Semi-Supervised Learning and Text Analysis

Machine Learning 10-701
November 29, 2005

Tom M. Mitchell
Carnegie Mellon University

Document Classification: Bag of Words Approach


the world of
TOTAL

All About The Company
Global Activities
Corporate Structure
TOTAL's Story
Upstream Strategy
Downstream Strategy
Chemicals Strategy
TOTAL Foundation
Homepage

all about the
company

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

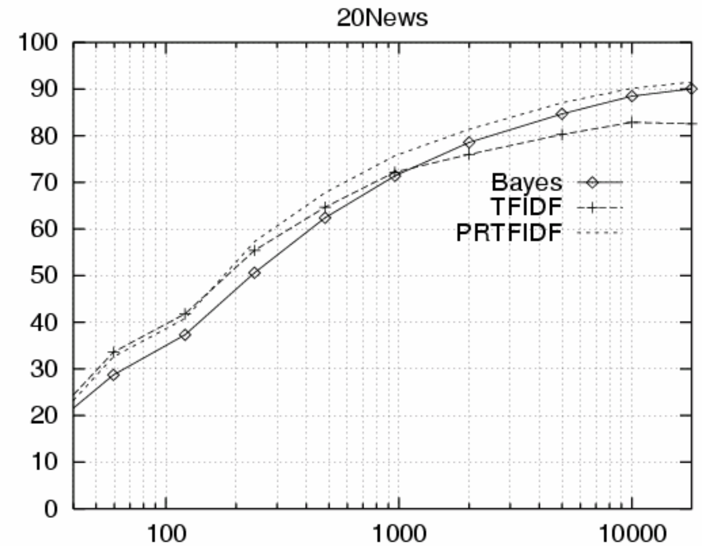
aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy



Accuracy vs. Training set size (1/3 withheld for test)

For code, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"

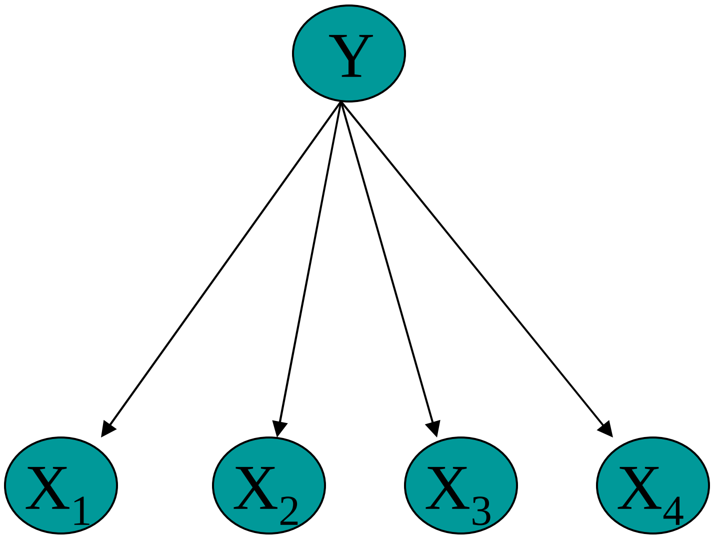
Supervised Training for Document Classification

- Common algorithms:
  - Logistic regression, Support Vector Machines, Bayesian classifiers
- Quite successful in practice
  - Email classification (spam, foldering, ...)
  - Web page classification (product description, publication, ...)
  - Intranet document organization
- Research directions:
  - More elaborate, domain-specific classification models (e.g., for email)
  - Using unlabeled data too → semi-supervised methods

EM for Semi-supervised document classification

Using Unlabeled Data to Help Train Naïve Bayes Classifier

Learn P(Y|X)



Y	X1	X2	X3	X4
1	0	0	1	1
0	0	1	0	0
0	0	0	1	0
?	0	1	1	0
?	0	1	0	1

- **Inputs:** Collections $\mathcal{D}^l$ of labeled documents and $\mathcal{D}^u$ of unlabeled documents.
- Build an initial naive Bayes classifier, $\hat{\theta}$, from the labeled documents, $\mathcal{D}^l$, only. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta)$ (see Equations 5 and 6).
- Loop while classifier parameters improve, as measured by the change in $l_c(\theta|\mathcal{D};\mathbf{z})$ (the complete log probability of the labeled and unlabeled data
  - **(E-step)** Use the current classifier, $\hat{\theta}$, to estimate component membership of each unlabeled document, *i.e.*, the probability that each mixture component (and class) generated each document, $\mathrm{P}(c_j|d_i;\hat{\theta})$ (see Equation 7).
  - **(M-step)** Re-estimate the classifier, $\hat{\theta}$, given the estimated component membership of each document. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \arg\max_\theta \mathrm{P}(\mathcal{D}|\theta)\mathrm{P}(\theta)$ (see Equations 5 and 6).
- **Output:** A classifier, $\hat{\theta}$, that takes an unlabeled document and predicts a class label.

From [Nigam et al., 2000]

E Step:

$$\mathrm{P}(y_i = c_j | d_i; \hat{\theta}) = \frac{\mathrm{P}(c_j|\hat{\theta})\mathrm{P}(d_i|c_j;\hat{\theta})}{\mathrm{P}(d_i|\hat{\theta})}$$

$$= \frac{\mathrm{P}(c_j|\hat{\theta}) \prod_{k=1}^{|d_i|} \mathrm{P}(w_{d_{i,k}}|c_j;\hat{\theta})}{\sum_{r=1}^{|\mathcal{C}|} \mathrm{P}(c_r|\hat{\theta}) \prod_{k=1}^{|d_i|} \mathrm{P}(w_{d_{i,k}}|c_r;\hat{\theta})}.$$

M Step:

w_t is t-th word in vocabulary

$$\hat{\theta}_{w_t|c_j} \equiv \mathrm{P}(w_t|c_j;\hat{\theta}) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} N(w_t, d_i)\mathrm{P}(y_i = c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|\mathcal{D}|} N(w_s, d_i)\mathrm{P}(y_i = c_j|d_i)},$$

$$\hat{\theta}_{c_j} \equiv \mathrm{P}(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} \mathrm{P}(y_i = c_j|d_i)}{|\mathcal{C}| + |\mathcal{D}|}.$$

Elaboration 1: Downweight the influence of unlabeled examples by factor λ

$$
\begin{aligned}
l_c(\theta|\mathcal{D};\mathbf{z}) \;=\; & \log(\mathrm{P}(\theta)) + \sum_{d_i \in \mathcal{D}^l} \sum_{j=1}^{|\mathcal{C}|} z_{ij} \log\left(\mathrm{P}(c_j|\theta)\mathrm{P}(d_i|c_j;\theta)\right) \\
& + \lambda \left(\sum_{d_i \in \mathcal{D}^u} \sum_{j=1}^{|\mathcal{C}|} z_{ij} \log\left(\mathrm{P}(c_j|\theta)\mathrm{P}(d_i|c_j;\theta)\right) \right).
\end{aligned}
$$

Chosen by cross validation

New M step:

$$
\hat{\theta}_{w_t|c_j} \equiv \mathrm{P}(w_t|c_j;\hat{\theta}) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} \Lambda(i) N(w_t, d_i) \mathrm{P}(y_i = c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|\mathcal{D}|} \Lambda(i) N(w_s, d_i) \mathrm{P}(y_i = c_j|d_i)}.
$$

$$
\hat{\theta}_{c_j} \equiv \mathrm{P}(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|\mathcal{D}|} \Lambda(i) \mathrm{P}(y_i = c_j|d_i)}{|\mathcal{C}| + |\mathcal{D}^l| + \lambda|\mathcal{D}^u|}
$$

