Email Urgency
It is unlikely you will respond to this email Yes
Meeting Request
It is unlikely that this email contains a meeting request Yes
History
thoughts about future of IRIS... (9/21/04 1
Notification window (9/21/04 10:30 AM)
DEX Update (5/23/05 1:30 PM)
Sold Time Report (5/23/05 12:18 PM)
start
CALO Desktop
IRIS
doc
Slides
Microsoft Excel...
Microsoft Powe...
68%
11:56 AM


IRIS
File View Go Tools Help
Graph
Look for: TYPE
Enter your Query:
Ask Me...
Calo Chat
CALO Actions
Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
File Browser: calo-signature.txt
My Computer
C:\
Adam
bin
BJPrinter
caloevents
CtrlCenter
cygwin
Dejima
Documents and Settings
DRIVERS
ECLiPSe 5.7
I386
IBMSHARE
IBMTOOLS
ICONS
lang
lj1141
lj632en
My Downloads
NortonAV
OAA
PlugInCtrlC
PlugInCtrlC_old
Program Files
Recycled
RECYCLER
SRI
CALO
Company
Demos
Projects
TOnline
Vanguard
SUPPORT
System Volume Information
temp
tmp
Tools
VALUEADD
WINDOWS
D:\
C:\SRI\CALO\calo-signature.txt
Name
Size
Type
Modified
Data
Folder
1/10/2005 11:46 AM
Docs
Folder
5/25/2005 5:27 PM
IET Test
Folder
2/9/2005 10:39 AM
Papers
Folder
4/5/2005 9:02 PM
Project Plans
Folder
2/2/2005 4:40 PM
Projects
Folder
5/23/2005 10:39 AM
Releases
Folder
4/6/2005 6:53 AM
Team
Folder
3/9/2005 10:53 AM
Videos
Folder
2/2/2005 4:40 PM
Vision
Folder
3/12/2005 12:33 PM
calo-signature.txt
336 bytes
txt File
8/16/2004 12:52 PM
PAL_480_Hi.mov
44.9 MB
mov File
3/1/2004 11:54 AM
Thumbs.db
10.5 KB
db File
2/2/2005 4:40 PM
YR1ARCH.GIF
110.6 KB
GIF File
5/10/2004 6:25 AM
Summary
Connections
Notes
Related Projects
Related Tasks
CALO Suggests
Related Projects
CATS Testing - 17%
Yes
Task Setup, Dashboard ...
Yes
Vanguard - 11%
Yes
History
calo-signature.txt
YR1ARCH.GIF
thoughts about future of IRIS... (9/21/04 1
Notification window (9/21/04 10:30 AM)
start
CALO Desk...
IRIS
doc
Slides
Microsoft E...
Microsoft P...
untitled (1...
67%
11:58 AM


IRIS
File View Go Tools Help
Graph
Look for: TYPE
Enter your Query:
Ask Me...
Calo Chat
CALO Actions
Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
Projects: Task Setup, Dashboard
New
Delete
Merge
End Date
Name
Status
Summary
URL
Task Discussion
CC: Task Discussion Meeting
Architecture
CC: Architecture for Calo
Task Setup, Dashboard
CC: Calo-iris Graphics for Dashboard
Vanguard
CC: Vanguard Extreme
Development
CC: CVS Commit Calo-iris CALO-LSI Apps
OAA
CC: Replay Oaa
IRIS
CC: Iris BLOG Iris Workalike
Task Fulfillment
CC: Task Fulfillment Suggestionstask Fulfillment Sugg...
CATS Testing
CC: Availability for Cats
Military Transition
CC: Commandworld CPOF
Query Manager
CC: Calo-query-manager Assumptions
Name
Task Setup, Dashboard
Summary
CC: Calo-iris Graphics for Dashboard
URL
Status
Initiation date
End date
Participants
First name
Last name
Rich Giuli
Colin Evans
Jim Carpenter
Alyssa Glass
Mark Drummond
James Arnold
Chris Brigham
Melinda Gervasio
Jeffrey Davitz
Valerie Wagner
Adam Cheyer
Kenneth Nitz
Jim Wissner
Girish Acharya
David Dunkley
Janet Murdock
Thomas Dietterich
Calo Iris
Nova Spivack
Jack Park
Keywords
Summary
Connections
Background Information
Notes
Participants
Giuli, Rich
Evans, Colin
Carpenter, Jim
Glass, Alyssa
Sub-Projects
Tasks
CALO Suggests
History
Task Setup, Dashboard
thoughts about future of IRIS... (9/21/04 1
calo-signature.txt
YR1ARCH.GIF
start
CALO Desktop
IRIS
2 Windows Explorer
Microsoft Excel - CA...
Microsoft PowerPoin...
65%
12:01 PM


IRIS
File View Go Tools Help
Graph
Look for: TYPE
Enter your Query:
Ask Me...
Calo Chat
CALO Actions
Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
People: McCallum, Andrew
New
Delete
Merge
Work Email
First Name
Last Name
Work Phone
irvine@eecs.oregonstate.edu
Jed
Irvine
-
alon@cs.washington.edu
Alon
Halevy
-
mccallum@cs.umass.edu
Andrew
McCallum
(413) 545-1323
Sophie.Wang@cs.cmu.edu
Sophie
Wang
-
enemes@cs.washington.edu
Ema
Nemes
-
First name
Andrew
Last name
McCallum
Also known as
Home page
http://www.cs.umass.edu/~mccallum/
Company
University of Massachusetts Amherst
Title
UMass ML Seminar
Contact information:
Work
Address
140 Governors Drive
City
Amherst
State
MA
Zip
01003
Email:
mccallum@cs.umass.edu
Phone:
(413) 545-1323
Cell:
Fax:
(413) 545-1789
Home
Address
City
State
Zip
Email:
Phone:
Cell:
Fax:
Keywords
conditional random, give higher, labeling multiple, relational models, upenn xiaoyang, separate serial, mccallum ieee, inter dependent, retrieval journal, mehran sahami, s research, urcs tech, robot learning, experimental results, natural language, language modeling, similar era, hand clustered, scheme interpreter, mccallum nineteenth, corrada emmanuel, cmu text, belare ph, tight integration, page numbers, handle additional, sequence models, baby pictures, happy situation, significant mining, incomplete selective, role discovery, foundation s, probabilistic approaches, data cleaning, kamal nigam, recipient structure, building umass, institutions venue, technology conference, learning algorithms, current deployments, cs umass, word segmentation, papers classified, contra dancing, crf paper, science research,
Articles
Article Title
Learning to Extract Symbolic Knowledge from the World-Wide Web
Projects
Project name
Project summary
Architecture
CC: Architecture for Calo
Military Transition
CC: Commandworld CPOF
Task Discussion
CC: Task Discussion Meeting
Summary
Connections
Notes
Projects
Architecture
Military Transition
Task Discussion
Tasks
CALO Suggests
History
McCallum, Andrew
Mishra, Sunil
Guzzoni, Didier
Towards the Meaning of LIFE
start
CALO Desktop
IRIS
2 Windows Explorer
Microsoft Excel - CA...
Microsoft PowerPoin...
61%
12:09 PM


IRIS
File Edit View Go Message Tools Help
Graph Look for: TYPE Enter your Query: Ask Me... Calo Chat CALO Actions Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
Mail: [JIRA] Assigned: (CLO-529) Answer anywhere doesn't fetch all relevant data (5/27/05 11:41 AM)
Get Msgs Compose Reply Reply All Forward Next Junk Delete Synchronize
View: All Subject or Sender contains: Clear
Subject Sender Date
Rate Your Transaction (058-1995752-0528303) at Amazon.com Amazon Marketpl... 11:45 AM
[JIRA] Assigned: (CLO-529) Answer anywhere doesn't fetch all relevan... Adam Cheyer (JI... 11:41 AM
Re: [Fwd: Updated: Correction on Avaya - SRI Concall] James Arnold 11:23 AM
[Calo-iris] here's are screenshots of KSL semantic search Nova Spivack 11:14 AM
[Calo-iris] semantic search engine from Stanford KSL Nova Spivack 11:10 AM
Demo scripts for DT. James Arnold 10:57 AM
Transfer test Bill Mark 8:24 AM
Re: [CALO-announce] CALO ver 2.0 Released! Boyle Edward S Ci... 6:57 AM
Re: DARPATech Slides Tom Martin 6:53 AM
Subject: [JIRA] Assigned: (CLO-529) Answer anywhere doesn't fetch all relevant data
From: Adam Cheyer (JIRA) <support@swep.sri.com>
Date: 11:41 AM
To: adam.cheyer@sri.com
[https://jira.esd.sri.com:443/browse/CLO-529?page=history]
Adam Cheyer reassigned CLO-529:
Assign To: Siamak Hodjat (was: Adam Cheyer)
Answer anywhere doesn't fetch all relevant data
Key: CLO-529
URL: https://jira.esd.sri.com:443/browse/CLO-529
Project: CALO
Type: Bug
Versions: CALO 2.0
Reporter: Sunil Mishra
Assignee: Siamak Hodjat
Priority: Minor
Asking for emails containg [software development] showed different results from selecting all emails and asking for those with subject [software development]. This came up during the May CLP.
This message is automatically generated by JIRA.
If you think it was sent incorrectly contact one of the administrators:
https://jira.esd.sri.com:443/secure/Administrators.jspa
Unread: 9 Total: 14444
Summary
Connections
Notes
Related Projects
Related Tasks
CALO Suggests
Related Projects
Task Setup, Dashboard - 15% Yes
Task Fulfillment - 13% Yes
CATS Testing - 11% Yes
Email Urgency
It is unlikely you will respond to this email Yes
Meeting Request
It is unlikely that this email contains a meeting request Yes
History
[JIRA] Assigned: (CLO-529) Answer anywhere doesn
thoughts about future of IRIS... (9/21/04 10:49 AM)
CALO
Task Setup, Dashboard
start
CALO Desktop
IRIS
Semantic Grap...
2 Windows ...
Microsoft Excel...
Microsoft Powe...
57%
12:16 PM


IRIS
File View Go Tools Help
Graph
Look for: TYPE
Enter your Query: find people with expertise in learning
Ask Me...
Explanation ...
Calo Chat
CALO Actions
Current Project: - None -
Applications
Applications
Mail
Web Browser
Calendar
File Browser
Chat
Web Search
Data
People
Teams
Projects
Tasks
Articles
Teach Calo
Tailor Tasks
Web Training
Debug
Calendar: Budget spreadsheet review (5/27/05 3:00 PM)
Go to Today
Go: May 27, 2005
Day
Work Week
Week
Month
Refresh
Friday, May 27 2005
ACTIVE Demo
Hoops
Test Debrief
Budget spreadsheet review
Summary
Connections
Background Information
Notes
Projects
CALO
Tasks
CALO Suggests
Background Information
spreadsheet (5/26/05 1:16 PM) - 9...
Yes
No
History
Budget spreadsheet review (5/27/05 3:00 PM)
ACTIVE Demo (5/27/05 9:00 AM)
spreadsheet (5/26/05 1:16 PM)
[Calo-iris] here's are screenshots of KSL semantic se
start
CALO Desktop
IRIS
2 Windows Explorer
Microsoft Excel - CA...
Microsoft PowerPoin...
47%
12:32 PM

CALO Target Functionality

- Organize & Manage Information
  - Manage email, documents, web info
  - Organize information by tasks and user activities
- Prepare Information Products
  - Prepare meeting, event info packages
  - Organize and assemble reports, summaries
- Observe & Mediate Interactions
  - Monitor meetings, email threads, chat
  - Record meeting discussion, events, actions
- Monitor & Manage Tasks
  - Organize and monitor task execution
  - Monitor due dates, perform time management
- Schedule & Organize in Time
  - Schedule meetings, events, tasks
  - Organize task dependencies, preconditions
- Acquire, Allocate Resources
  - Locate, acquire, allocate resources (equipment, facilities, people) in response to needs

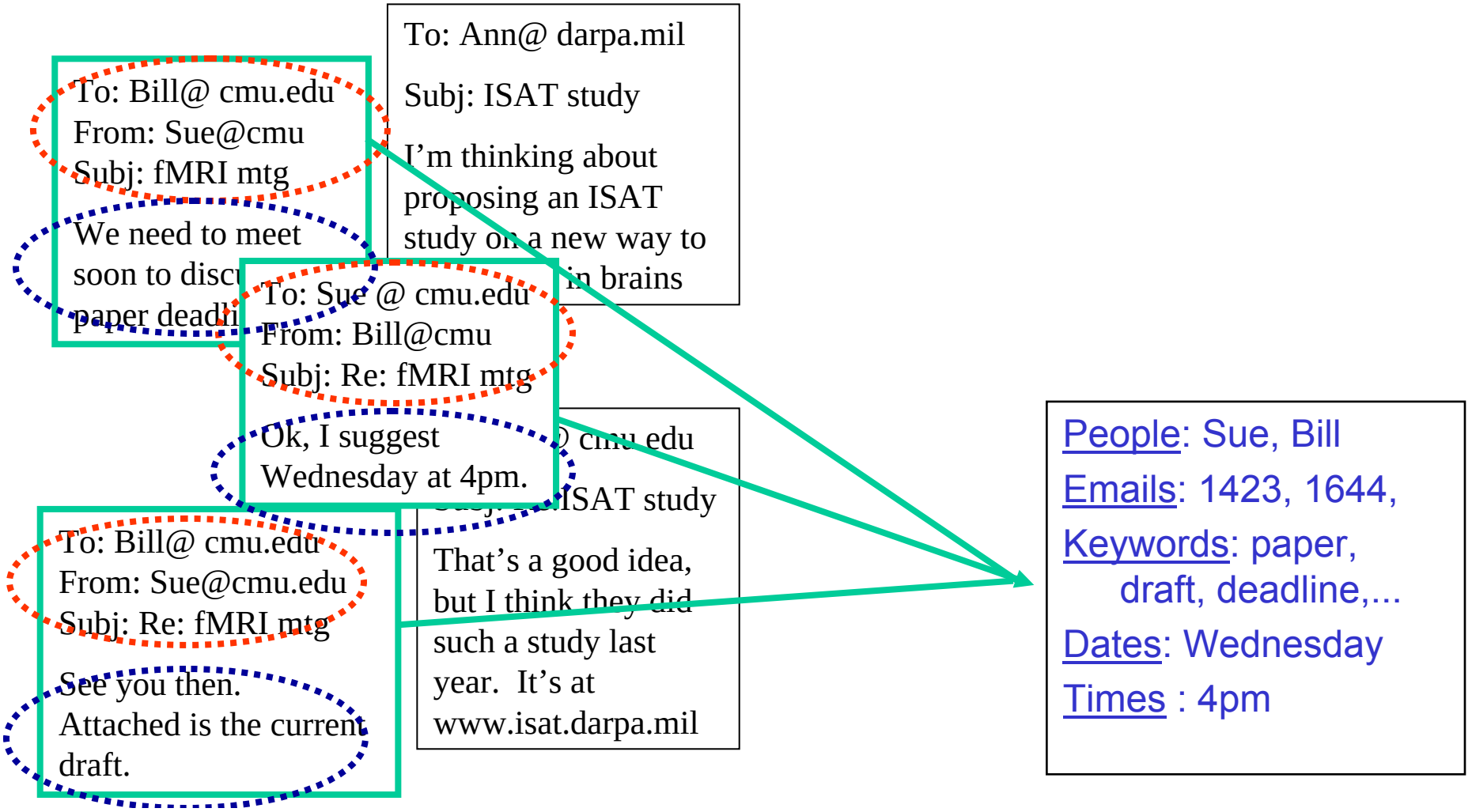
1. Can we extract structured descriptions of ongoing projects from user's email?