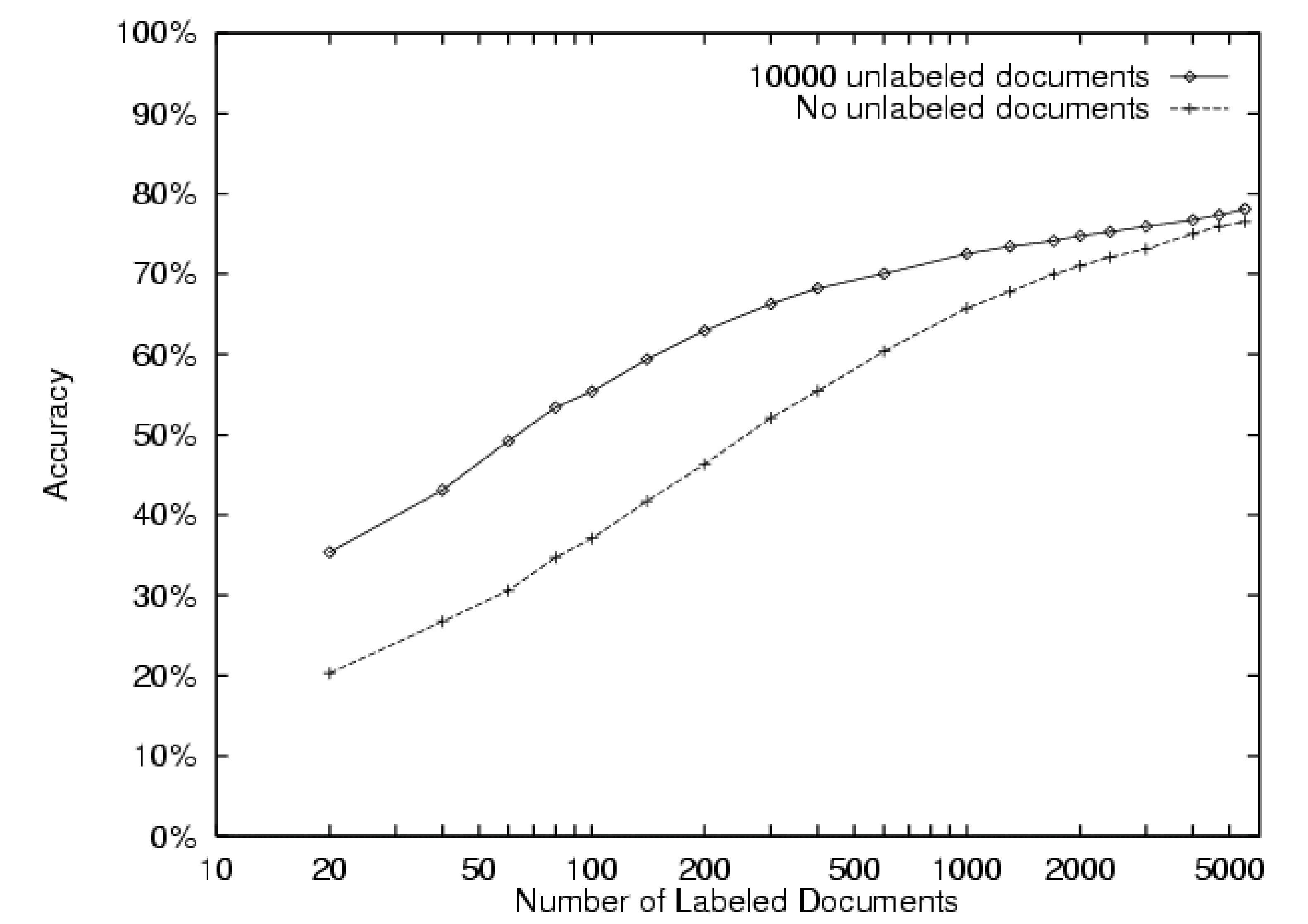
$$
\Lambda(i) = \begin{cases} \lambda & \text{if } d_i \in \mathcal{D}^u \\ 1 & \text{if } d_i \in \mathcal{D}^l. \end{cases}
$$

Table 3. Lists of the words most predictive of the course class in the WebKB data set, as they change over iterations of EM for a specific trial. By the second iteration of EM, many common course-related words appear. The symbol D indicates an arbitrary digit.

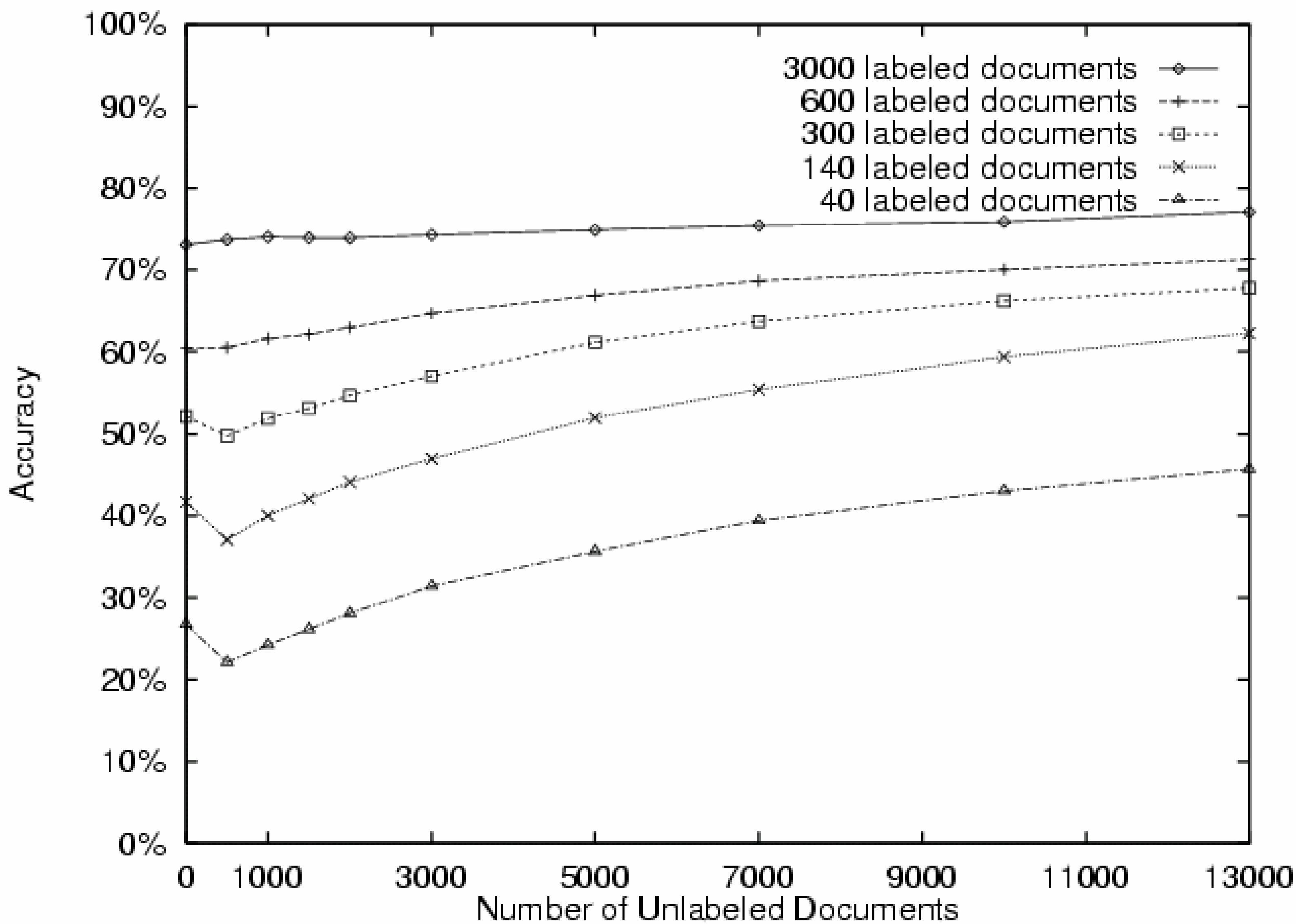
Iteration 0	Iteration 1	Iteration 2
intelligence	DD	D
DD	D	DD
artificial	lecture	lecture
understanding	cc	cc
DDw	$D^{\star}$	$DD{:}DD$
dist	$DD{:}DD$	due
identical	handout	$D^{\star}$
rus	due	homework
arrange	problem	assignment
games	set	handout
dartmouth	tay	set
natural	DDam	hw
cognitive	yurttas	exam
logic	homework	problem
proving	kfoury	DDam
prolog	sec	postscript
knowledge	postscript	solution
human	exam	quiz
representation	solution	chapter
field	assaf	ascii

Using one labeled example per class

20 Newsgroups



20 Newsgroups



EM for Semi-Supervised Doc Classification

- If all data is labeled, corresponds to Naïve Bayes classifier
- If all data unlabeled, corresponds to mixture-of-multinomial clustering
- If both labeled and unlabeled data, it helps if and only if the mixture-of-multinomial modeling assumption is correct

- Of course we could extend this to Bayes net models other than Naïve Bayes (e.g., TAN tree)

Bags of Words, or Bags of Topics?

LDA: Generative model for documents

[Blei, Ng, Jordan 2003]

$$p(D\,|\,\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d\,|\,\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}\,|\,\theta_d)\, p(w_{dn}\,|\,z_{dn},\beta) \right) d\theta_d.$$