Approach

[with Y. Huang, S. Wang]

1. Cluster emails by headers
2. Cluster emails by body

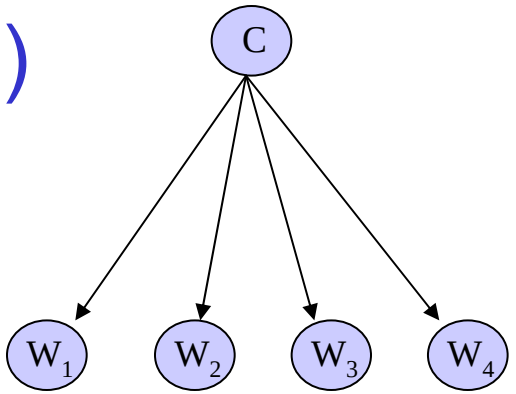
↔ Use both for better performance



Unsupervised Learning of Projects

1. Cluster emails
   - (Headers) Initialize clusters bottom-up
     - group emails with similar subject lines, then select initial groups with greatest TFIDF distance
   - (Body) Refine clusters by applying EM algorithm,
     - Represent email by bag of words in subject and body
   - (Social network) Subdivide each cluster based on graph of email co-recipients
     - Make each clique of co-recipients a subcluster
2. For each cluster, extract information from the email text and headers

Naïve Bayes classifier (supervised)



Train:

For each class c_j of documents

1. Estimate $P(c_j)$

2. For each word w_i estimate $P(w_i \mid c_j)$

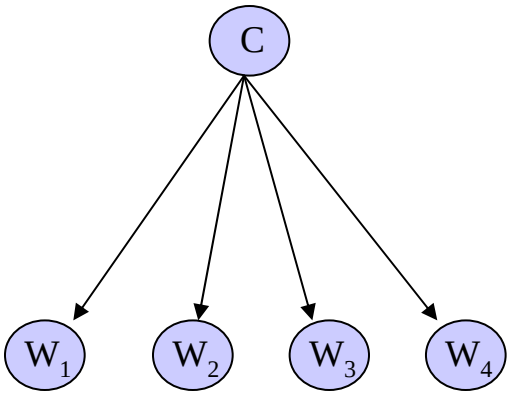
Classify (doc):

Assign *doc* to most probable* class

$$\arg\max_j P(c_j) \prod_{w_i \in doc} P(w_i \mid c_j)$$

* assuming words are conditionally independent, given class

EM for Text Clustering



- Like supervised learning, but no training labels

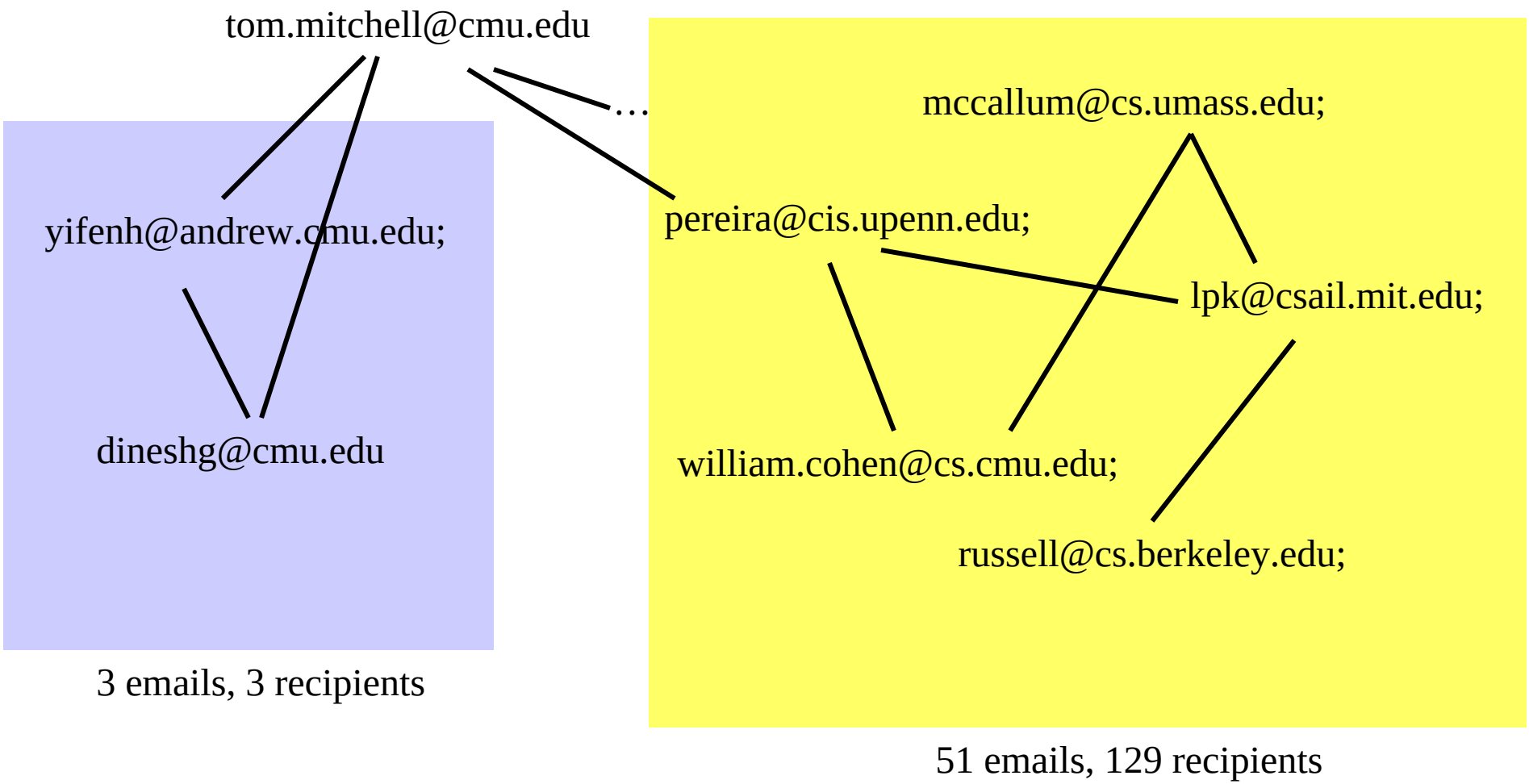
EM:

- Initialize labels, then train Naïve Bayes classifier
- Repeat until convergence
  - E: use current classifier to assign probabilistic labels
  - M: retrain Naïve Bayes classifier using these probabilistic labels

Results: Discovered topic words and User labels
(from 1481 emails during one month)

- CS faculty discussions (158)
  - Faculty, SCS, Qatar, LTI, wishlist
- Recruiting in AI (137)
  - CSD, host, hiring, RI, interviews
- Research1 (93)
  - RADAR, fluid, Siemens, lead, Bloomberg, IDA, letter, parkway
- Research2 (105)
  - CALO, TFC, SRI, examples, heads, labeled, Leslie, HMM
- CALD management (54)
  - SAS, software, color, GSIA, ATT, license, consulting, CALD, NYU
- Family and friends (63)
  - Joan, Paris, Petra, wet, water, towels, restaurant, night, weekend
- Professional conversations (299)
  - UnitedTechnologies, Diane, Howard, house, Dewey, research
- Professional organizations (45)
  - PASCAL, Hamilton, Southampton, Amari, Shawe, network, Carol
- AAAI Fellows committee (38)
  - AAAI, fellows, selection, committee, nominations, Hamilton, Carol
- Writing activities (109)
  - Vitor, Melissa, nouns, noun, paper, Beers, Vincent, verbs, classification
- Computer facilities support (59)
  - DVD, root, purchase, upgrade, Natural, bookkeeping, hardware, toolbox
- Seminar announcements (123)
  - CNBC, cognitive, fMRI, code, activation, brain, data, theory, model
- Research props, grant reporting (148)
  - Lecture, news, miles, Keck, grade, hall
- Health-related seminars/RFP's (6)
  - NIH, RFA, grants, pathways, scientific
- Recommendation letters (11)

Email co-recipient subgraph cliques (cluster 4)



Cluster 4: 105 emails total.

21 subcliques, containing from 1 to 51 emails

Activity Clustering Algorithm

1. Cluster emails
   - (Headers) Initialize clusters bottom-up
     - group emails with similar subject lines, then select initial groups with greatest TFIDF distance
   - (Body) Refine clusters by applying EM algorithm to all emails,
     - Represent email by bag of words in subject and body
   - (Social network) Subdivide each cluster based on graph of email co-recipients
     - Make each clique of co-recipients a subcluster
2. For each cluster, extract information from the email bodies

Example: Learned Project Description

ActivityCluster4.1 (51 emails)

- *Keywords*: TFC, heads, CALO, Leslie, estimation, capabilities, baseline, capture, DarpaTech, calendar, SRI, goals, HMM, extraction
- *PrimarySenders:* Mitchell(21), McCallum(6), Leslie(5), Cohen(3),...
- *UserActivityFraction*: 71/1210 of total email (0.058 of total)
- *IntensityOfUserInvolvement:* created 29% of traffic; (default 31%)
- *ExtractedNames:* Tom (64), Leslie(11), Andrew (11), Dave (9), …
- *ExtractedDates:* 2004(32), today(12), tomorrow(11), Wednesday(10), February 18(9), Tuesday(8), Monday (8)
- *ExtractedTimes:* 5(4), 11:30am(3), 5pm(2), morning(2), 2PM(2), 4pm(2), about 2PM(2), 9:15 PM(1), 4:30-6 p.m.(1), this morning (1)...
- *RequestEmails:* <emailA>, <emailB>, …

Example: Learned Activity Fra… email corpus

ActivityCluster4.1 (51 emails, from initial clus…

• *Keywords*: TFC, heads, CALO, Leslie, estimation, baseline, capture, DarpaTech, calendar, SRI, goal…

• *PrimarySenders:* Mitchell(21), McCallum(6), Lesl…

• *UserActivityFraction*: 71/1210 of total email (0.05…

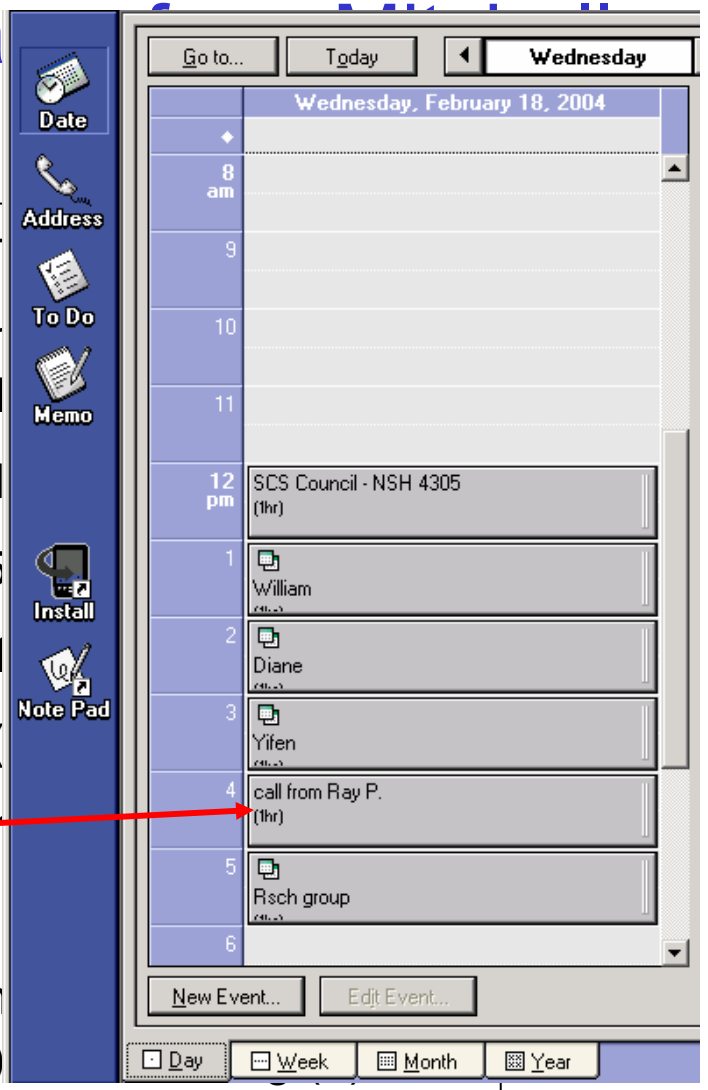
• *IntensityOfUserInvolvement:* created 29% of traff…

• *ExtractedNames:* Tom (64), Leslie(11), Andrew (…

• *ExtractedDates:* 2004(32), today(12), tomorrow(… February 18(9), Tuesday(8), Monday (8)

•*ExtractedTimes:* 5(4), 11:30am(3), 5pm(2), morn… 4pm(2), about 2PM(2), 9:15 PM(1), 4:30-6 p.m.(1)

• *RequestEmails:* <emailA>, <emailB>, …



Example: Learned Activity Frame from Mitchell email corpus

ActivityCluster4.1 (51 emails, from initial cluster containing 105)

• *Keywords*: TFC, heads, CALO, Leslie, estimation, capabilities, baseline, capture, DarpaTech, calendar, SRI, goals, HMM, extraction

• *PrimarySenders:* Mitchell(21), McCallum(6), Leslie(5), Cohen(3),

• *UserActivityFraction*: 71/1210 c

• *IntensityOfUserInvolvement:* cr

• *ExtractedNames:* Tom (64), Le

• *ExtractedDates:* 2004(32), toda
February 18(9), Tuesday(8), Mo

•*ExtractedTimes:* 5(4), 11:30am
4pm(2), about 2PM(2), 9:15 PM(

• *RequestEmails:* <emailA>, <emailB>, …

I need to get to DARPA by COB tomorrow a list of CALO participants who need access to the IPTO booth. It seems to me we should ask for this for any of you who is likely to be there. Could you let me know asap if you *might* be there? No big deal if you end up not going.

THanks, --r

1. Can we extract structured descriptions of ongoing projects from user's email?

- Yes! (though plenty of room to improve)

Accuracy is imperfect (e.g., in TM's "CALO" cluster, only ~half the emails were truly relevant)

Nevertheless, the project descriptions are often useful, because they describe aggregate statistics

headers,
body,
attachments

times, dates,
people,
places

file content,
subdirectories,
creator,
last edit date

page content,
hyperlinks,
host site structure

Email

To: Bill@ cmu.edu

Subj: fMRI meeting

We need to meet soon to discuss

To: Sue @ cmu.edu

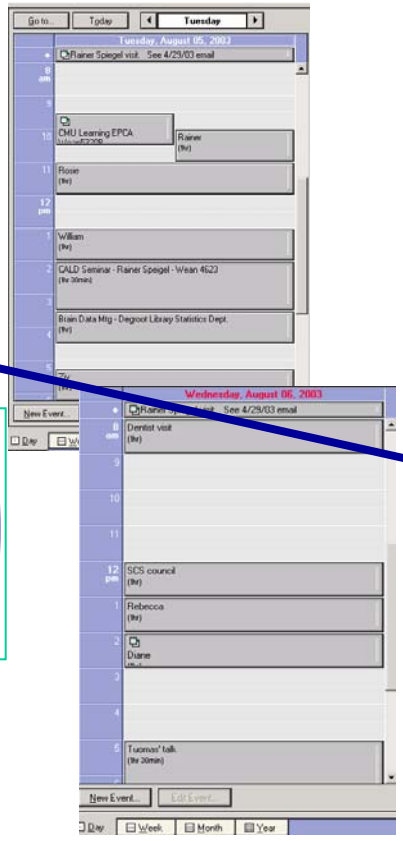
Subj: Re: fMRI meeting

To: Bill@ cmu.edu

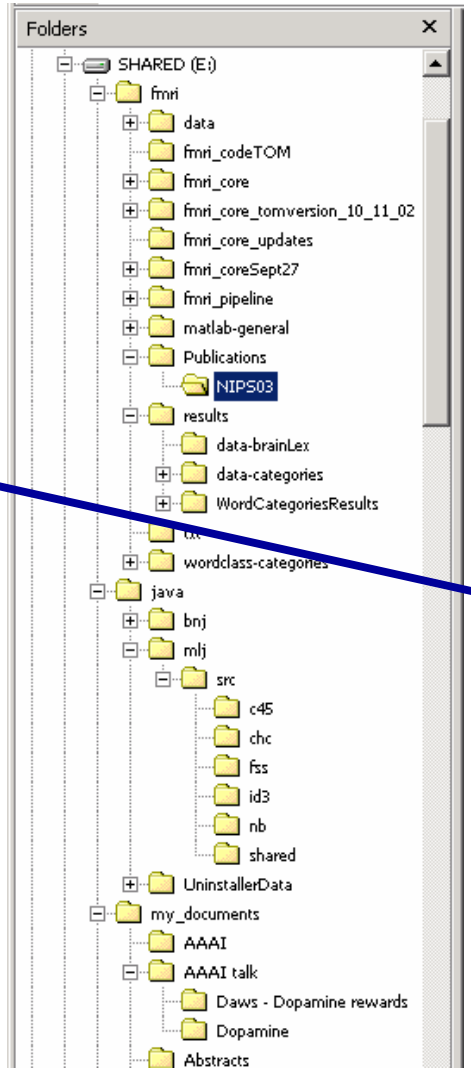
Subj: Re: fMRI meeting

See you then. Attached is the current draft.

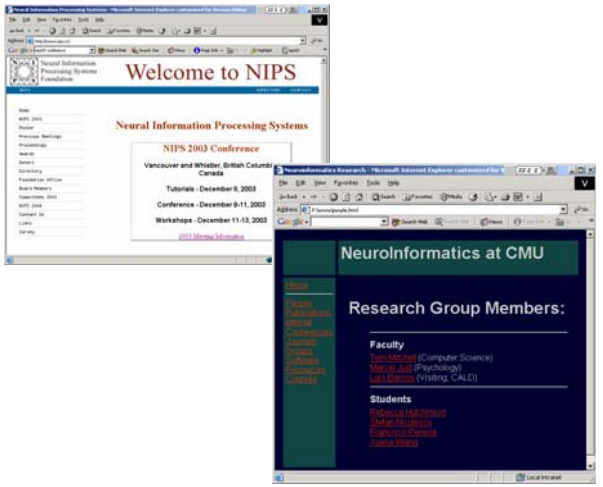
Calendar



Directories



Web Activity



fMRI paper writing

People: Sue, Bill

Document: <fileptr>

Meetings: Aug 24,

Emails: 1423, 1644,

Leader: Bill

Deadline: Jan 15

Vision

headers,
body,
attachments

times, dates,
people,
places

file content,
subdirectories,
creator,
last edit date

page content,
hyperlinks,
host site structure

Email

To: Bill@ cmu.edu

Subj: fMRI meeting

We need to meet soon to discuss

To: Sue @ cmu.edu

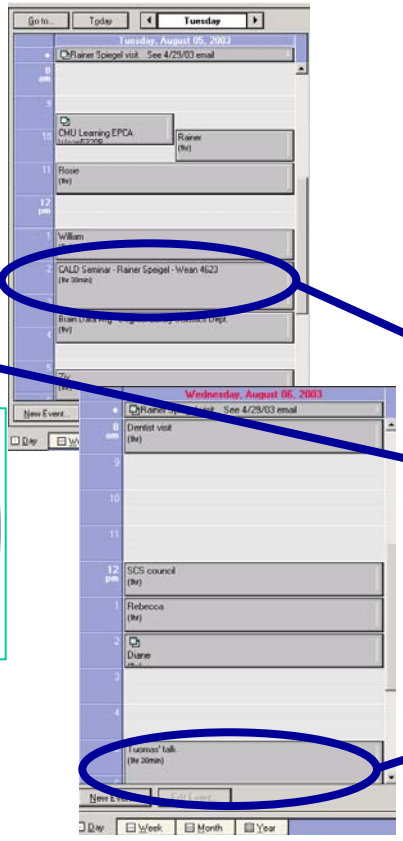
Subj: Re: fMRI meeting

To: Bill@ cmu.edu

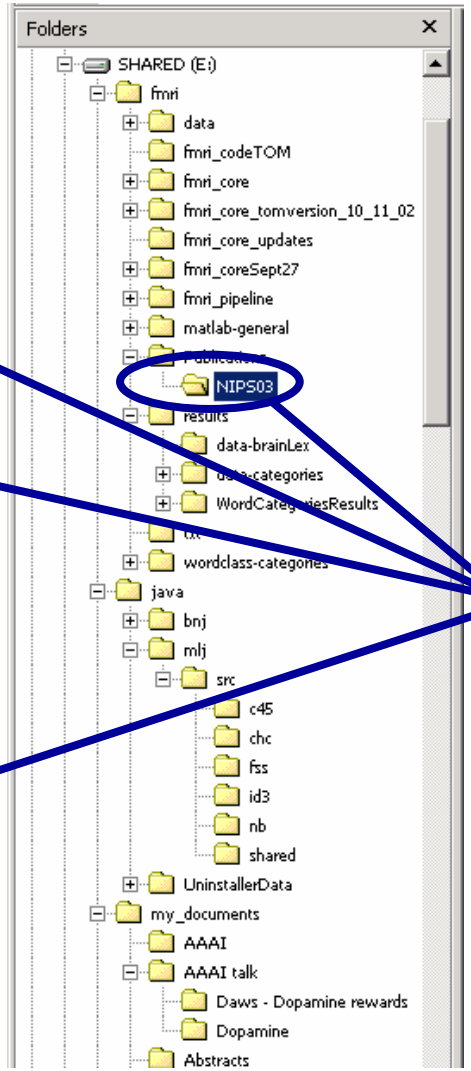
Subj: Re: fMRI meeting

See you then. Attached is the current draft.

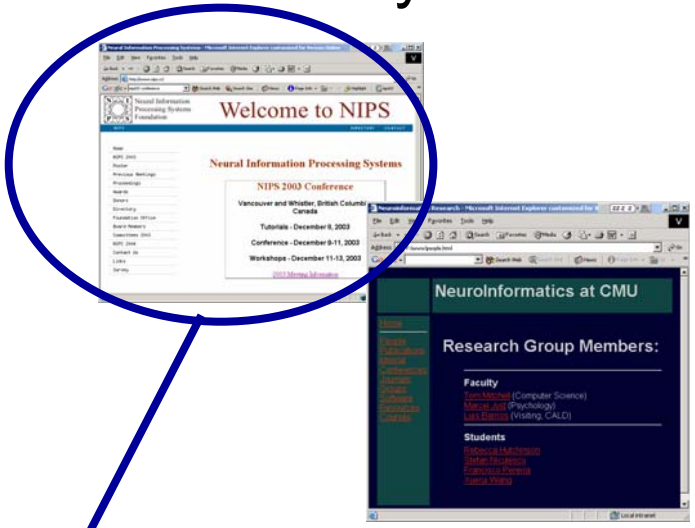
Calendar



Directories



Web Activity



fMRI paper writing

People: Sue, Bill

Document: <fileptr>

Meetings: Aug 24,

Emails: 1423, 1644,

Leader: Bill

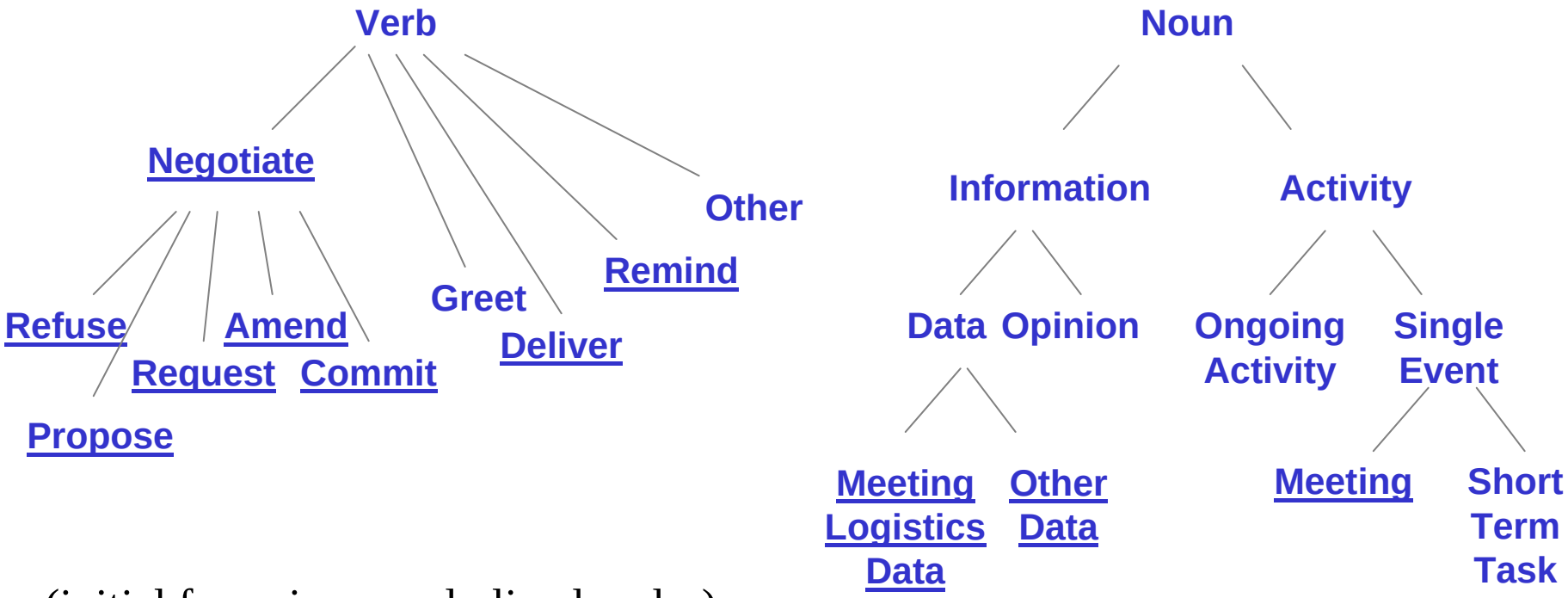
Deadline: Jan 15

2. How can we classify email according to the senders intent?

Emails as noun-verb “speech acts”