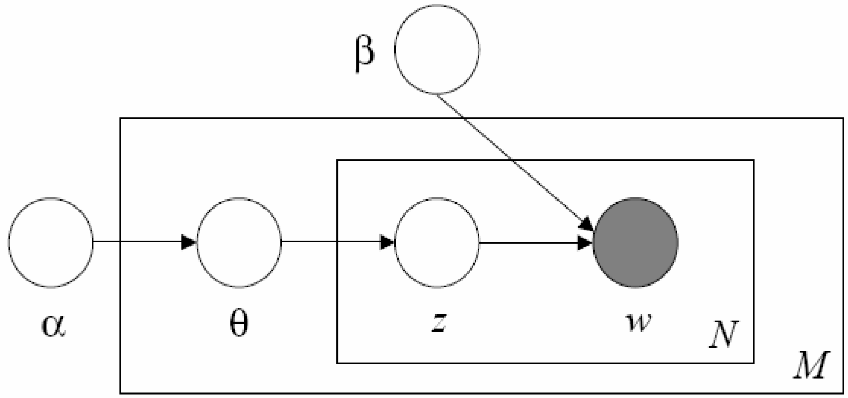
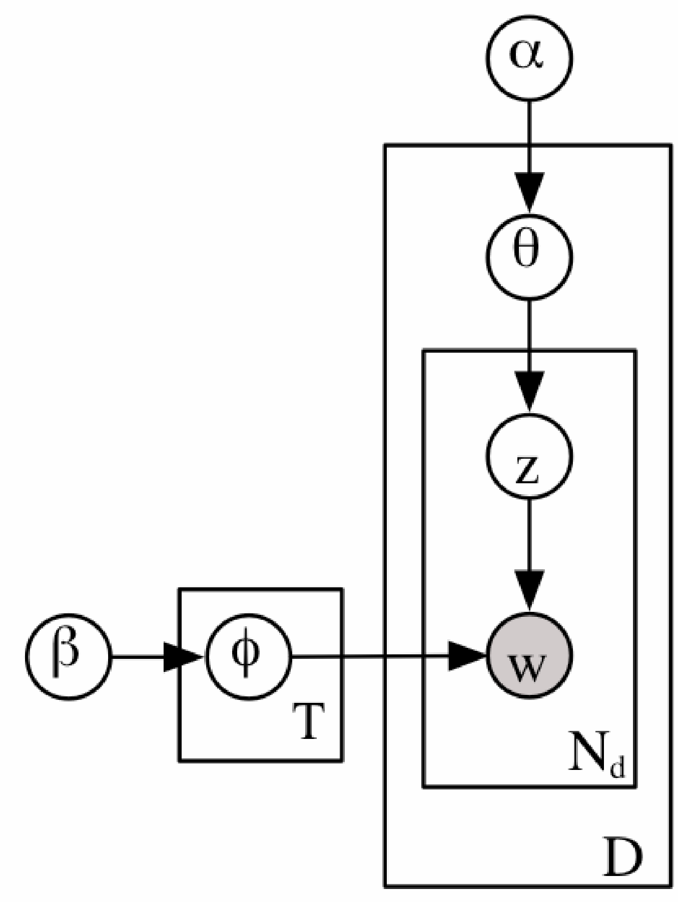


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Also extended to case where number of topics is not known in advance (hierarchical Dirichlet processes – [Blei et al, 2004])

Clustering words into topics with Hierarchical Topic Models (unknown number of clusters)

[Blei, Ng, Jordan 2003]



Probabilistic model for generating document D:

1. Pick a distribution $P(z|\theta)$ of topics according to $P(\theta|\alpha)$
2. For each word w
    - Pick topic z from $P(z \mid \theta)$
    - Pick word w from $P(w \mid z, \phi)$

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]

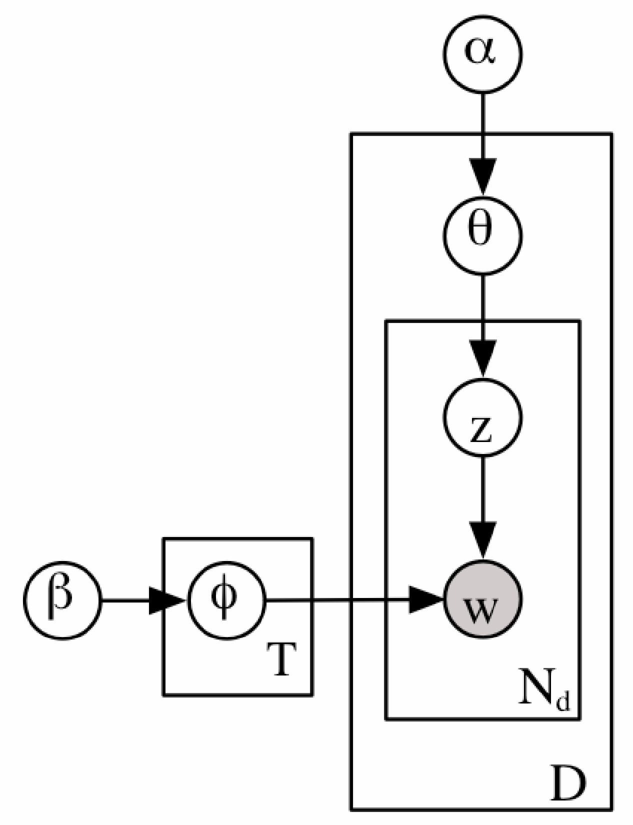
Can we analyze roles and relationships between people by analyzing email word or topic distributions?

Author-Recipient-Topic model for Email

Latent Dirichlet Allocation

(LDA)

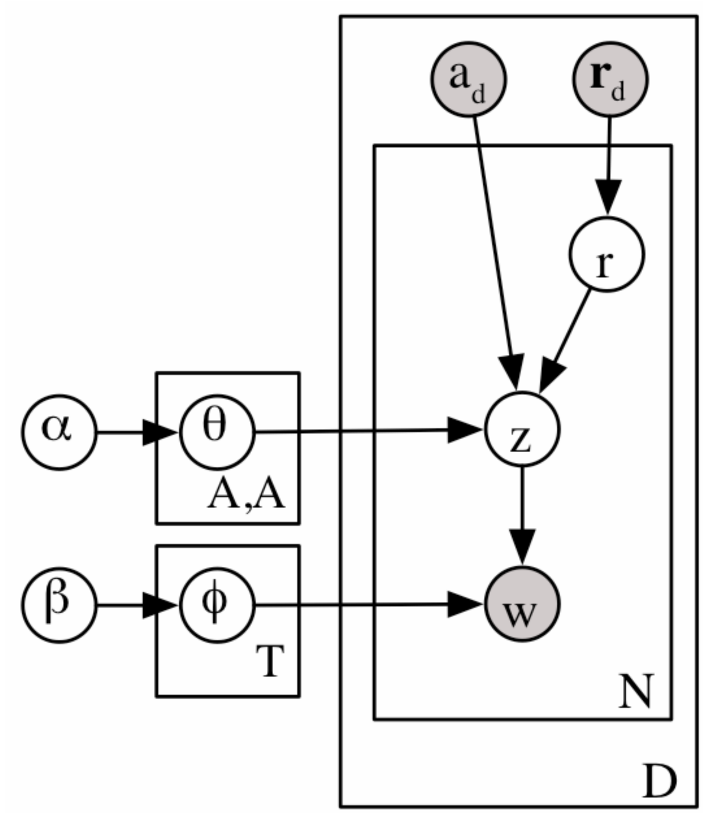
[Blei, Ng, Jordan, 2003]



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

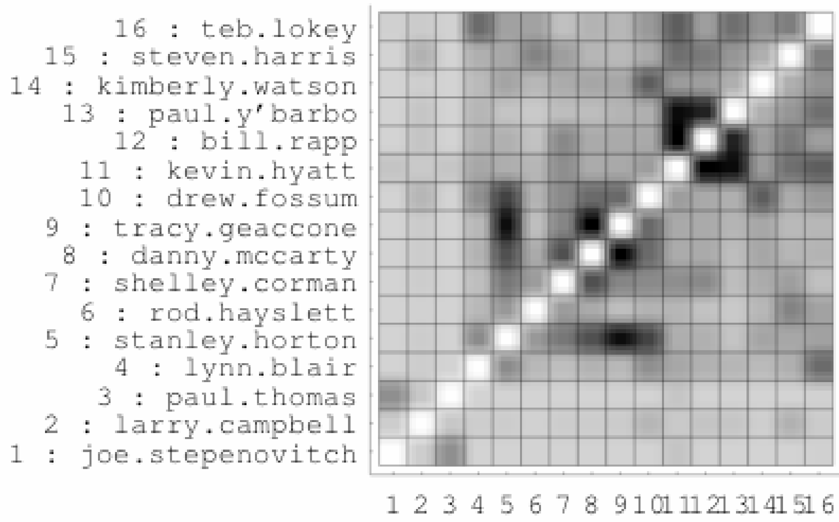
Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	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"

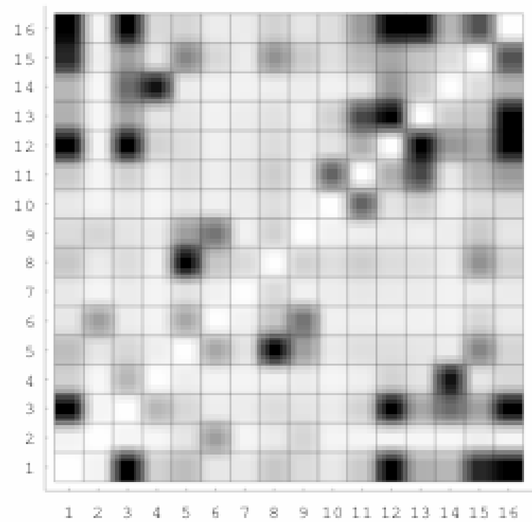
Dasovich = "Government Relations Executive"
Shapiro = "Vice Presidence of Regulatory Affairs"
Steffes = "Vice President of Government Affairs"