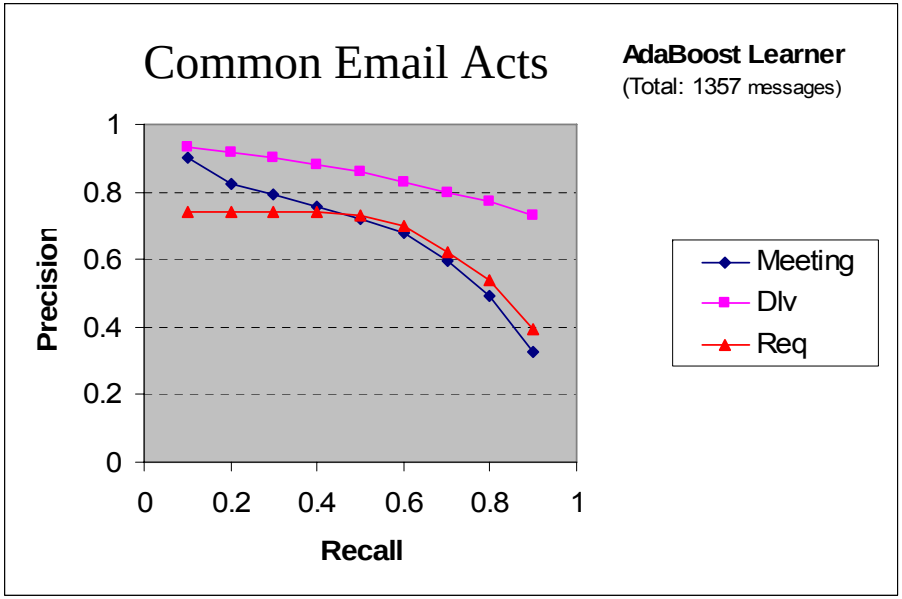
[Cohen, et al. 2004]

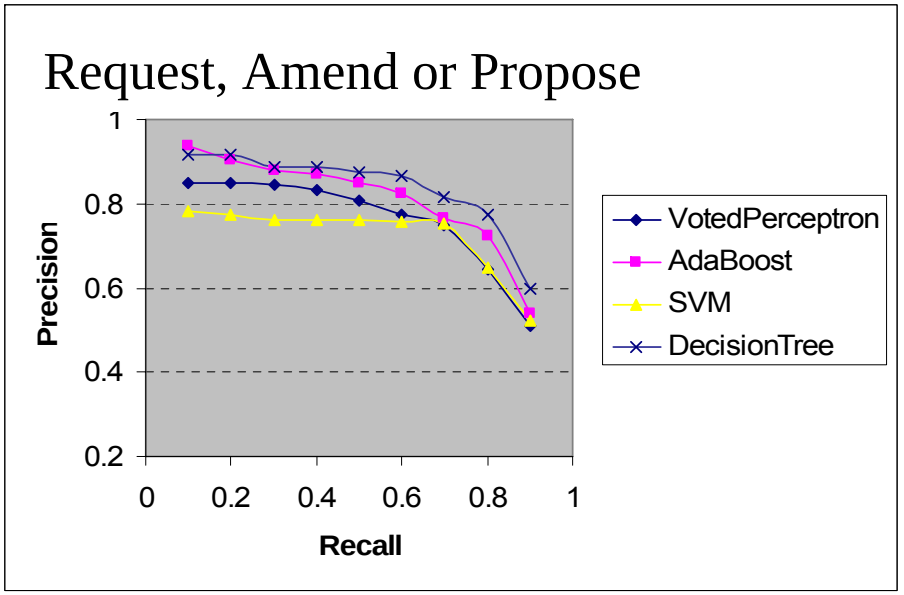
Learn to classify email sender’s *intent*, according to a taxonomy of “email speech acts” defined by <verb,noun> (e.g., “Request meeting,” “Deliver data”)



(initial focus is on underlined nodes)

Accuracies for Best-Learned Email Acts





Classification Errors and F1 Scores

VP: voted perceptron

AB: AdaBoost

SVM: linear Support Vector Machine

DT: decision tree

‘request’ ‘amend’ or ‘propose’

‘commit’ or ‘deliver’

Act		VP	AB	SVM	DT
Request (450/907)	*Error*	0.25	0.22	0.23	**0.20**
	F1	0.58	0.65	0.64	**0.69**
Proposal (140/1217)	*Error*	0.11	0.12	**0.12**	0.10
	F1	0.19	0.26	**0.44**	0.13
Delivery (873/484)	*Error*	**0.26**	0.28	0.27	0.30
	F1	**0.80**	0.78	0.78	0.76
Commit-ment (208/1149)	*Error*	0.15	0.14	**0.17**	0.15
	F1	0.21	0.44	**0.47**	0.11
Directive (605/752)	*Error*	0.25	0.23	0.23	**0.19**
	F1	0.72	0.73	0.73	**0.78**
Commis-sive (993/364)	*Error*	0.23	0.23	0.24	**0.22**
	F1	0.84	0.84	0.83	**0.85**
Meet (345/1012)	*Error*	0.187	0.17	**0.14**	0.18
	F1	0.573	0.62	**0.72**	0.60

Speech Acts and Factored Classification

Standard classification: learn f: X → Y

Factored classification: learn f: X → Y1 x Y2

e.g., f: email → Noun x Verb

Our initial approach:

- Learn f_Y: email → Y, for each $Y \in \{N \cup V\}$
- i.e., treat as |N|+|V| independent classifiers

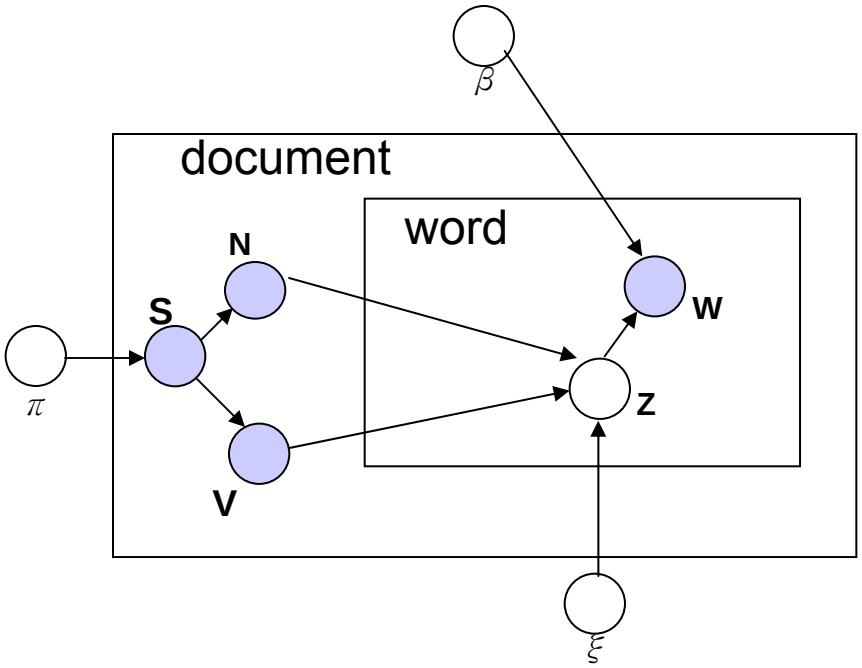
Idea:

- Assume Noun and Verb are not independent, and each email word is generated either by the Noun or Verb

Emails Words Generated by SpeechAct Nouns and Verbs

Current work, Qiong Chen, Yifen Huang

Motivation: (1) more accurate classification, (2) localize relevant text segments



1. Each document has speech act S, which specifies a noun N, verb V
2. Each word W in a document, is generated either by its N or V.
3. The hidden variable Z takes on either the value of N or value of V, for each word W,
4. Word W is generated by P(W|Z).

$P(W_i|N,V,\theta) = P(W_i|Z_i)P(Z_i|N,V)$

4. Parameters can be estimated using an EM algorithm.

Parameters to be learned:

β_{wx} = p(W=w |Z=x)

α_{nv} = p(X=n | N=n,V=v)

π_{nv} = p(S=<n,v>)

Variant of LDA [Blei, Ng, Jordan, 2003]

Preliminary results:
Words associated with Commit vs. Meeting

Speech act = <commit, meeting>

Subject: Re: Monday's ***meeting***

commit

meeting

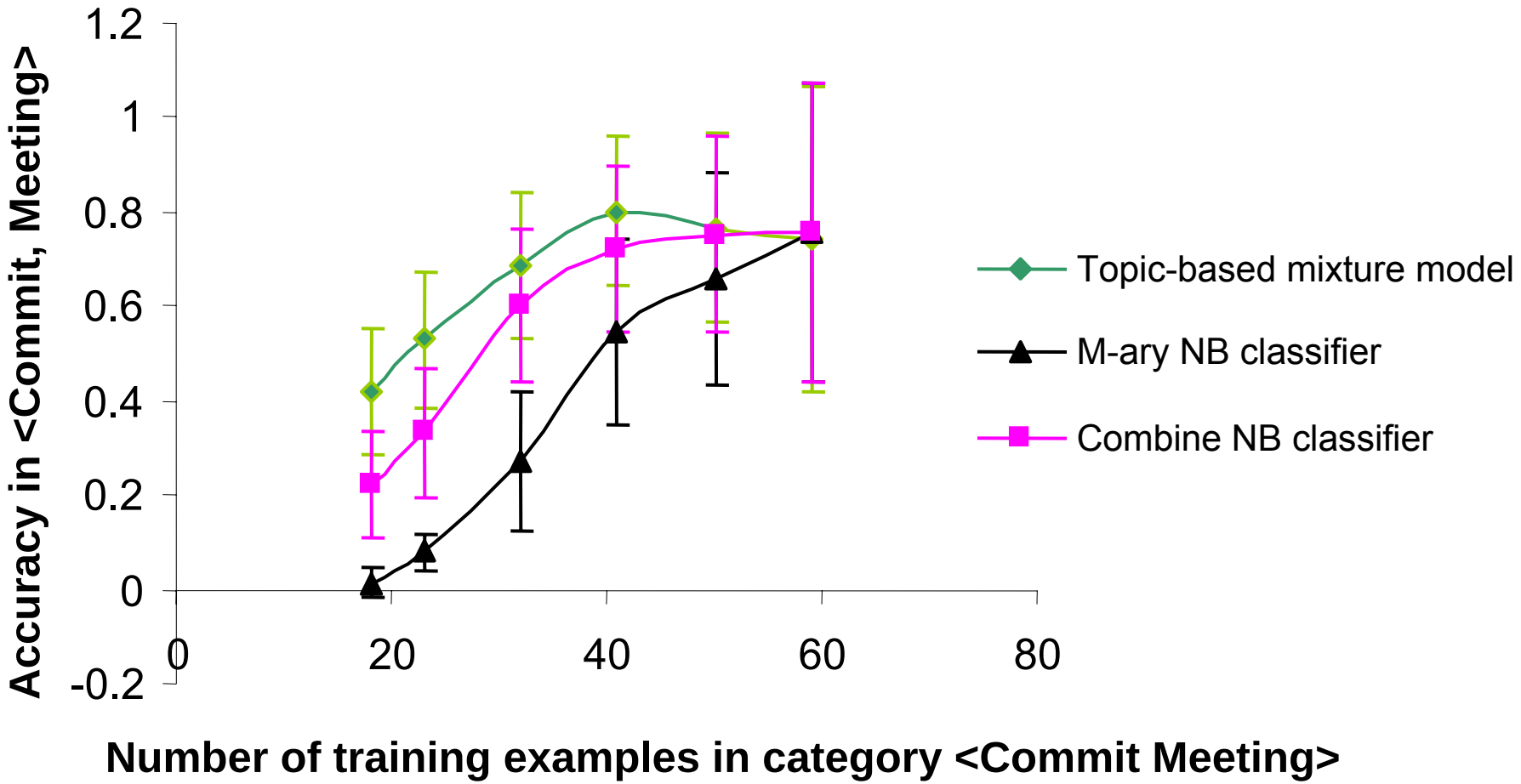
Hi , ***everyone***

Monday ***at*** 10 : ***30*** am is okay with me. ***Let*** 's just plan ***on*** ***meeting*** in ***the*** commons ***at*** 10: ***30***. Like I ***mentioned*** previously, ***we*** should only take about 45 minutes just to decide what each of ***us*** ***plans*** to accomplish ***over*** the summer.

Good luck ***on*** all of you finals.

Supervised learning of Noun-Verb topic models

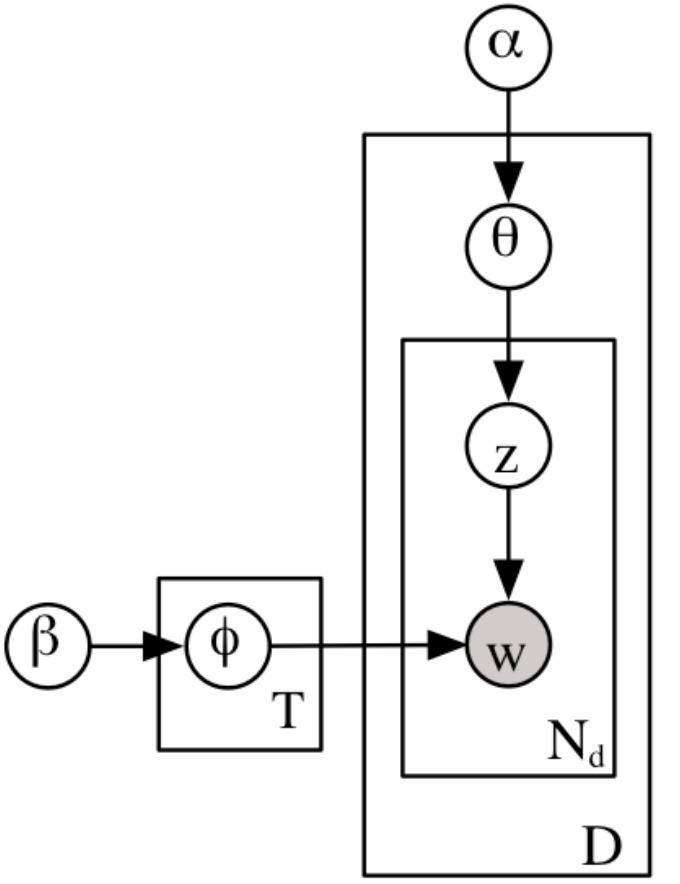
[Preliminary results, Qiong, 2005]



Bags of Words, or Bags of Topics?

Clustering words into topics with Latent Dirichlet Allocation

[Blei, Ng, Jordan 2003]



Probabilistic model for generating document D:

1. Pick a topic distribution P(z|θ) according to P(θ|α)
2. For each word w
   - Pick topic z from P(z | θ)
   - Pick word w from P(w |z, ϕ)

Training this model defines topics (i.e., ϕ which defines P(W|Z))

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

Significance:

- Learned topics reveal hidden, implicit semantic categories in the corpus
- In many cases, we can represent documents with 10^2 topics instead of 10^5 words
- Especially important for short documents (e.g., emails). Topics overlap when words don't !

FIELD	SCIENCE	BALL	JOB
MAGNETIC	STUDY	GAME	WORK
MAGNET	SCIENTISTS	TEAM	JOBS
WIRE	SCIENTIFIC	FOOTBALL	CAREER
NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CURRENT	WORK	PLAYERS	EMPLOYMENT
COIL	RESEARCH	PLAY	OPPORTUNITIES
POLES	CHEMISTRY	**FIELD**	WORKING
IRON	TECHNOLOGY	PLAYER	TRAINING
COMPASS	MANY	BASKETBALL	SKILLS
LINES	MATHEMATICS	COACH	CAREERS
CORE	BIOLOGY	PLAYED	POSITIONS
ELECTRIC	**FIELD**	PLAYING	FIND
DIRECTION	PHYSICS	HIT	POSITION
FORCE	LABORATORY	TENNIS	**FIELD**
MAGNETS	STUDIES	TEAMS	OCCUPATIONS
BE	WORLD	GAMES	REQUIRE
MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
POLE	STUDYING	BAT	EARN
INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

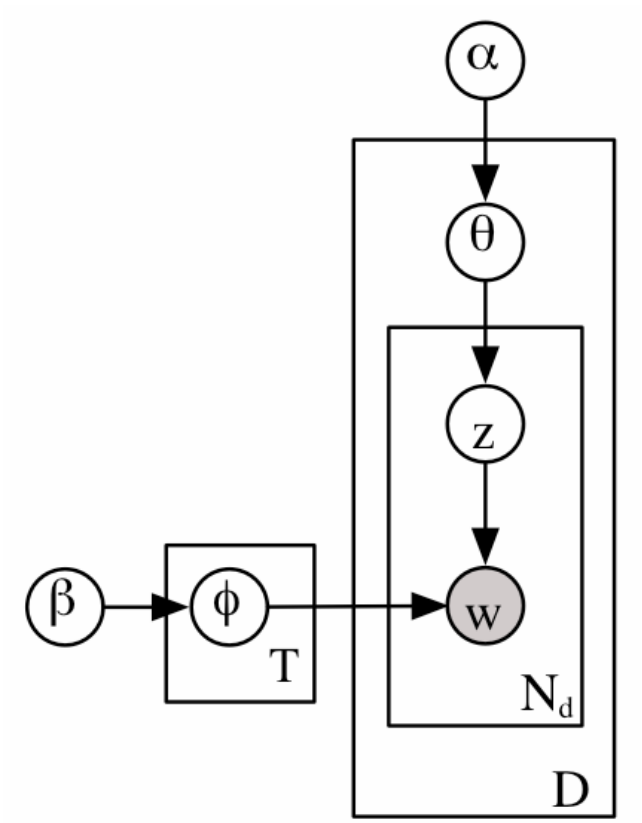
3. Can we analyze roles and relationships between people by analyzing email word or topic distributions?

Author-Recipient-Topic model for Email

Latent Dirichlet Allocation

(LDA)

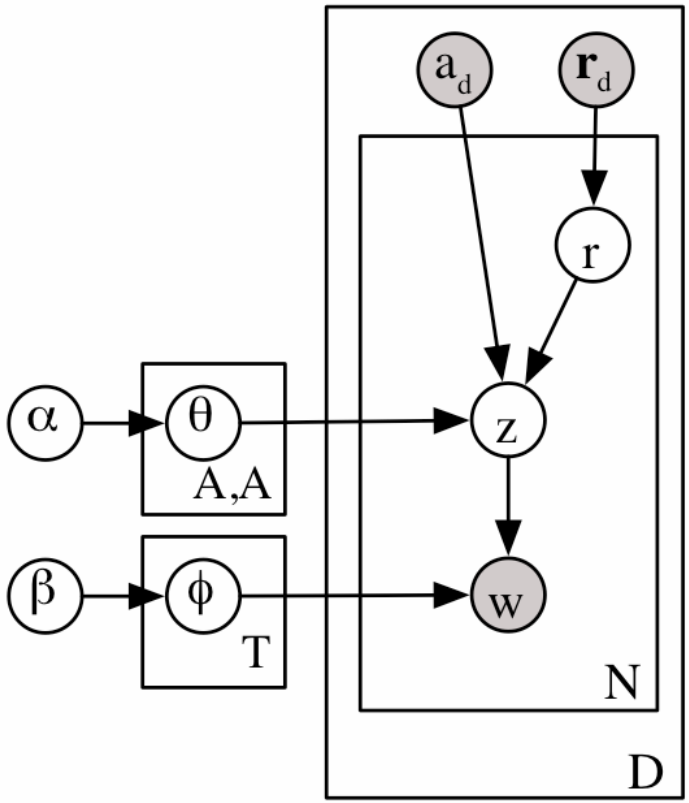
[Blei, Ng, Jordan, 2003]



Author-Recipient Topic

(ART)

[McCallum, Corrada, Wang, 2004]



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 sender/receivers discovered by ART [McCallum et al, 2004]

Top words within topic :

Top author-recipients exhibiting this topic

Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
attached	0.0742	day	0.0419	game	0.0170
agreement	0.0493	friday	0.0418	draft	0.0156
review	0.0340	morning	0.0369	week	0.0135
questions	0.0257	monday	0.0282	team	0.0135
draft	0.0245	office	0.0282	eric	0.0130
letter	0.0239	wednesday	0.0267	make	0.0125
comments	0.0207	tuesday	0.0261	free	0.0107
copy	0.0165	time	0.0218	year	0.0106
revised	0.0161	good	0.0214	pick	0.0097
document	0.0156	thursday	0.0191	phillip	0.0095
G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

Topics, and prominent sender/receivers discovered by ART

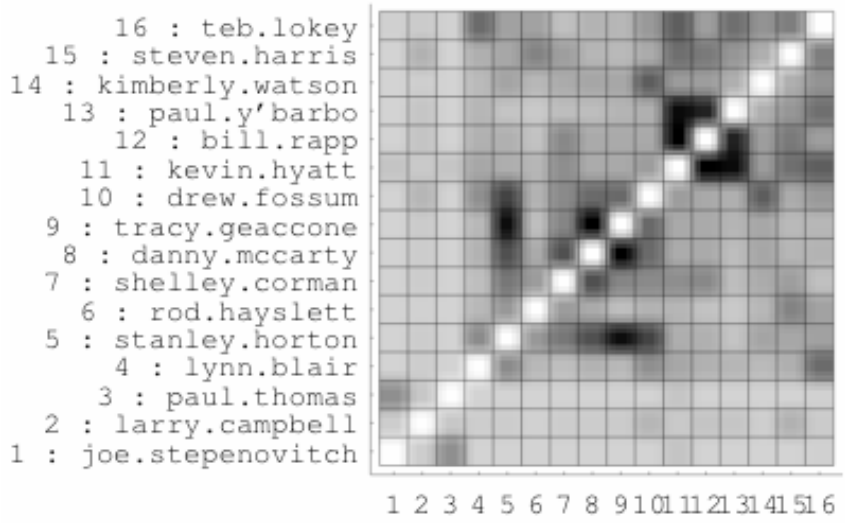
Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	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 L.Kitchen	0.2158	J.Dasovich J.Steffes	0.1231	J.Dasovich R.Shapiro	0.3338	R.Haylett T.Geaccone	0.1432
S.Beck J.Lavorato	0.0826	J.Dasovich R.Shapiro	0.1133	J.Dasovich J.Steffes	0.2440	T.Geaccone R.Haylett	0.0737
S.Beck S.White	0.0530	M.Taylor E.Sager	0.0218	J.Dasovich R.Sanders	0.1394	R.Haylett D.Fossum	0.0420

Beck = "Chief Operations Officer"

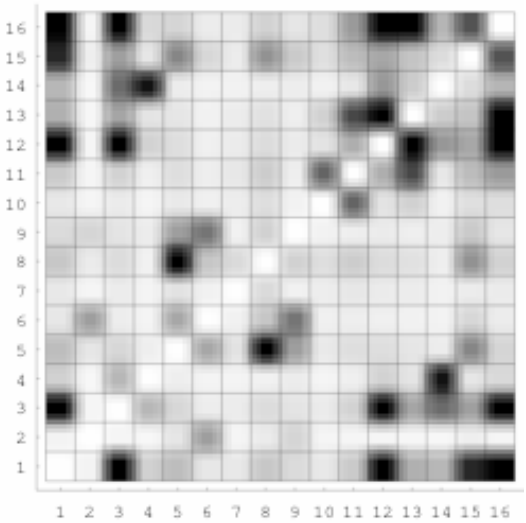
Dasovich = "Government Relations Executive"
Shapiro = "Vice Presidence of Regulatory Affairs"
Steffes = "Vice President of Government Affairs"

Discovering Role Similarity

Traditional SNA



ART



connection strength (A,B) =

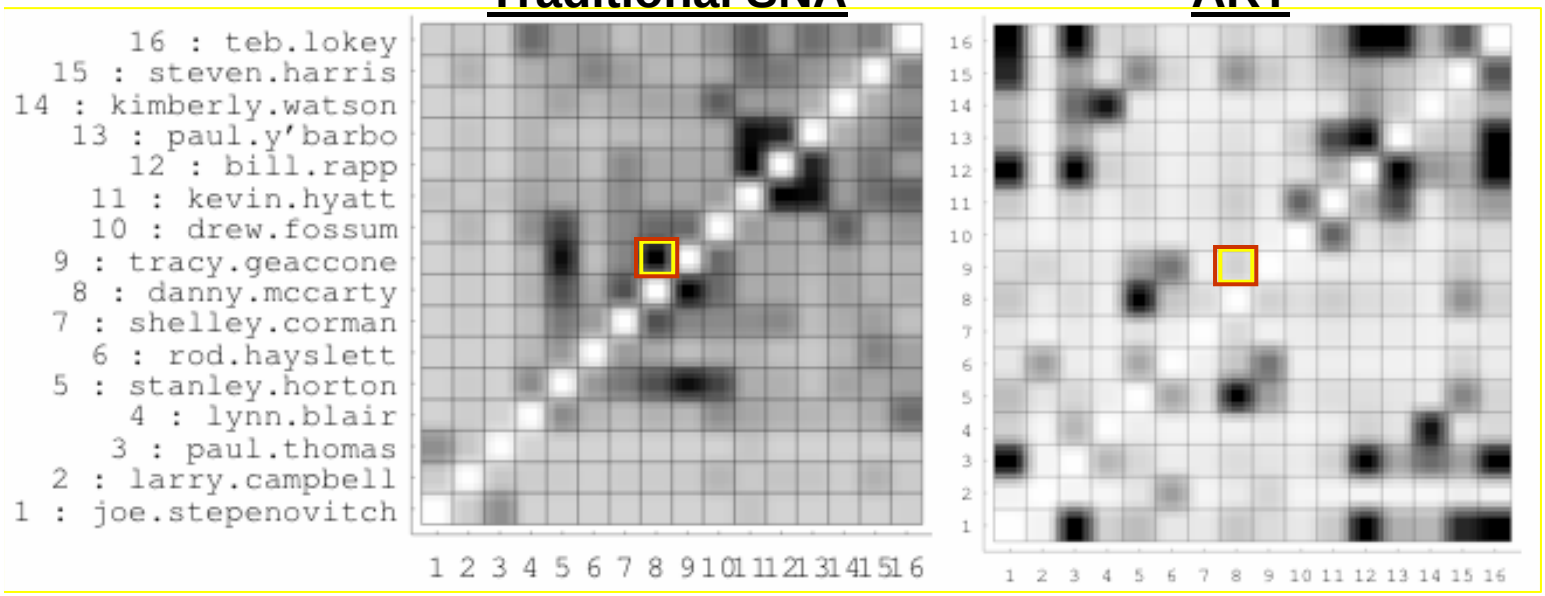
Similarity in recipients they sent email to