Discovering Role Similarity

Traditional SNA



ART



connection strength (A,B) =

Similarity in recipients they sent email to

Similarity in authored topics, conditioned on recipient

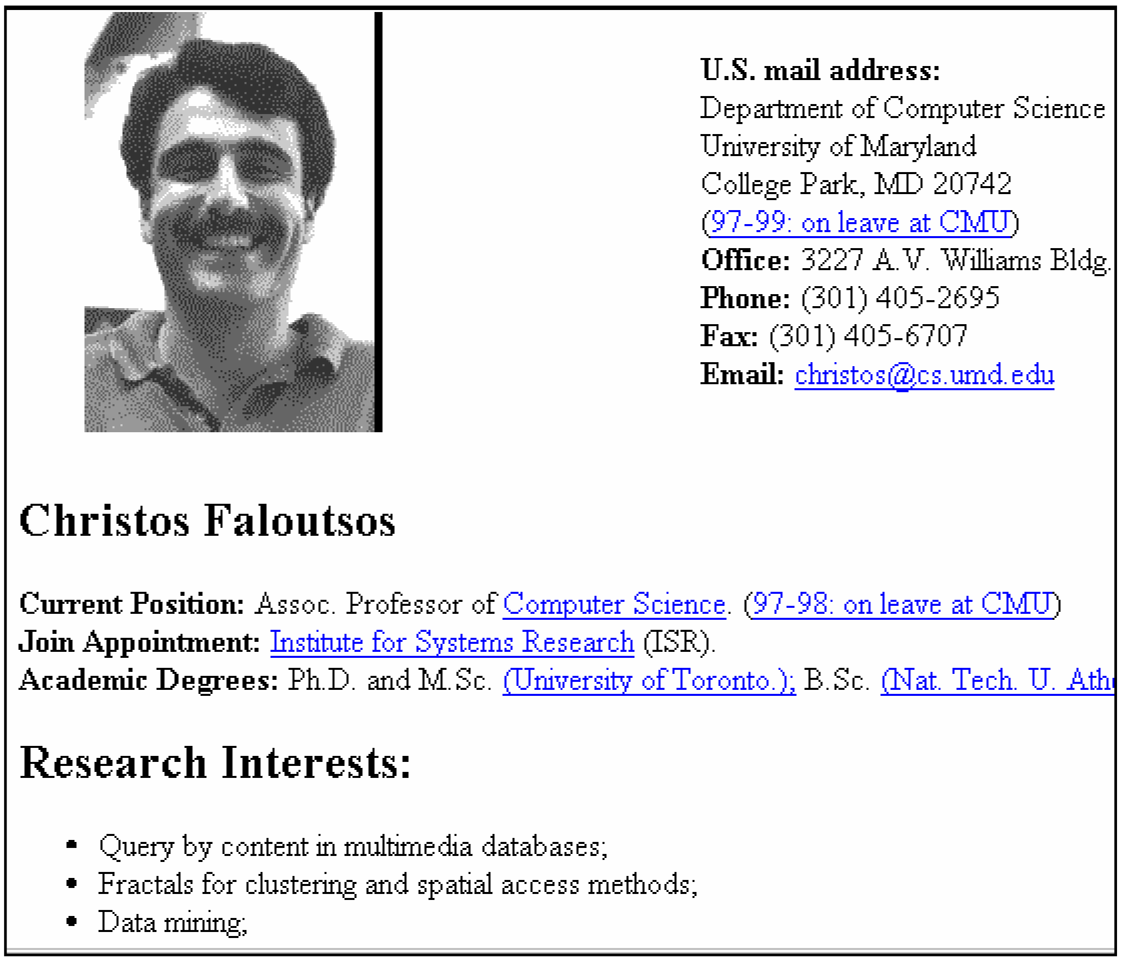
Co-Training for Semi-supervised document classification

Idea: take advantage of *redundancy*

Redundantly Sufficient Features

Professor Faloutsos

my advisor



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

Research Interests:

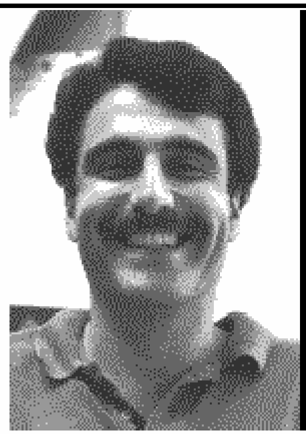
- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Redundantly Sufficient Features

Professor Faloutsos

my advisor

Redundantly Sufficient Features



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

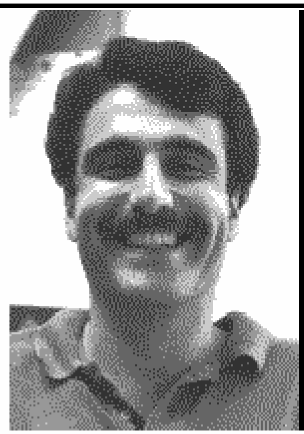
Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Redundantly Sufficient Features

Professor Faloutsos

my advisor



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

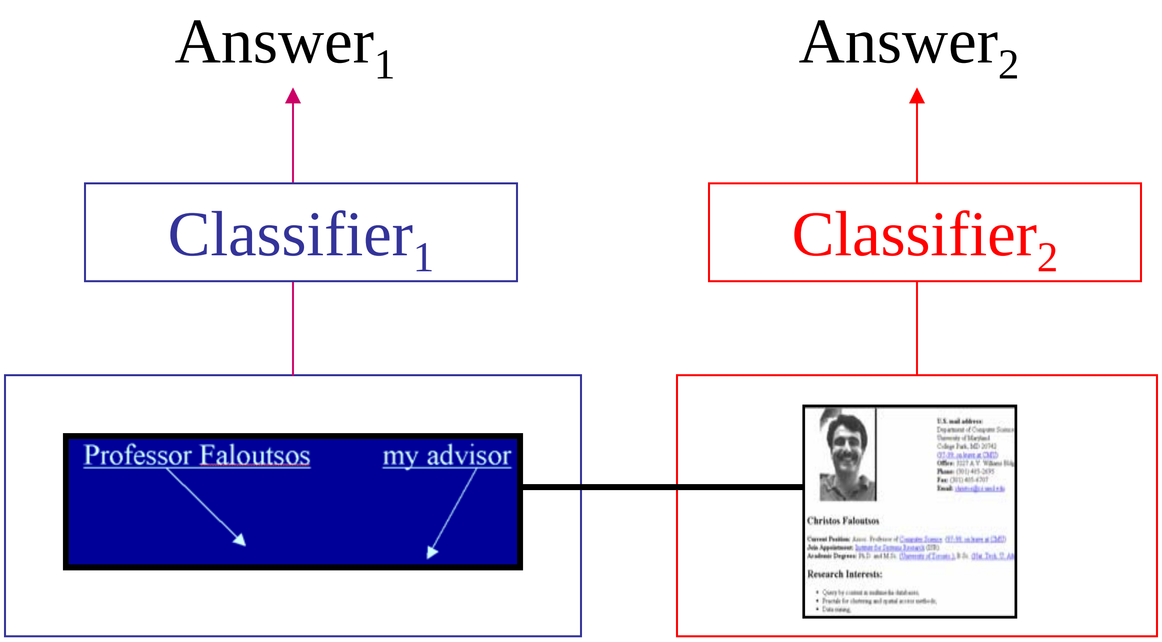
Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Co-Training

Key idea: Classifier$_1$ and Classifier$_J$ must:

1. Correctly classify labeled examples
2. <u>Agree</u> on classification of unlabeled



CoTraining Algorithm #1

[Blum&Mitchell, 1998]

Given: labeled data L,

unlabeled data U

Loop:

Train g1 (hyperlink classifier) using L

Train g2 (page classifier) using L

Allow g1 to label *p* positive, *n* negative examps from U

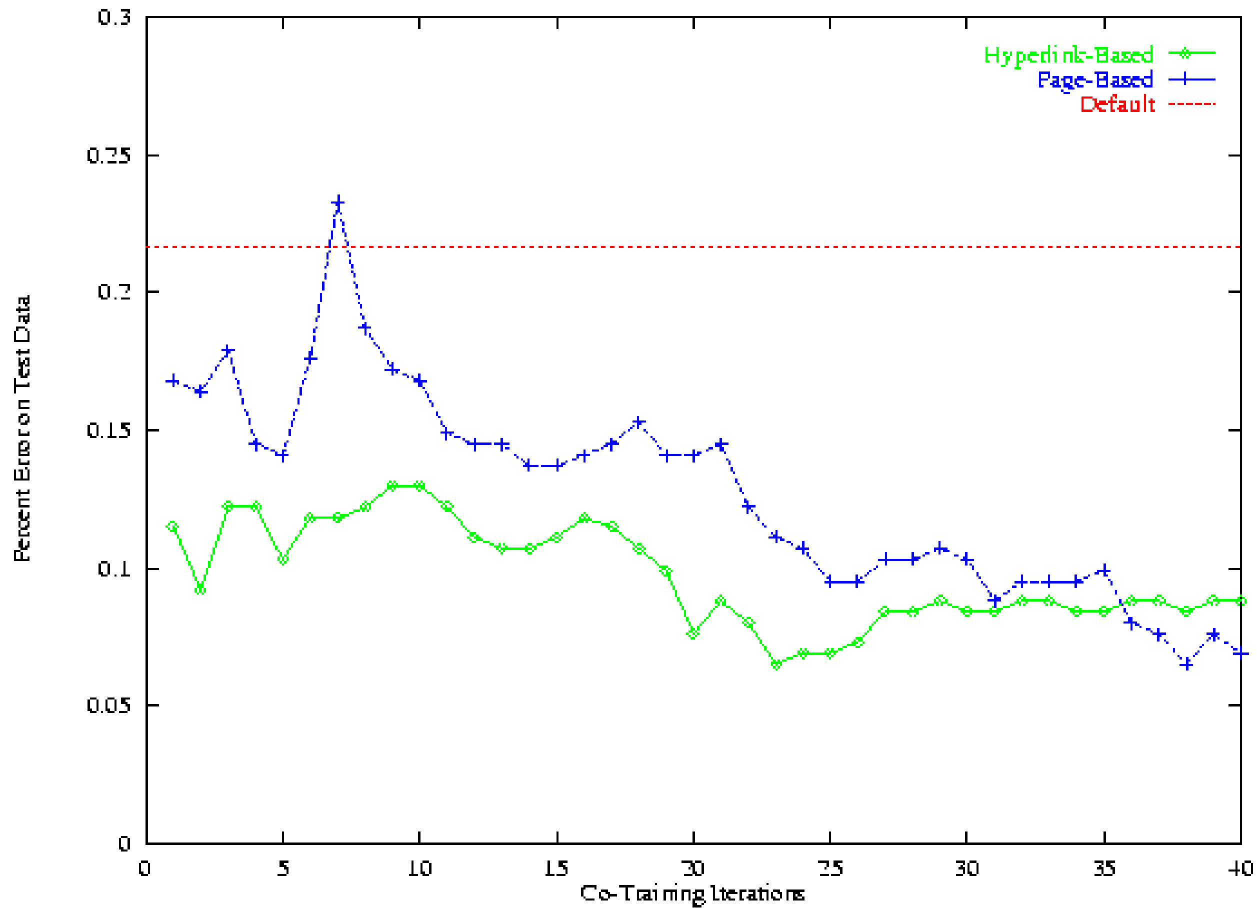
Allow g2 to label *p* positive, *n* negative examps from U

Add these self-labeled examples to L

CoTraining: Experimental Results

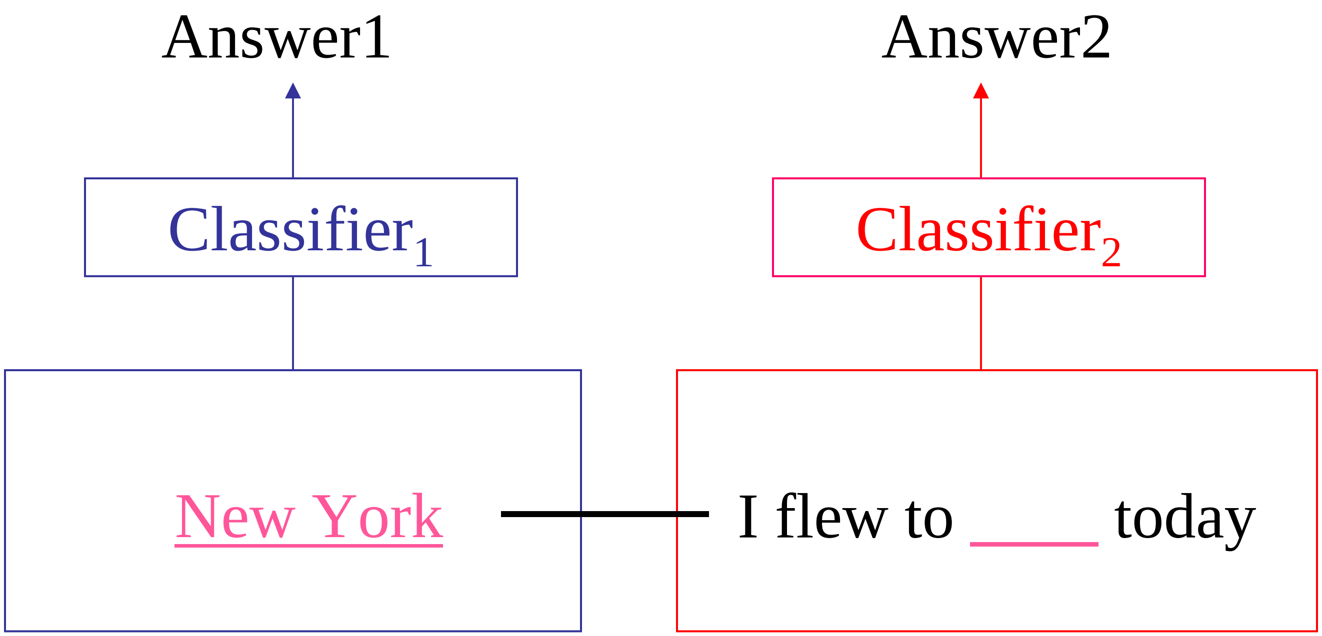
- begin with 12 labeled web pages (academic course)
- provide 1,000 additional unlabeled web pages
- average error: learning from labeled data 11.1%;
- average error: cotraining 5.0%

Typical run:



Co-Training for Named Entity Extraction (i.e.,classifying which strings refer to people, places, dates, etc.)

[Riloff&Jones 98; Collins et al., 98; Jones 05]



I flew to **New York** today.

CoTraining setting:

- wish to learn f: X → Y, given L and U drawn from P(X)
- features describing X can be partitioned (X = X1 x X2)

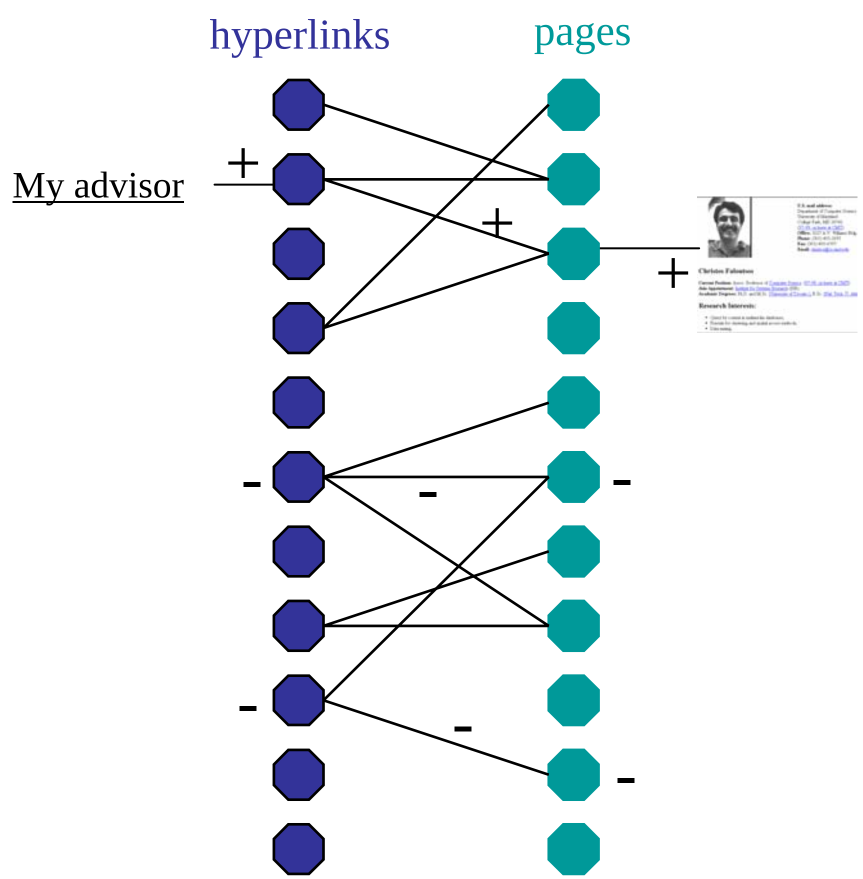
  such that f can be computed from either X1 or X2