Similarity in authored topics, conditioned on recipient

Discovering Role Similarity

Tracy Geaconne ⇔ Dan McCarty

Traditional SNA **ART**



Similar (send email to same individuals)

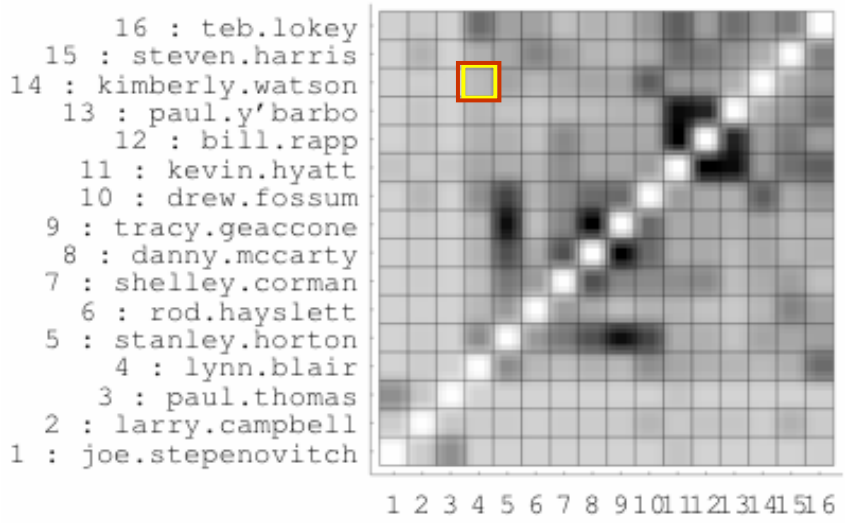
Different (discuss different topics)

Geaconne = "Secretary"
McCarty = "Vice President"

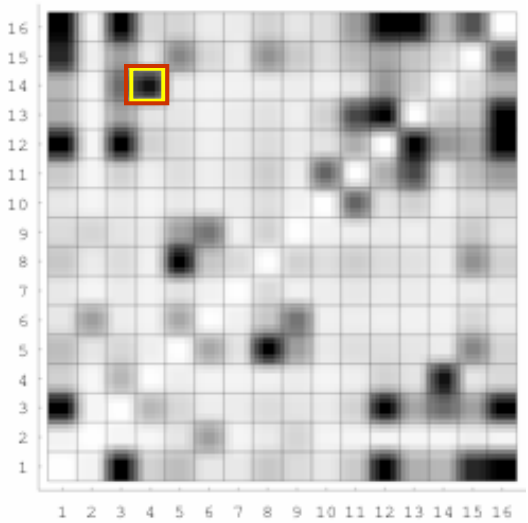
Discovering Role Similarity

Lynn Blair ⇔ Kimberly Watson

Traditional SNA



ART



Different (send to different individuals)

Similar (discuss same topics)

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

So far: what structure can we extract from email?

1. Projects
2. Email intent
3. Latent semantic topics
4. User relationships/roles

4. Why not use entire workstation contents? and the web...

Entity Descriptions from Workstation

[with S. Wang, W. Cohen]

Idea:

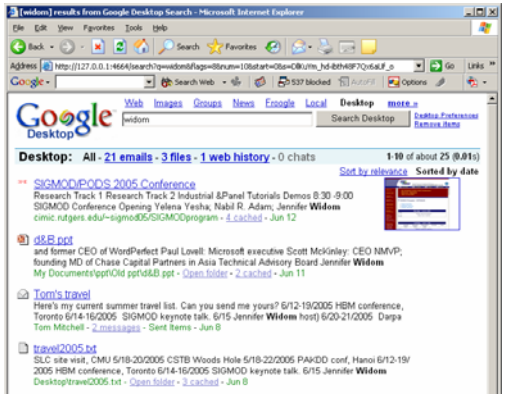
- Represent ***every*** entity (person, project, organization,...) by the distribution of words associated with it across entire workstation

- Simple implementation: type the name of the entity into Google Desktop Search, and form histogram of all associated words in all returned files, emails, webpages on the workstation

- Use this to find which entities are related, how

Word Vector Representation for Entities

- Type name of entity into Google desktop search, collect the returned snippets of text



- Create vector with one feature per word. It's value is based on the number of word occurrences in the returned Google snippets.
- Assign each feature its TFIDF value. Then normalize vectors to unit length.

Measuring Distance Between Entity Word Vectors

- Distance between vectors is their dot product:

$$dist(x, y) = \sum_i x_i y_i$$

- Use distance to automatically construct descriptions for every entity, describing it by its most closely related organization, department, discipline, conference, funder, and person

Simple Experiment

Create word distribution vectors for 20k words. Type in simple ontology and instances:

People: widom faloutsos wcohen mccallum yifen huang indra rustandi rebecca hutchinson stefan niculescu john ramish jay pujara sharon cavlovich woodside diane stidle randy bryant jeannette wing wei wang sophie zhenzhen daniel neill kaustav carlos guestrin murphy marcel

Organizations: cmu stanford mit pitt upenn sri umass mitre lockheed

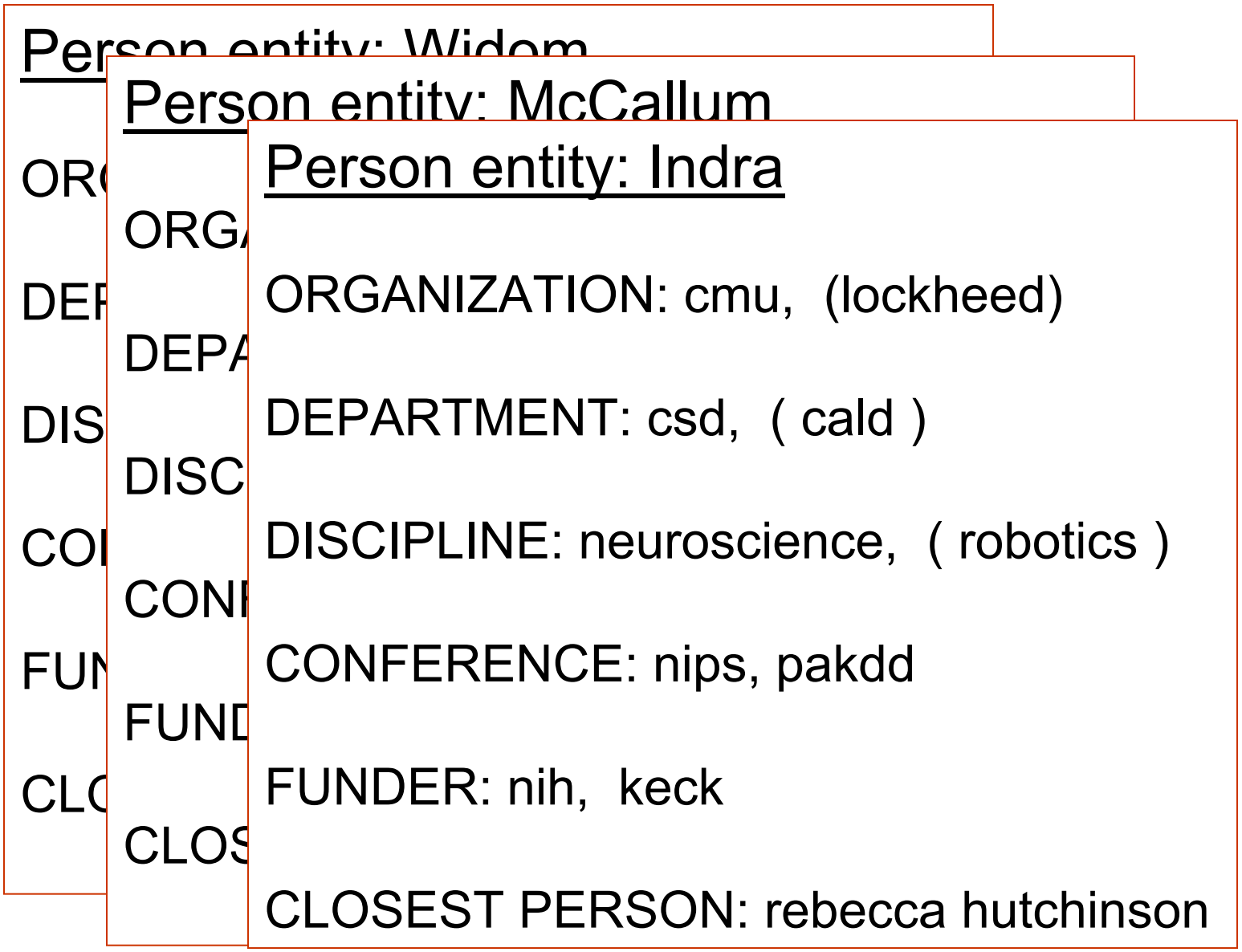
Departments: csd cald lti ri hci cnbc

Disciplines: ai databases biology cogsci neuroscience datamining robotics psychology

Conferences: aaai sigmod nips pakdd hbm ijcai

Funders: darpa nih keck nsf

Automatically Constructed Descriptions



Automatically Constructed Descriptions

Confere

PERSON

ORGANI

DEPART

DISCIPL

CLOSES

FUNDER

Conference entity: IJCAI

PERSON: indra, (guestrin)

ORGANIZATION: lockheed, (pitt)

DEPARTMENT: csd, (hci)

DISCIPLINE: ai (databases)

CLOSEST CONFERENCE: aaai, sigmod

FUNDER: nsf, nih

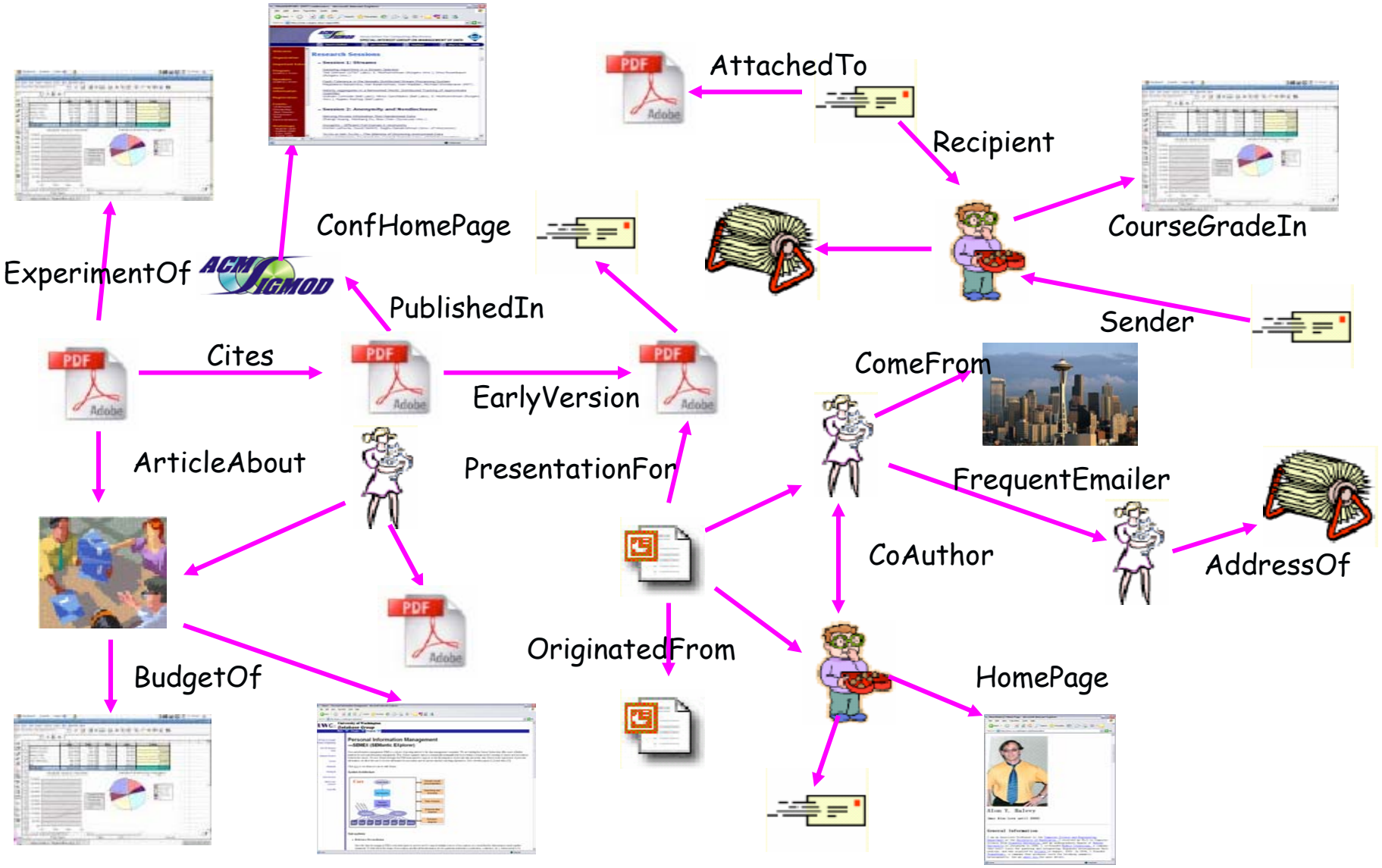
Summary: Workstation Word Vectors

- Surprisingly accurate descriptions of entities by simply typing their *names* into Desktop Search engine
- Leverages huge redundancy in workstation contents
- Future work: represent and infer more subtle relations