$$(\exists g_1, g_2)(\forall x \in X) \quad g_1(x_1) = f(x) = g_2(x_2)$$

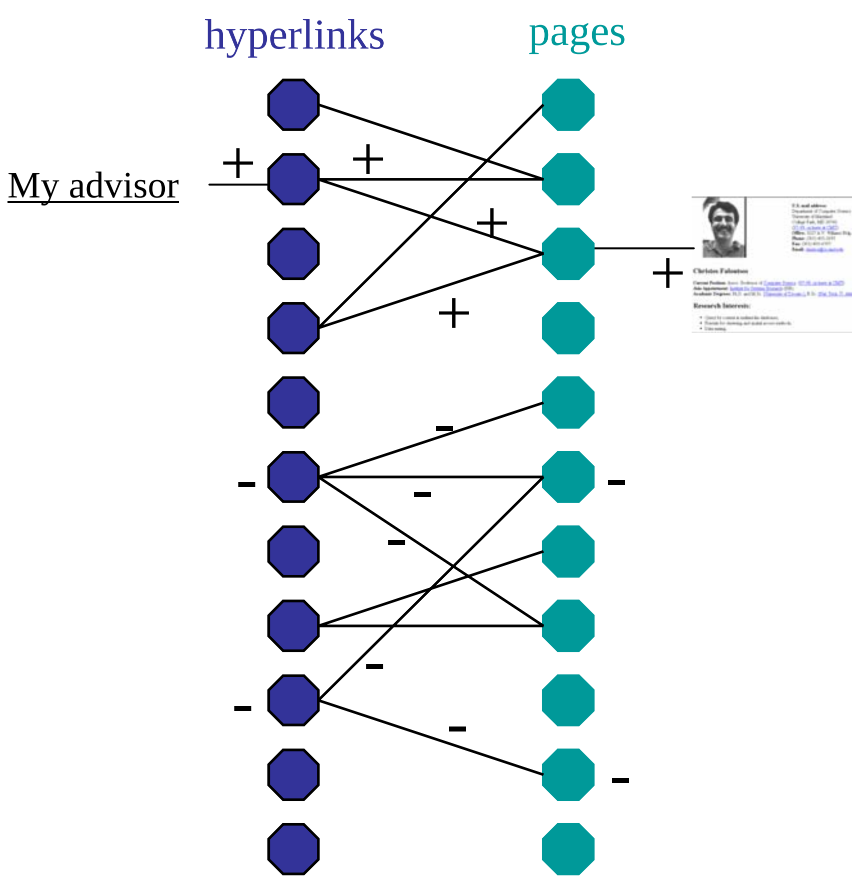
One result [Blum&Mitchell 1998]:

- If
  - X1 and X2 are conditionally independent given Y
  - f is PAC learnable from noisy *labeled* data
- Then
  - f is PAC learnable from weak initial classifier plus *unlabeled* data

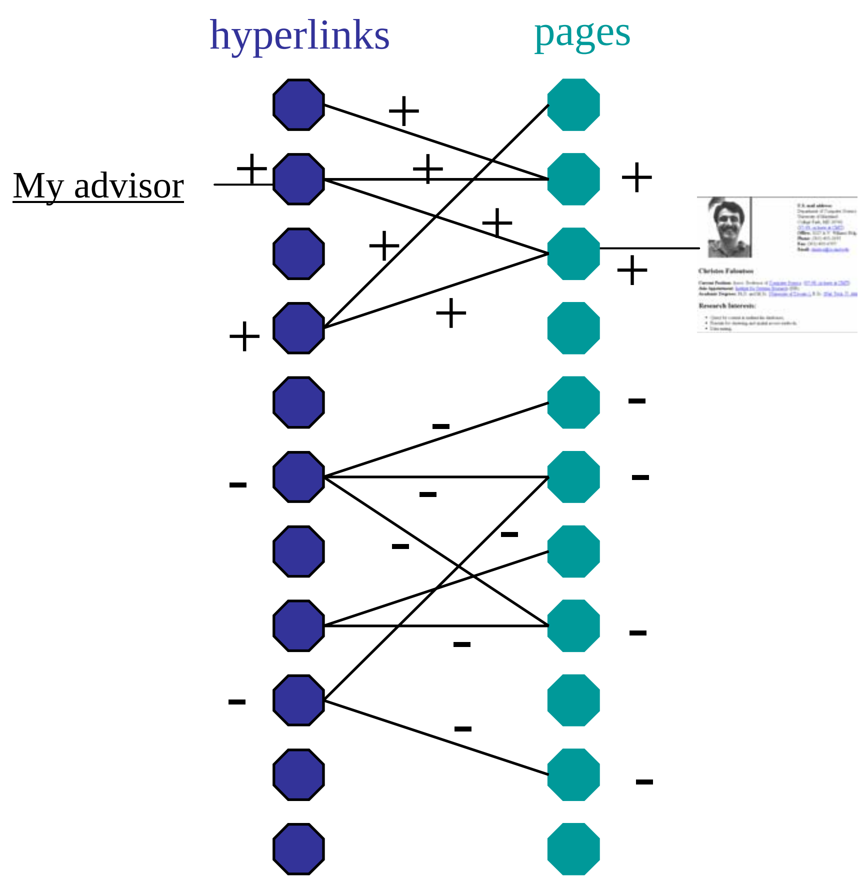
Co-Training Rote Learner



Co-Training Rote Learner



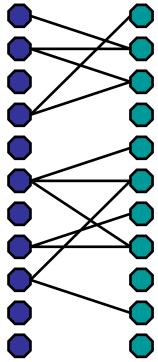
Co-Training Rote Learner



Expected Rote CoTraining error given *m* examples

$$
\begin{aligned}
&CoTraining \quad setting:\\
&learn \quad f : X \rightarrow Y\\
&where \quad X = X_1 \times X_2\\
&where \quad x \quad drawn \quad from \quad unknown \quad distribution\\
&and \quad \exists g_1, g_2 \quad (\forall x) g_1(x_1) = g_2(x_2) = f(x)
\end{aligned}
$$

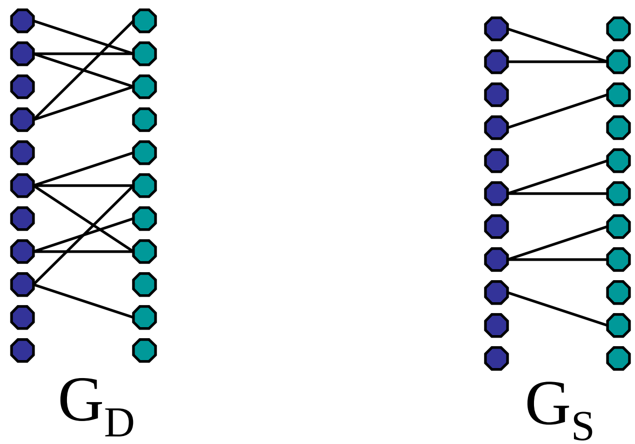
$$E[error] = \sum_j P(x \in g_j)(1 - P(x \in g_j))^m$$



Where g_j is the *j*th connected component of graph of L+U, *m* is number of labeled examples

How many *unlabeled* examples suffice?

Want to assure that connected components in the underlying distribution, G_D, are connected components in the observed sample, G_S



O(log(N)/α) examples assure that with high probability, G_S has same connected components as G_D [Karger, 94]

N is size of G_D, α is min cut over all connected components of G_D

PAC Generalization Bounds on CoTraining

[Dasgupta et al., NIPS 2001]

This theorem assumes X1 and X2 are conditionally independent given Y

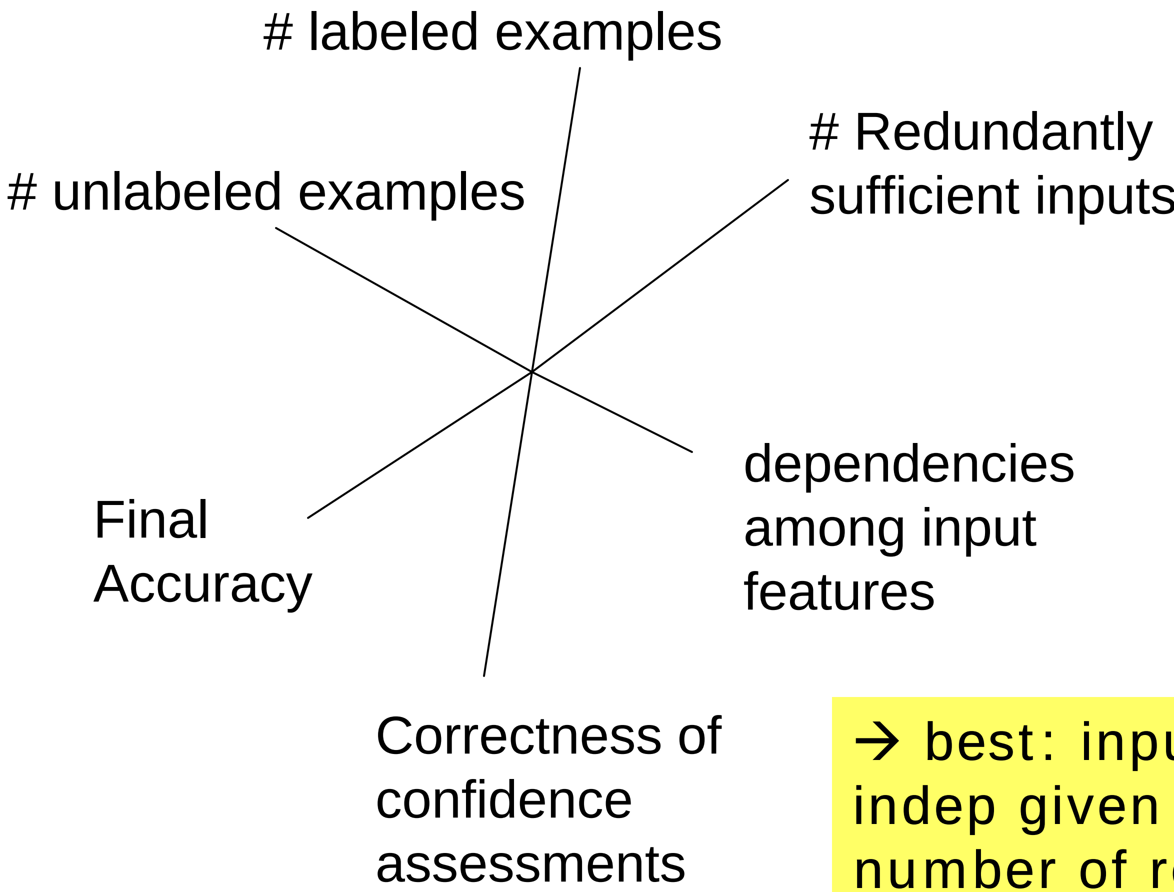
Theorem 1 *With probability at least* $1-\delta$ *over the choice of the sample* S, *we have that for all* h_1 *and* h_2, *if* $\gamma_i(h_1, h_2, \delta) > 0$ *for* $1 \le i \le k$ *then (a)* f *is a permutation and (b) for all* $1 \le i \le k$,

$$P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp) \;\; \le \;\; \frac{\widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp) + \epsilon_i(h_1, h_2, \delta)}{\gamma_i(h_1, h_2, \delta)}.$$

The theorem states, in essence, that if the sample size is large, and h_1 and h_2 largely agree on the unlabeled data, then $\widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp)$ is a good estimate of the error rate $P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp)$.

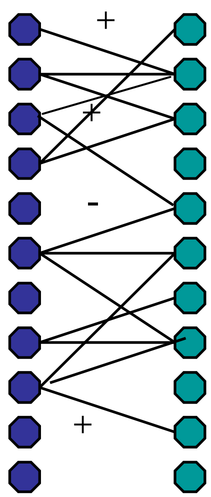
Co-Training Theory

How can we tune learning environment to enhance effectiveness of Co-Training?



→ best: inputs conditionally indep given class, increased number of redundant inputs, ...

What if CoTraining Assumption Not Perfectly Satisfied?



- Idea: Want classifiers that produce a *maximally consistent* labeling of the data
- If learning is an optimization problem, what function should we optimize?

What Objective Function?

$$E = E1 + E2 + c_3 E3 + c_4 E4$$

Error on labeled examples

$$E1 = \sum_{<x,y>\in L} (y - \hat{g}_1(x_1))^2$$

$$E2 = \sum_{<x,y>\in L} (y - \hat{g}_2(x_2))^2$$

Disagreement over unlabeled

$$E3 = \sum_{x\in U} (\hat{g}_1(x_1) - \hat{g}_2(x_2))^2$$

Misfit to estimated class priors

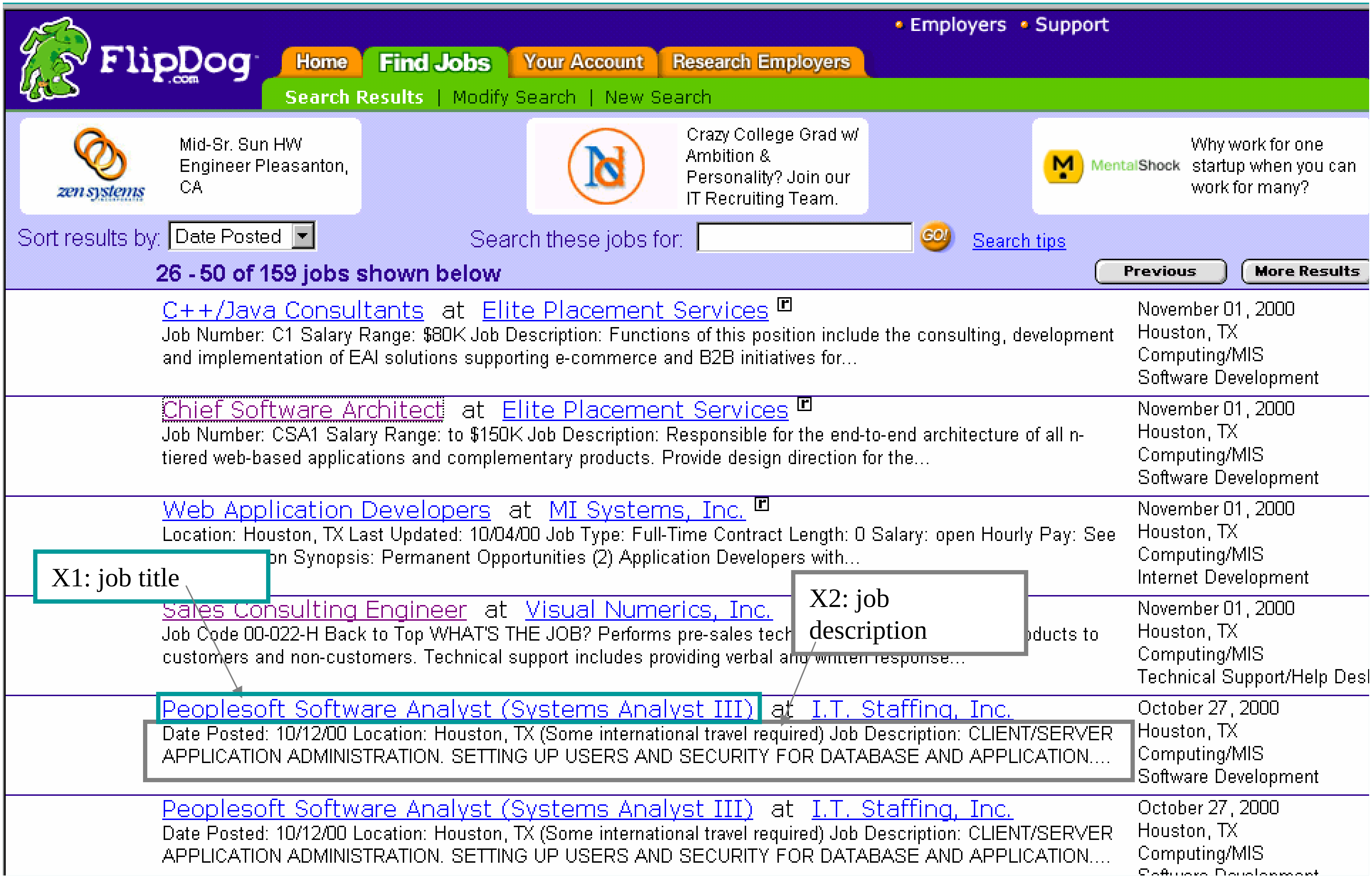
$$E4 = \left(\left(\frac{1}{|L|} \sum_{<x,y>\in L} y \right) - \left(\frac{1}{|L|+|U|} \sum_{x\in L\cup U} \frac{\hat{g}_1(x_1) + \hat{g}_2(x_2)}{2} \right) \right)^2$$

What Function Approximators?

$$\hat{g}_1(x) = \frac{1}{1 + e^{\sum_j w_{j,1} x_j}} \qquad \hat{g}_2(x) = \frac{1}{1 + e^{\sum_j w_{j,2} x_j}}$$

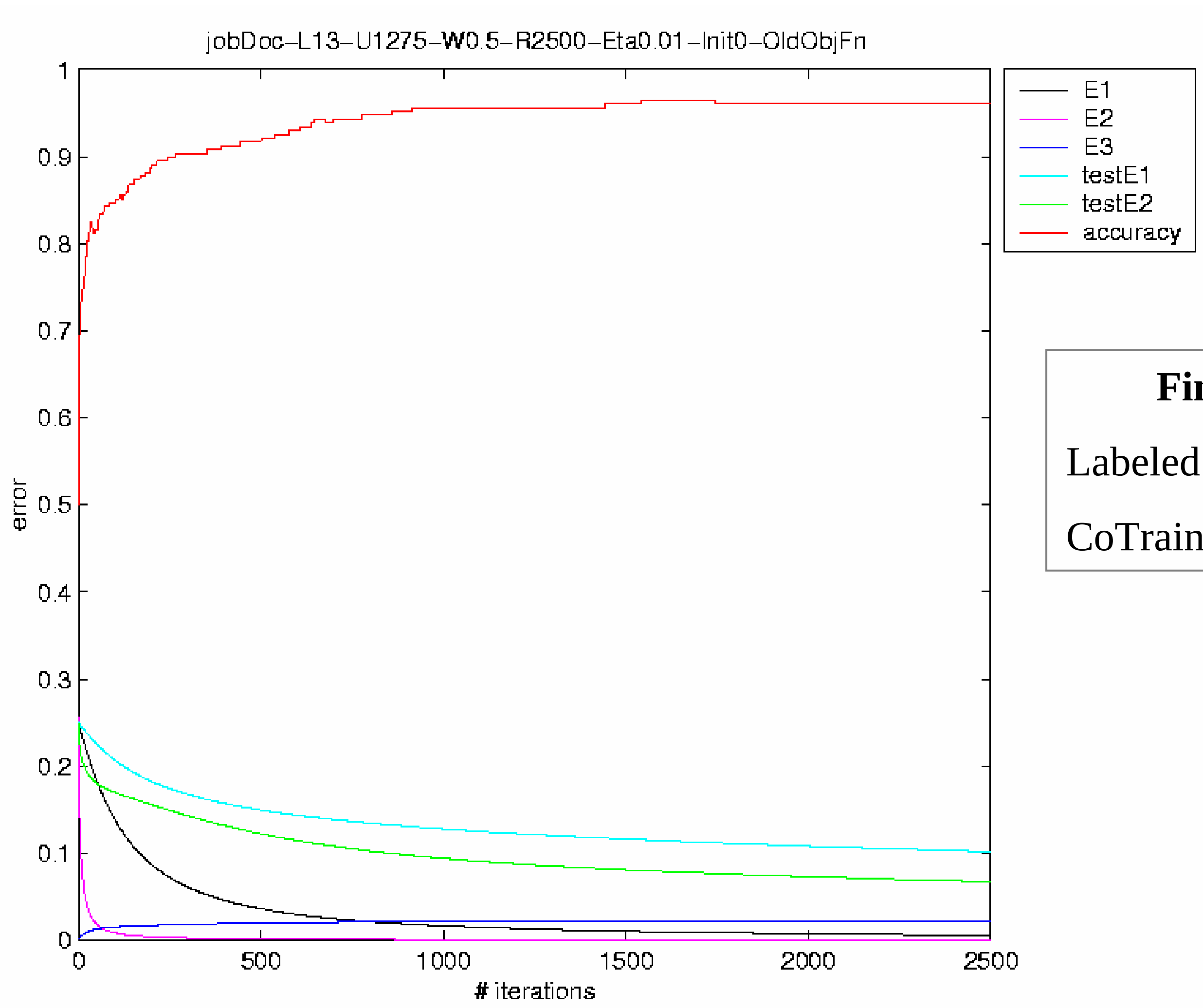
- Same functional form as logistic regression
- Use gradient descent to simultaneously learn g1 and g2, directly minimizing E = E1 + E2 + E3 + E4
- No word independence assumption, use both labeled and unlabeled data