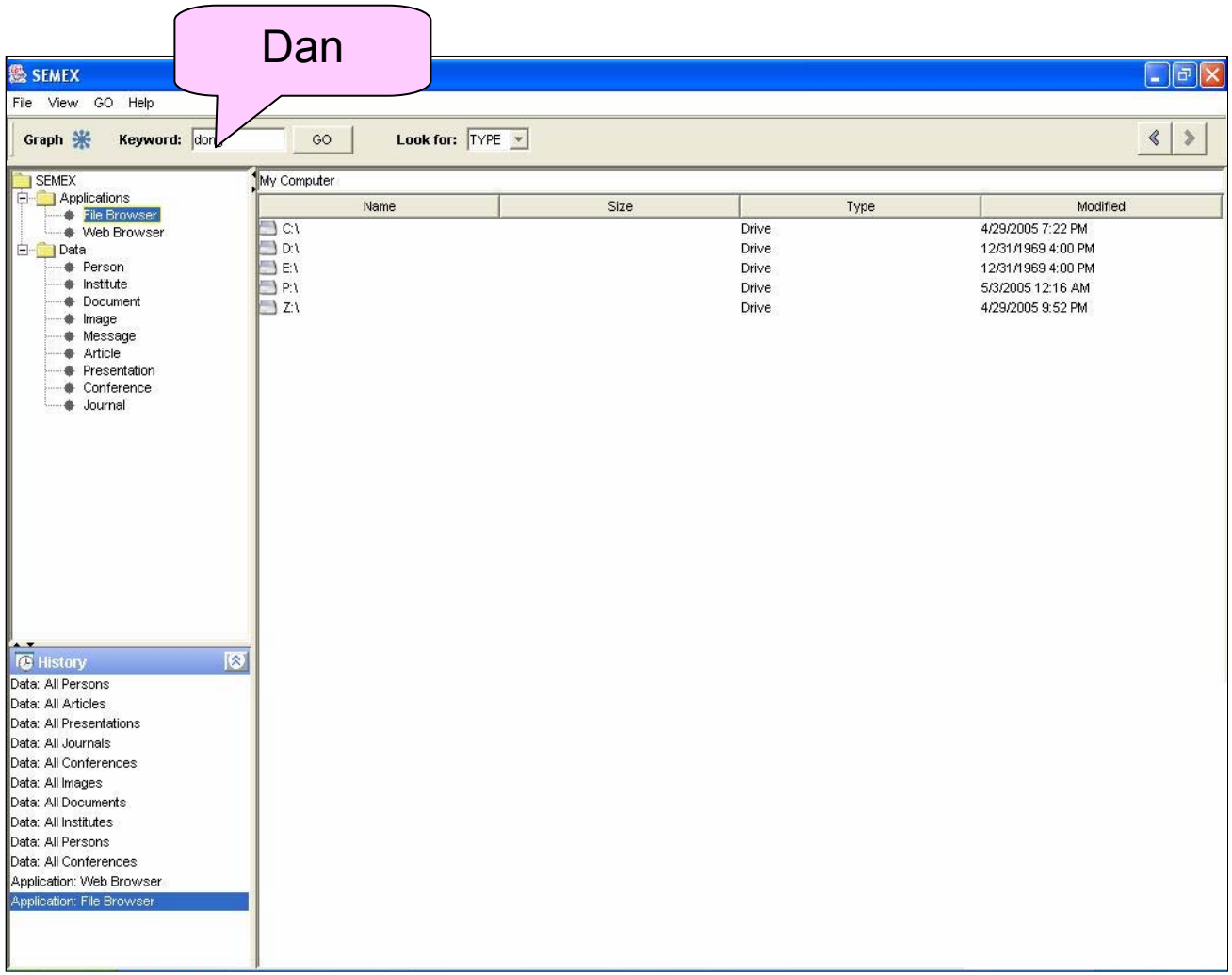
SEMEX – Alon Halevy

SEMEX Goal: Network of Associations Between Object Instances on One's Desktop

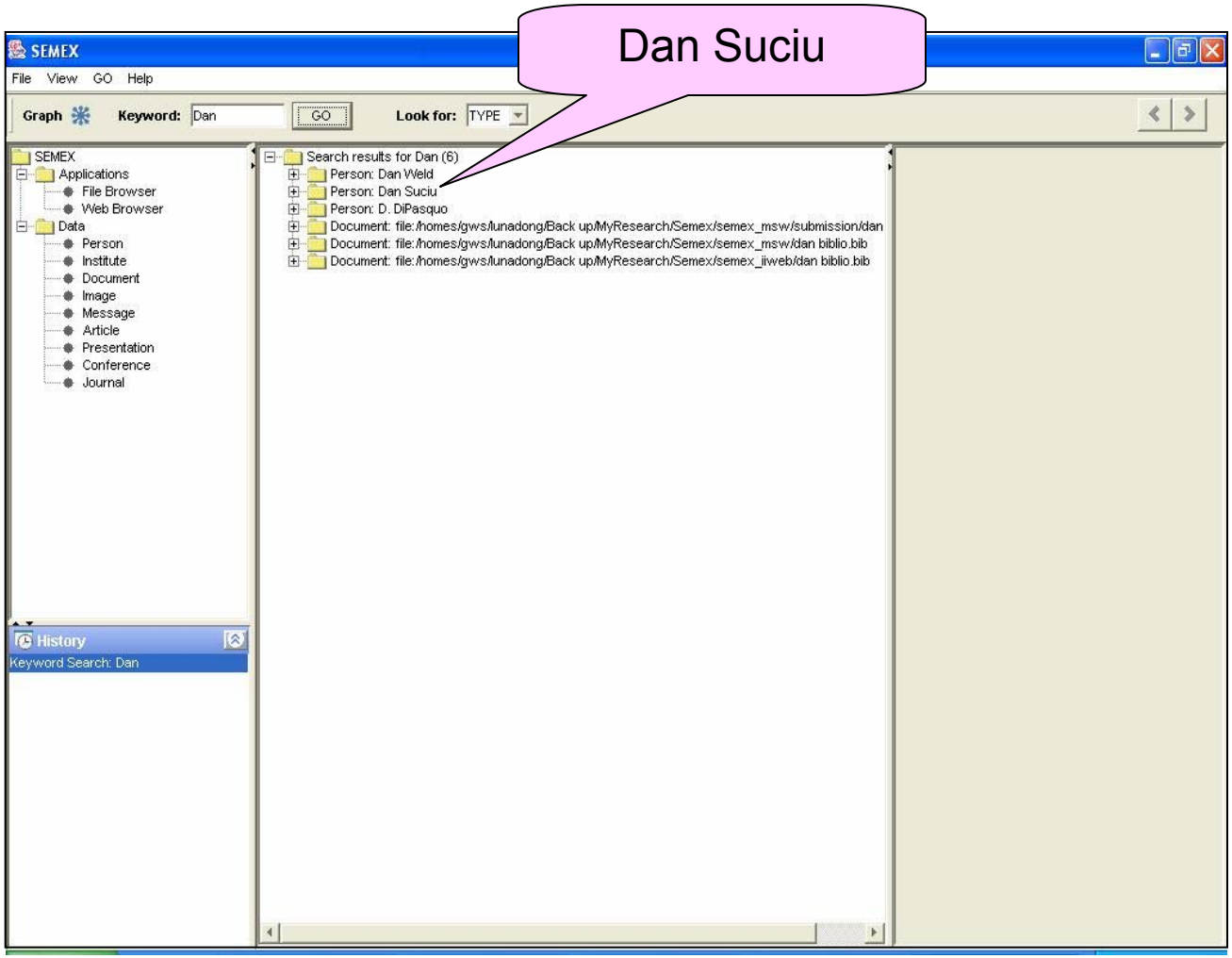
[Halevy, 2004]



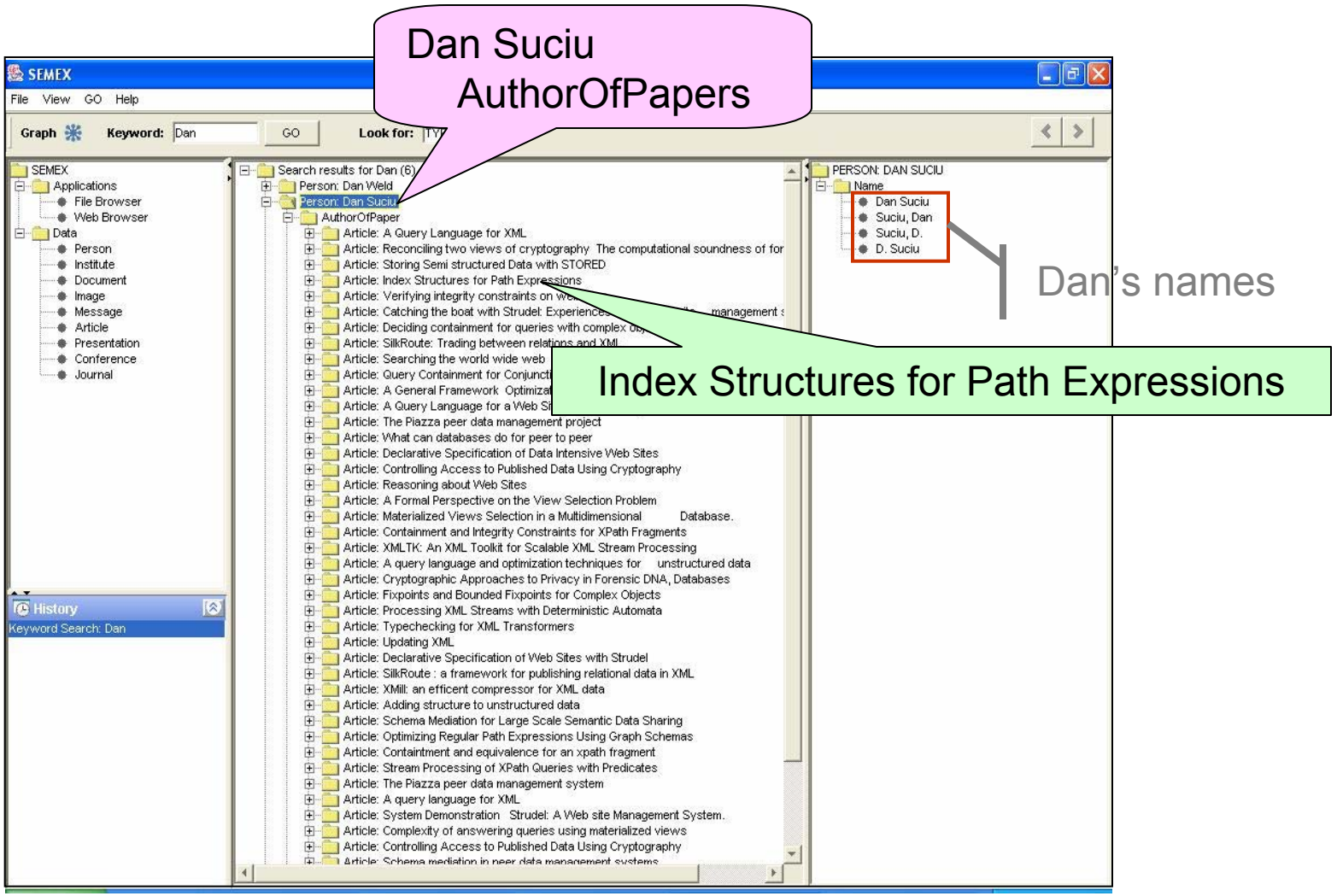
SEMEX (SEMantic EXplorer) – III. Association Search and Browsing



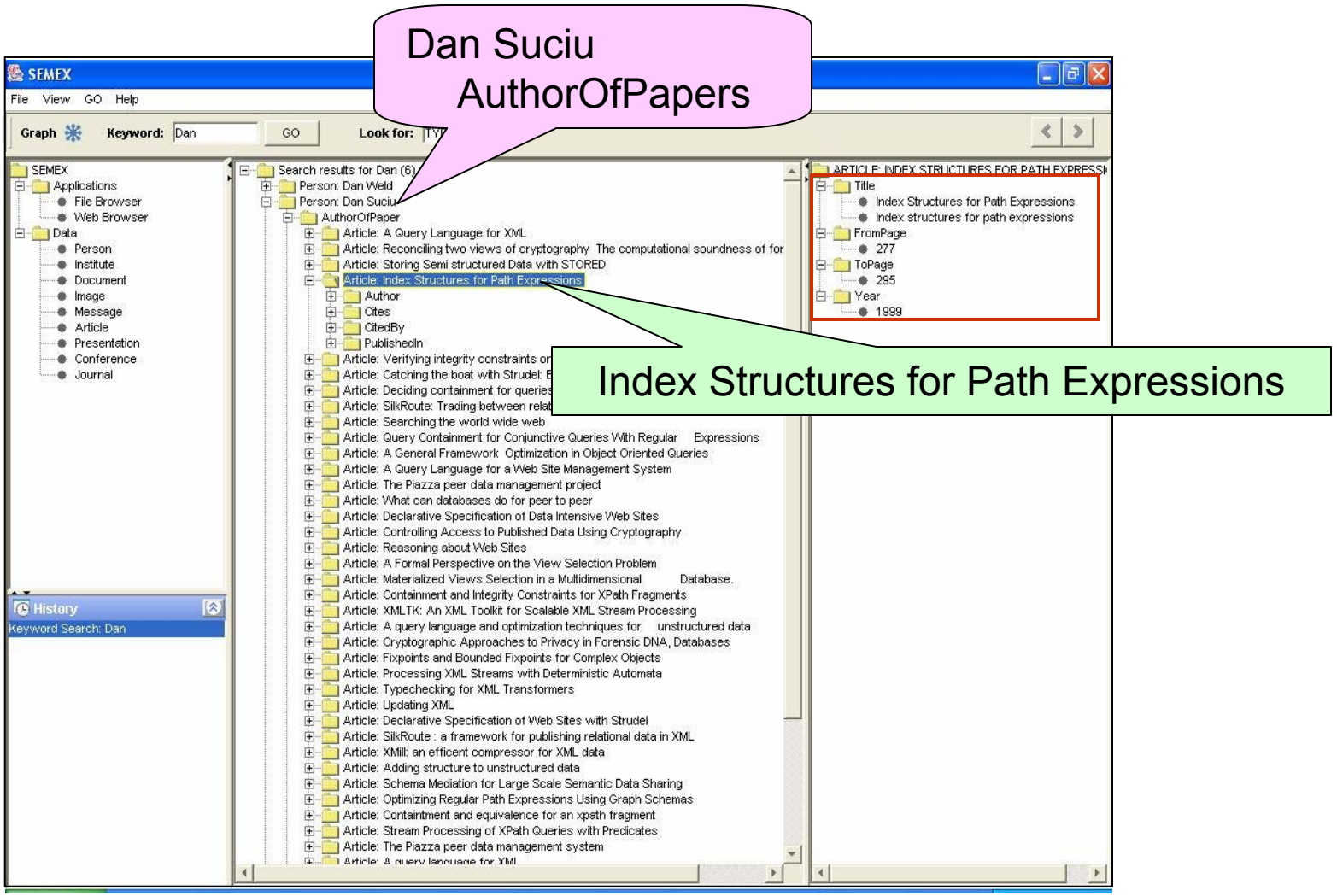
SEMEX (SEMantic EXplorer) – III. Association Search and Browsing