Classifying Jobs for FlipDog



Gradient CoTraining

Classifying FlipDog job descriptions: SysAdmin vs. WebProgrammer



Final Accuracy

Labeled data alone: 86%

CoTraining: 96%

Gradient CoTraining

Classifying Capitalized sequences as Person Names

Eg., "Company president <u>Mary Smith</u> said today…"

x1 x2 x1

Error Rates

	25 labeled 5000 unlabeled	*2300 labeled 5000 unlabeled*
Using labeled data only	.24	.13
Cotraining	.15 *	.11 *
Cotraining without fitting class priors (E4)	.27 *	

* sensitive to weights of error terms E3 and E4

CoTraining Summary

- Unlabeled data improves supervised learning when example features are redundantly sufficient
  - Family of algorithms that train multiple classifiers

- Theoretical results
  - Expected error for rote learning
  - If X1,X2 conditionally independent given Y, Then
    - PAC learnable from weak initial classifier plus unlabeled data
    - disagreement between g1(x1) and g2(x2) bounds final classifier error

- Many real-world problems of this type
  - Semantic lexicon generation [Riloff, Jones 99], [Collins, Singer 99], [Jones, 05]
  - Web page classification [Blum, Mitchell 98]
  - Word sense disambiguation [Yarowsky 95]
  - Speech recognition [de Sa, Ballard 98]
  - Visual classification of cars [Levin, Viola, Freund 03]

Bootstrap learning algorithms that leverage *redundancy*

- Classifying web pages [Blum&Mitchell 98; Slattery 99]
- Classifying email [Kiritchenko&Matwin 01; Chan et al. 04]
- Named entity extraction [Collins&Singer 99; Jones&Riloff 99]
- Wrapper induction [Muslea et al., 01; Mohapatra et al. 04]
- Word sense disambiguation [Yarowsky 96]
- Discovering new word senses [Pantel&Lin 02]
- Synonym discovery [Lin et al., 03]
- Relation extraction [Brin et al.; Yangarber et al. 00]
- Statistical parsing [Sarkar 01]

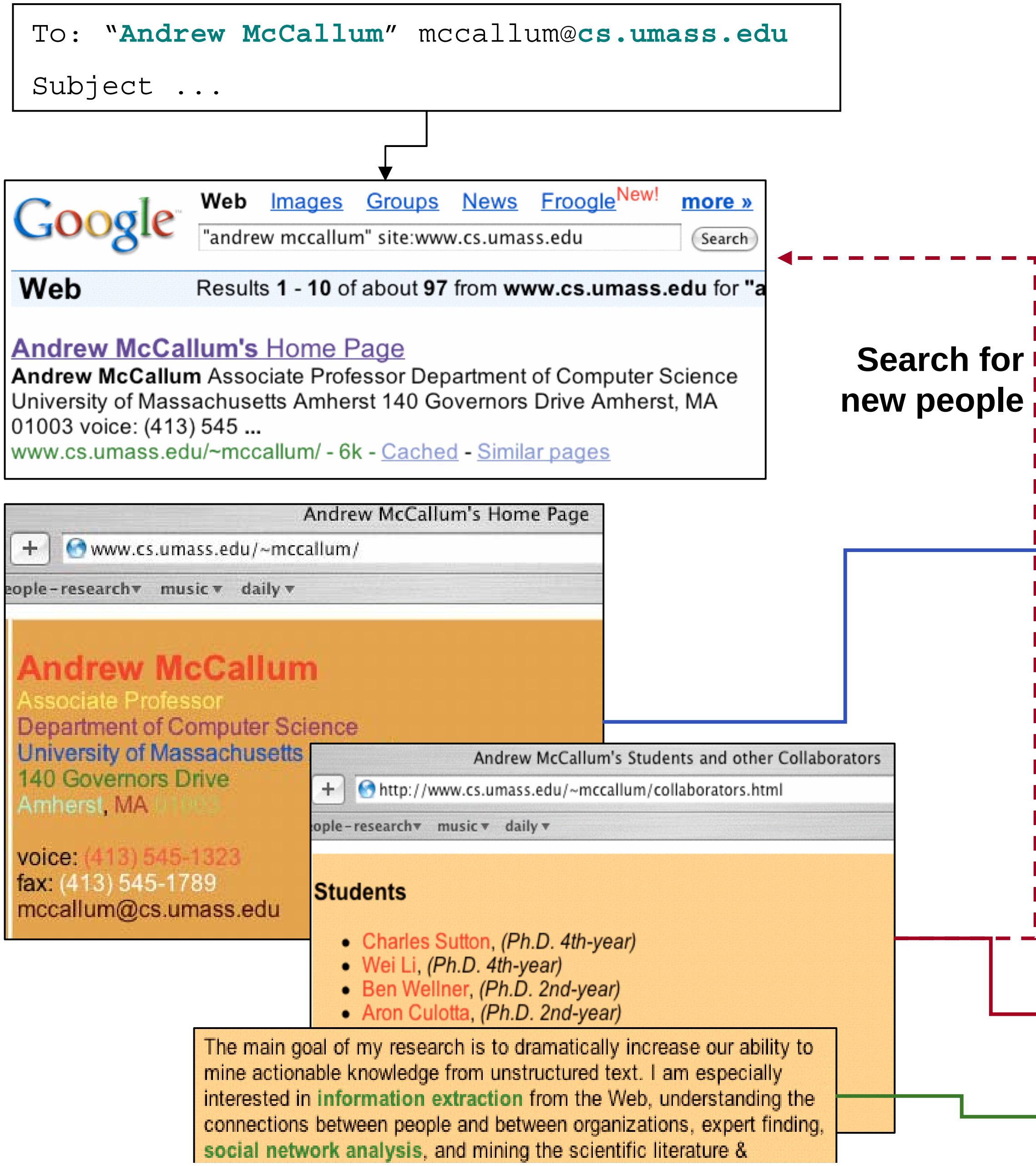
Read The Web course 10-709

1. Cover current research literature
2. Build a system that continuously bootstrap learns from web

-Large scale web information extraction [Etzioni, et al. 05]
-Graphical models for information extraction [Rosario, 05]
-Statistical parsing [Collins, et al. 05]
-Cotraining for web classification [Blum&Mitchell 98]
-Bootstrapping for natural language learning [Eisner&Karakos, 05]
-Semi-supervised learning for named entity extraction [Collins&Singer 99; Jones 05]
-Automatic learning of hypernyms [Ng, 05]
-Wrapper induction for extraction from structured web pages [Muslea et al., 01; Mohapatra et al. 04]
-Learning to disambiguate word senses [Yarowsky 96]
-Discovering new word senses [Pantel&Lin 02]
-Synonym and ontology discovery [Lin et al., 03]
-Relation extraction [Brin et al.; Yangarber et al. 00]
-Latent Dirichlet Allocation [Blei, 03]

Extracting Contact Information from the Web

[McCallum 2004]



Automatically extracted

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

What you should know

- Statistical machine learning having major impact on Natural Language Processing
  - Doc classification, Named entity extraction, Relation extraction, parsing, co-reference resolution, ontology generation, ...
- Semi-supervised methods rely heavily on unlabeled data and redundancy
- Potential for a never-ending language learning system?