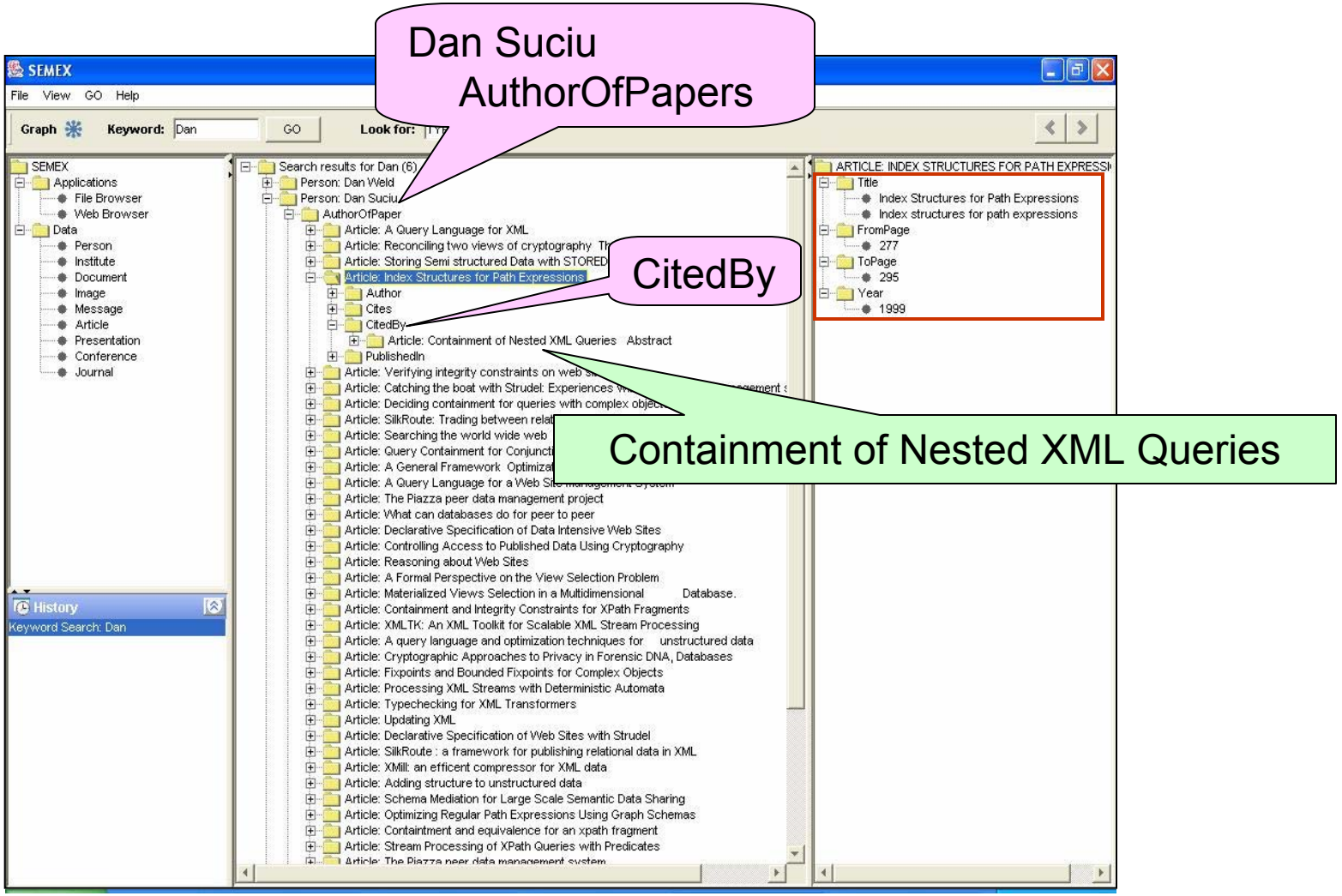
SEMEX (SEMantic EXplorer) – III. Association Search and Browsing



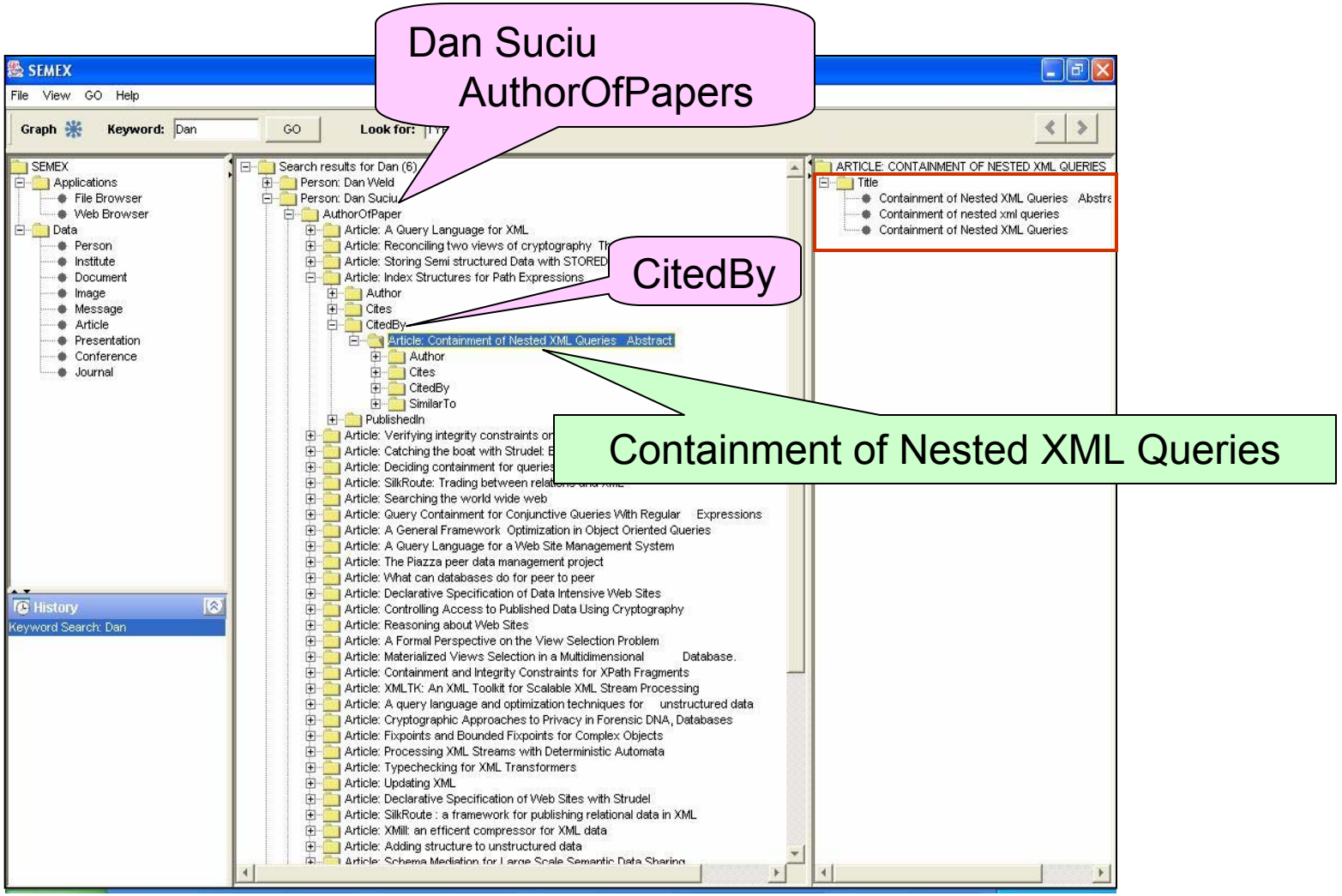
SEMEX (SEMantic EXplorer) – III. Association Search and Browsing



SEMEX (SEMantic EXplorer) – III. Association Search and Browsing



SEMEX (SEMantic EXplorer) – III. Association Search and Browsing

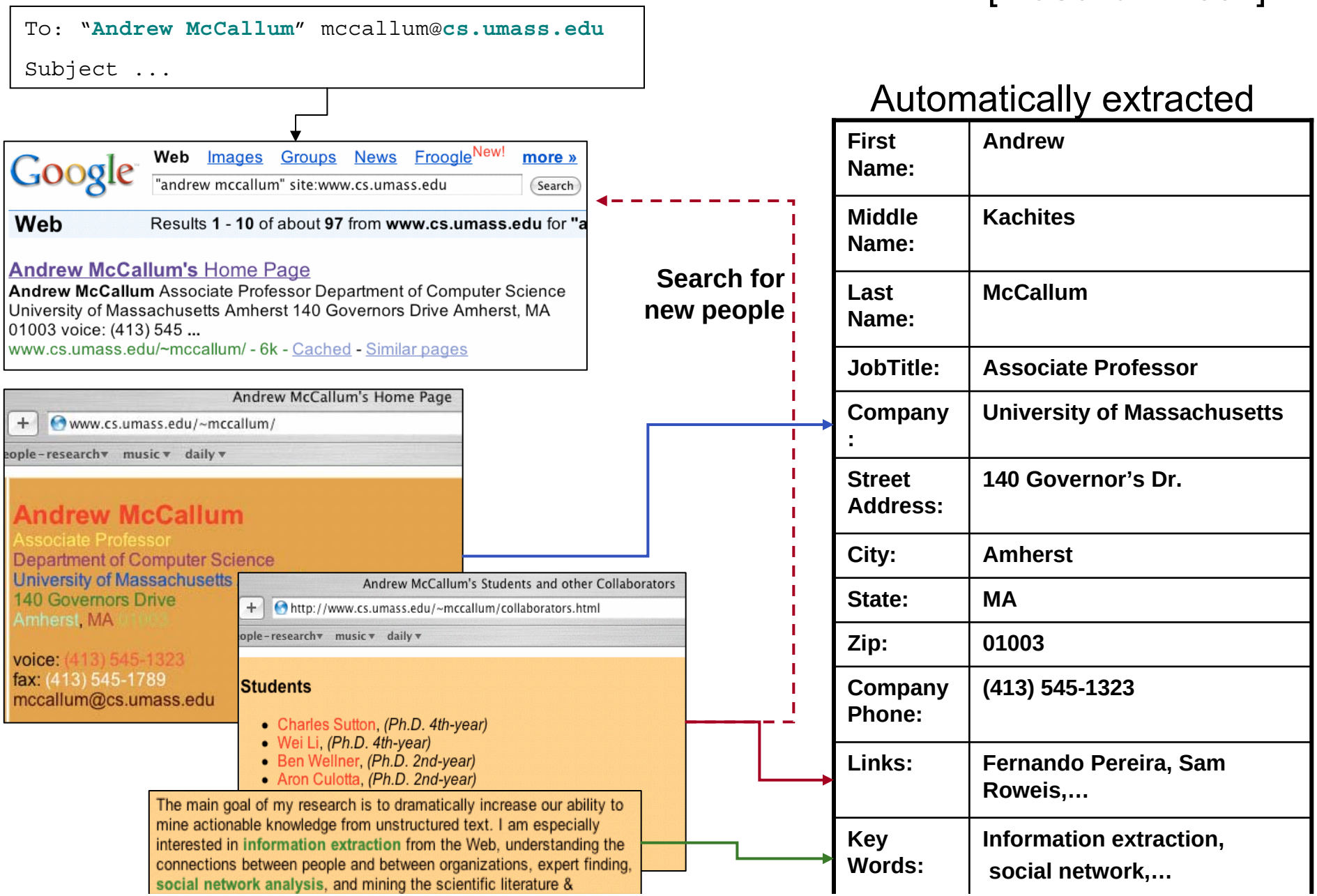


DEX

Andrew McCallum

Extracting Contact Information from the Web

[McCallum 2004]



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,…

Results Summary

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
	94.50	85.73	76.33	80.76

Summary: Much Progress, More Needed

- Extracting structured knowledge
  - Email classification
  - Information extraction from text
  - Social network analysis
  - Discovering latent structures (projects, semantic topics, ...)
  - Entity semantics as word distribution across workstation
  - Linking workstation and web information
  - Coreference resolution
  - Leading toward structured database/knowledge base of
    - people, projects, tasks, roles, deadlines, …
- Tools like Google Desktop Search suddenly make it easy to do this kind of experimental research